ABSTRACT

Title of dissertation: INVESTIGATING CONVERSATIONAL ACCESS TO HISTORICAL CONTENT

Xin Qian, Doctor of Philosophy, 2025

Dissertation directed by: Professor Douglas W. Oard
College of Information and UMIACS

Since the inception of writing systems, people have left behind written traces of their lives. Modern media systems further extend these traces to include spoken and digital records, raising the possibility that future systems could reconstruct aspects of a person's life. This dissertation supports that vision by investigating conversational interaction with representations of historical figures that are grounded in historical content, commonly referred to as *virtual immortality*. This dissertation leverages recent advances in retrieval methods for conversational search, large language models (LLMs), and prompting techniques to propose a conceptual framework for a conversational agent comprising three stages: retrieval, contextualization, and style transfer. Rather than building a full end-to-end system, the dissertation begins with an investigation into each component individually.

First, foundational techniques for conversational search are developed through participation in TREC CAsT 2021, using web-scale document collections. Experiments demonstrate that automatic query rewriting improves retrieval performance and user-perceived utility. Second, the work examines a domain-specific setting using a curated Ronald Reagan collection—diaries, interview transcripts, and public papers. A retrieval-based prototype for single-turn interactions supports semi-structured interviews with experts from libraries,

archives, and museums (LAM). The study reveals implications for immersive experiences, archival assurance, and engagement of inquiring visitors. To help bridge this gap, the third part of the dissertation focuses on text rewriting. Specifically, the task of contextualization aims to generate clause-level *elaborations* for *uncontextualized mentions*. Reagan's letters are used as input due to their brevity, and human annotations are collected to investigate system performance. Results show that prompting LLMs with detailed task instructions and few-shot examples can achieve near-human performance, confirmed by manual inspection. Finally, the dissertation investigates text style transfer—rewriting historical content to match a conversational style reflective of Reagan's public representation. This is framed as a data-driven task, in which LLMs are prompted using comparable examples drawn from texts with topical overlap. Experiments across four sets of paired corpora show that single-turn, rather than chain-of-thought prompting, combined with short comparable examples, yields the best results. Automatic measures for entailment align well with human assessment of content preservation, while style classifiers struggle to capture the subtleties of stylistic variation when used to assess style strength. In sum, this dissertation offers an empirical investigation into components that could support conversational interaction with historical figures. It concludes with a discussion of the opportunities and limitations of such systems and aims to inspire future interdisciplinary work with born-digital historical collections.

INVESTIGATING CONVERSATIONAL ACCESS TO HISTORICAL CONTENT

by

Xin Qian

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2025

Committee:
Professor Douglas W. Oard, Chair/Advisor
Professor Wei Ai
Professor Wayne G. Lutters
Professor Louiqa Raschid, Dean's Representative
Professor Victoria van Hyning

Acknowledgments

The past seven years of my PhD journey have been an invaluable experience.

First and foremost, I extend my earnest gratitude to my advisor, Doug Oard. Doug is an exceptional scholar from whom I have learned the intricacies of research. His guidance went far beyond general advice—he provided feedback that was practical, constructive, patient, and kind—all of which maximized my chances of success and sustained my morale throughout the dissertation process. I am deeply indebted to him, as this advising relationship has inevitably demanded a significant share of his time and energy. Despite his responsibilities as a prolific researcher and now the Interim Dean of the iSchool, he has remained consistently supportive and generous in mentoring me.

I would like to thank my dissertation committee members—Professors Wei Ai, Jen Golbeck, Wayne Lutters, Louiqa Raschid, and Victoria Van Hyning—for their thoughtful and thought-provoking feedback throughout the course of this dissertation. I am also grateful to my proposal committee, including David Kirsch and David Traum, as well as my former advisor, Joel Chan, who advised me on the qualitative research methodology used in Chapter 4, resulting in a substantial contribution to this dissertation, along with other guidance.

I wish to acknowledge other faculty and collaborators who have mentored me or offered valuable suggestions: Eunyee Koh, Fan Du, Ryan Rossi, Sungchul Kim, Sana Malik, and Tak Yeon Lee during my two internships at Adobe Research; Ramya Ramakrishnan and Felix Wu from ASAPP Inc., where I had an intellectually stimulating and welcoming internship filled with novel research ideas. I am especially indebted to them for their support and collegiality during a personally difficult time then. I appreciate having been mentored by Rui Wang before beginning my PhD studies—an experience that offered not only an engineering practicum but also extensive guidance in NLP research. I also thank Katrina

Fenlon, Diana Marsh, Jason R. Baron, Kari Kraus, and Ricky Punzalan, among others, for their insights related to Libraries, Archives, and Museums (LAM) topics. Faculty members in the CLIP and HCIL labs have offered valuable research inspiration throughout the course of my PhD program.

During my nine years in graduate school—across two institutions and the years of preparation leading up to it—I have had the opportunity to meet many amazing people. At Maryland, Nilla Karunakaran and Ruth Ayele contributed their time and interest to my human annotation project. I am grateful to Suraj Nair, Petra Galuscakova, and Peter Rankel for fostering a supportive learning environment in information retrieval; Pramod Chundury, Utkarsh Dwivedi, Shawn Jenzen, Alisha Pradhan, and others for being encouraging peers in my cohort; Yuhan Luo for sharing enjoyable moments exploring the DMV area; and Yimin Xiao for helping me rehearse two of my PhD milestone talks; and to many others for their support. In addition, I thank Charles Chen for being a more experienced collaborator as we tackled challenges in research during the 2019 Adobe internship.

I also appreciate Ting Lu, who boosted my confidence in managing defense anxiety at the final stretch of this marathon; Xiaoshuang Wei, whose self-discipline and pragmatic mindset had an impact on me; and Hongyi Sheng, who encouraged me to begin GRE preparation early in my undergraduate years, among many other long-term relationships. Buttercup and Zhen'e are two dear friends who uplifted me with their wit and warmth.

Special thanks are due to the iSchool administrative and managerial team, as well as the international student advisor: Emily Dacquisto, Jeff Waters, Yen Lin, Daisy Mason, Jessica Vitak, Bijoya Chakraborty, and others for their help with procedures related to the Graduate School. I gratefully acknowledge the generous support from the Doctoral Students Research Awards (DSRA) and the Dean's Fellowship from the UMD iSchool, the Ann G. Wylie Dissertation Fellowship from the UMD Graduate School, Graduate Assistantships at UMD, and NSF Grant IIS-1816242.

Life is like an onion, and so is this acknowledgment. There are layers of gratitude, growth, and personal reflections. The content that follows may be sensitive and reflects solely my own views.

I want to thank myself for the courage to grow beyond the coercive control and retaliatory harm I experienced in the past. It took enormous strength and resilience to keep going. What happened to me is not an isolated case. Similar patterns of violence, violation, and harm have disturbed many within academia and beyond, disproportionately and most severely impacting bright female professionals. Since 2021, several tragedies and cases have unsettled academic and professional communities. I feel a personal obligation to remember and speak about three incidents as disheartening examples.

The first case involves substantiated allegations of sexual misconduct by a computer science professor at a public research university in the US. In 2020, I learned that a friend of mine was required by the school to change advisors due to the investigation into the professor that was coming to light. In 2021, the Association for Computing Machinery (ACM) concluded that the professor had violated its Policy Against Harassment, and banned him from ACM events for at least five years. According to the University's Daily, this decision was based on evidence provided by 12 individuals, with some dating back to 2016, and a School of Information PhD student played a key role throughout the process.

The second is the tragic death of Dr. Liu, a computer vision researcher affiliated with ETH Zurich and the University of Adelaide, and a former Adobe Research intern, happened in September 2023. The loss of such a talented researcher as Dr. Liu was shocking to many junior researchers. In April 2025, ETH Zurich completed its clarification process and published a report summarizing its overall conclusions.

The third case is an ongoing criminal proceeding in California involving the tragic death of a Google engineer who died from brutal and fatal head injury in Jan 2024. According to the Santa Clara County District Attorney's Office, her husband has been charged with

murdering his wife. I thank two journalists, Ms. WT from Sing Tao Daily and Ms. LK from The Paper (a major news outlet in Shanghai), for inviting me to assist in reporting and to provide commentary in early 2024. Some of those early statements have since been supported as new evidence came to light during the preliminary hearing.

Each of these cases represents only the tip of the iceberg. I have no doubt that many more go unseen or unheard. Survivors may remain silent to protect themselves, guard against perpetrators' use of DARVO tactics ("deny, attack, and reverse victim and offender"), defend against unfounded victim-blaming, or navigate institutional inertia and vexatious litigation—all of which further undermines their dignity and sense of justice. The cost can be collective: the misconduct of collaborators may compromise the research efforts of junior scholars; communities may struggle to sustain trust when ethical standards are not upheld through decisive action; and at worst, families and friends may suffer irreparable loss of a loved one.

While fully acknowledging the severity of the harm already sustained by those affected, I stand with those who have faced such challenges and encourage them to find light even in the dimmest of places. Sublimation, in the psychological sense, can help transform adversity into a source of strength—a strength only reserved for the principled soul. For me, I would especially like to thank two music professionals (and maestros) for their positive influence in me: one composer, Qigang Chen, and one pianist, Haochen Zhang. Although my connection with them was parasocial, drawn simply by their public presence, it evolved into a profound journey of self-discovery. Through their artistry, I gained new perspectives on music as something prophetic and transcendental.

Two performances of the same contemporary chamber music piece by Zhang—one at the renowned Kennedy Center in 2021 and another at a church in Napa in 2023—sparked my interest in the landscape of 20th-century music, often referred to as the modern period, with the United States as its late center. I began to recognize how modernity permeates, de-

velops, matures, and perhaps later decays within musical writing. This train of thought inspired a new interpretation amid the unsettling yet paradoxical spaces of my life. In spring 2024, I created my own visual and audio interpretation of the *Spring Round Dances* episode from Igor Stravinsky's *Rite of Spring*, moved by the masterpiece's precise and vivid portrayal of a primitive, unexperienced tale—the sacrifice of a girl who dances until her final breath. On another occasion, I gifted Zhang a plush kingfisher (martin-pêcheur), wrapped in clear plastic and tightened by a gold ribbon, with an unspoken curiosity about whether a composer's original intent truly bears the philosophical interpretations of such music, such as ideas of *Becoming-animal* and *Becoming-woman*. Zhang responded with a programming of Olivier Messiaen's *Quatuor pour la fin du Temps* in a concert in September 2024—a challenging, intellectually demanding work with a remarkable compositional background. Messiaen's own annotations on two movements speak of *"ineffable harmonies of Heaven,"* through the piano's *"soft cascades of blue-orange chords"* and contrasting the abyss of Time with the birds' symbolization of a desire for light, stars, rainbows, and jubilant song. Zhang's choice affirmed that my artistic expressions carry meaning and purpose, and whether or not it also meant tacit support, that immense encouragement continues to move me deeply today.

Early this year, I came to know the composer Qigang Chen. The fact that Chen is a living composer of our era and shares a similar cultural context meant nothing more than my admiration and innate affinity for his work. Learning that Chen became a student of Olivier Messiaen in the 1980s—a meeting Chen vividly depicted in his memoir—felt less like coincidence and more like an exquisite episode of the broader trend of Western music assimilating non-Western elements, including the *Chinese pentatonic scale*, which Messiaen analyzed yearningly from work of his predecessors such as Maurice Ravel. In part because of this, his memoir contains firsthand insights from his teacher, including some less conventional opinions. This connection also reminded me of my own advising rela-

tionship with Doug, reflecting a postgraduate advising paradigm that I've felt fortunate to experience—one that reveals similarities across disciplines, whether *music* or *information studies*. Chen's dedication to all his teachers, expressed in the book's front matter, reminded me of the immense influence of my three piano teachers: Mei Yao in Hangzhou for eight years, Naira Babayan in Maryland for two years, and my peer teacher and friend Whitney Wang for two years. I would not have gained the same understanding of music without the practical experiences they shared with me. Another poignant moment arose from Chen's reflections on grief in one chapter of his book—not as a musician, but as a parent memorizing his late son. His candid portrayal of grief as an almost unbearable human emotion resonated with me, capturing the weight felt by families affected by any tragedies, including those involved in the second and third incidents I discussed. It taught me to cherish the love and support with my parents that have sustained all of us through difficult times since 2021. Equally importantly, I express my deepest gratitude to my parents for enabling me to achieve every accomplishment in life.

I would like to conclude this acknowledgment by quoting from the final chapter, "Sunken Cathedral," of the book *The Rest is Noise: Listening to the Twentieth Century*. When I read this passage a month before completing my dissertation, I was struck by how my visual interpretation of the Spring Round Dances coincided with the author's depiction of a landscape. It felt like a sanctuary for 20th-century music at the century's end, and for me as well: a nostalgic reflection of three decades of life through the rearview mirror, revealing a destiny that has brought enduring joy and fulfillment.

> *As Highway 1, the California coastal highway, goes north of San Francisco, it holds the eyes like a work of art. The landscape might have been devised by a trickster creater who delights in grand gestures and abrupt transitions. Rolling meadows end in cliffs; redwood trees rise above slender patches of*

beach. Towers of rock rest on the surface of the ocean like the ghosts of clipper ships. A lost cow sits on the shoulder, looking out to sea. Side roads head up the inland hills at odd angles, tempting the aimless driver to follow them to the end. One especially beguiling detour, the Meyers Grade Road, departs from Highway 1 shortly after the town of Jenner. The grade is 18 percent, and the steepness of the ascent causes dizzying distortions of perspective. The Pacific Ocean rises in the rearview mirror like a blue hill across a hidden valley.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

Since writing first emerged around 3400–3000 BCE, people have left behind traces of their lives—evolving from inscriptions to recordings to vast digital footprints. Today's abundance of digitized or digitizable human experiences, collectively referred to as *historical content*, bring forward the richness of human life. Individuals and organizations preserve selected records—writing, images, music, and even algorithms—as lasting expressions of ideas to future generations. Consequently, figures like Hammurabi, Aristotle, Martin Luther King Jr., Shakespeare, Jane Austen, Mozart, Rembrandt, and Euler, remain effectively immortal through their ideas. This impulse has even extended beyond Earth, as teams of scientists and cultural heritage specialists sent the Golden Records—carefully curated compilations of human ingenuity—aboard Voyager 1 and 2 to represent humanity to potential extraterrestrial life.

Historians and cultural heritage institutions such as libraries, archives and museums (LAM) have long worked to assemble, preserve, and make accessible historical document collections tied to individuals [40, 169]. Contemporary individuals are also generating comparable archives of personal data that may be conveyed to the future. Bell and Gray define the act of preserving and transmitting ideas to the future as *one-way immortality* [18].

At the other end of the spectrum is *two-way immortality*, where a person's experiences are not only digitally preserved but also "take on a life of their own," such as an avatar that can engage with future generations [18]. Humans have long been evolutionarily attuned to acquiring and assimilating information through conversation, which makes dialog-based

approaches especially promising. The idea to emulate a deceased individual, often described as *digital immortality* or *virtual immortality*, has shifted from science fiction to early-stage reality. For instance, early systems emulating Charles Darwin [149], Albert Einstein [18], or Richard Nixon [32] were manually engineered to respond to a limited range of questions from users. Not only were these early systems limited in their functionality, but they may also fall short in supporting rich information access, particularly for educational purposes. Acquiring new information in digital form is inherently challenging for many people [37, 56], and this difficulty also applies to iterative question-answering over historical document collections, as seen in early systems for virtual immortality. This dissertation pursues a more ambitious goal: enabling open-ended, conversational interactions with representations of historical figures, grounded in their actual historical content.

Several technologies are now converging to make this possible. Among them, retrieval technologies have evolved beyond traditional web search to support conversational scenarios. When systems handle user queries within a multi-turn conversation, emphasizing the interpretation and fulfillment of evolving information needs, the task is referred to as *conversational search* [42, 128, 166]. In contrast, when ad hoc retrieval is used to ground generated dialogue responses in external knowledge sources, the task is known as *retrieval–augmented generation* (RAG) [47, 94]. Both approaches share a central idea: when a conversational agent must provide informative and grounded responses, a retrieval component is invoked to locate relevant information from an indexed collection (or knowledge source). Conversational search emphasizes effective retrieval to meet user information needs [42], while RAG focuses on integrating retrieved content into generated responses. This dissertation builds on the conversational search perspective by treating retrieval as a core capability. However, it diverges from standard RAG approaches by introducing two additional post-retrieval requirements—contextualization and text style transfer—that are unique to virtual immortality. These requirements are more tailored than the general dialog

response generation stage typical of RAG systems. Therefore, I review related work of conversational search in Chapter 2, and develop foundational retrieval techniques in Chapter 3. Chapter 4 applies general conversational search to the domain of virtual immortality, where I create a historical document collection, centered on one historical figure, Ronald Reagan, and develop a retrieval-based prototype that supports semi-structured interviews with experts from libraries, archives, and museums (LAM).

Another emerging technological development is the rise of large language models (LLM). LLMs are built on the Transformer architecture, a neural network model introduced by Vaswani et al. [164], which has become foundational for numerous natural language processing (NLP) tasks. The Generative Pre-trained Transformer (GPT) family consists of Transformer-based models, among which GPT-3 [25] was the first to quickly adapt to new tasks using only a small number of labeled examples without requiring task-specific fine-tuning. This approach, which embeds instructions for a new task directly into the prompt sent to the LLMs, is called in-context learning, where the context often consists of task descriptions and examples. These advances provide a robust foundation for text generation tasks, enabling rapid adaptation to new, low-resource domains where adequate training data may be unavailable. Relevant tasks include elaborative simplification [171], which not only simplifies input text but also adds explanatory information, and text style transfer to arbitrary target styles [131]. For virtual immortality, I introduce the requirement of contextualization, drawing on related work in elaborative simplification, to enhance clarity, as detailed in Chapter 5. Additionally, I incorporate text style transfer to ensure the content exhibits desirable communicative traits for engaging conversational interaction with the general public, as detailed in Chapter 6. Contextualization and text style transfer build on top of retrieval. The next section outlines a conceptual framework for the envisioned virtual immortality system, which includes stages for retrieval, contextualization, and style transfer. Instead of building a full end-to-end system, this dissertation examines each com-

ponent separately.

Yet another category of enabling technologies involve video synthesis, which can be used to generate highly realistic avatars of the emulated historical figure, such as those created using online tools like Synthesia.[1] This dissertation focuses exclusively on text, and does not address eventual multimodal realization, in part because existing tools already offer viable solutions in that space. It is, however, important to acknowledge that advances in video synthesis raise serious ethical considerations for digital immortality, one of the most visible being the potential for misuse through malicious deepfake content [75, 106].

The ethical landscape extends well beyond concerns about deepfakes. The very fact that it is now possible to build these systems, at least in their text-based form, inevitably leads to a deeper question: should we? Potential issues related to authenticity [77, 158], consent [102], and emotional impact on audiences [157] all contribute to the complexity of designing such systems responsibly. As part of this dissertation's contribution, grounding conversational interaction in authentic source content through retrieval functionality represents one step toward responsible system design. The need-finding study offers some preliminary insight into this ethical question, situated in the specific context of history museums. A more systematic ethical investigation is outlined as future work in Chapter 7.

1.1 Overall Framework: Retrieval, Contextualization, and Style Transfer

Figure 1.1 illustrates the overall framework for my envisioned conversational virtual immortality system, organized into three stages: retrieval, contextualization, and style transfer. The dashed gray box represents the process of completing one conversation turn, which takes a user utterance (U2) as input and generates a system-generated response (R2) from the emulated historical figure. This turn may be the first in the conversation or may

[1] `https://www.synthesia.io/`

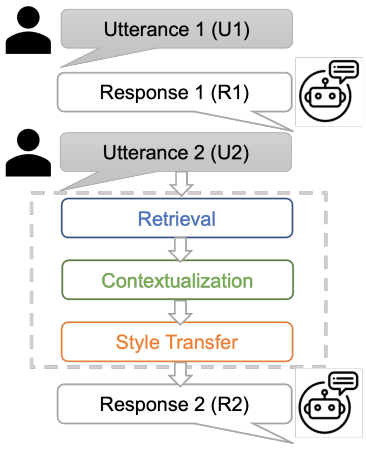


Figure 1.1: An overall framework with three stages: retrieval, contextualization, and style transfer. This dissertation explores the mechanisms of the individual components separately.

follow one or more prior turns (U1, R1, etc.). The three system components, one corresponding to each stage, are encompassed within the dashed gray box. This dissertation explores the mechanisms of each component individually and does not implement the system as a fully integrated pipeline.

A user who engages with the virtual immortality system has an evolving and potentially complex information need [42] pertaining to a historical figure. The retrieval component may optionally reformulate the utterance by incorporating relevant information from prior context (U1 and R1) using an automatic query rewriter (described in Section 3.2.1). If so, the reformulated utterance serves as the query to retrieve relevant passages from the collection; otherwise, U2 serves as the query.

The retrieved passages are segmented into sentences, with each sentence then rewritten to better suit conversational interaction. The first rewriting step I propose is contextualization. The anticipated conversational interaction takes place in a modern context that may be very different from, or even completely unrelated to, the original in terms of time, place, people, and other factors. This step identifies mentions in the retrieved sentence that require elaboration to support user comprehension, drawing on elaborative simplification [150]. Two sources of information may be selectively integrated into the origi-

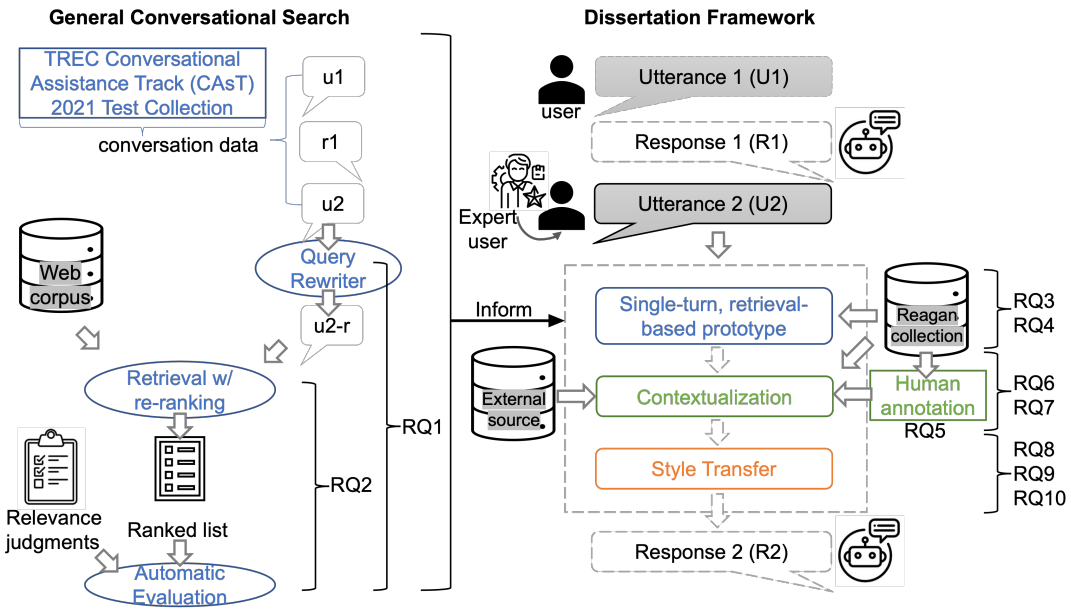


Figure 1.2: Dissertation Overview.

nal sentence: context from preceding sentences within the retrieved passage, and external sources—an approach inspired by knowledge-enhanced text generation [176].

Historical documents originate from diverse settings with varying communication goals. The second step, style transfer, rewrites the retrieved content to suit conversational interaction with the general public, such as museum visitors. This framework assumes contextualization precedes style transfer, as reversing the order may inadvertently alter the desired style. For simplicity, the current framework does not include future components such as dialogue management, which could regulate turn-taking and conversational flow.

1.2 Research Questions

The dissertation is organized into four parts: Retrieval (Chapter 3), Need-finding (Chapter 4), Contextualization (Chapter 5), and Style Transfer (Chapter 6). Figure 1.2 illustrates the overall design of this dissertation, and indicates where each specific research question fits within it.

1.2.1 Retrieval

Chapter 3 aims to develop foundational retrieval techniques applicable to conversational interactions. To support this goal, I participated in the 2021 TREC Conversational Assistance Track (CAsT), which provides a test collection of user utterances situated in a conversational context, along with corresponding relevance judgments across three large document collections. I submitted four systems to TREC CAsT 2021. All systems follow a multi-stage pipeline.[2] The pipeline begins with a query rewriter using the Transformer-based T5-base model, trained on CANARD, a dataset designed for question-in-context rewriting [51]. The pipeline then performs full-collection keyword-based retrieval using the BM25 algorithm, returning the top-1000 indexed units. The pipeline eventually performs neural network-based passage re-ranking, implemented using a Transformer-based, mono-T5 model trained on a general web search dataset named MS MARCO [119].

Query rewriting is critical to a conversational search system. Unlike traditional keyword search, conversational search needs to handle ambiguous, evolving queries, and understand context, including references and omissions [41]. Query rewriting is expected to turns brief user utterances into complete queries ready for retrieval. This motivates my first research question:

RQ1: Does query rewriting provide the retrieval system with more effective queries?

I answer RQ1 using mixed methods: quantitative evaluation and a qualitative case study. An ablation study compares submitted runs with and without query rewriting on the evaluation measures averaged over all queries (i.e., all evaluated conversational turns). TREC CAsT 2021 uses normalized Discounted Cumulative Gain at position 3 (nDCG@3)

[2]Note that the retrieval pipeline is only one part of the overall framework for this dissertation, shown in Figure 1.1.

as the primary evaluation measure, which reflects how satisfied a user would be if a voice assistant responded verbally with very few retrieved passages [41]. Results show a substantial increase in nDCG@3 with automatic query rewriting, indicating clear benefits. Further comparisons between manually rewritten queries, the CAsT organizers' baseline, and my implementations demonstrate that query rewriting achieves useful improvements over the original utterance in retrieval effectiveness.

In addition to query rewriting, I explore several configurations within the conversational search pipeline that may impact retrieval performance. These include: (1) document-level evidence augmentation, (2) indexing granularity, and (3) dense retrieval with sharding. Document-level evidence was found effective in the 2021 TREC Deep Learning track, likely because it introduces or reinforces key terms. Dense retrieval leverages contextualized language models such as BERT to encode documents and retrieve top candidates via Approximate Nearest Neighbor (ANN) search [85]. Scaling dense retrieval to a collection the size of CAsT 2021 requires some parallelism due to local computational constraints that I faced. To address this, I apply sharding, which distributes dense retrieval across smaller subsets of the collection. As these features may influence both retrieval effectiveness and efficiency, I ask my next research question:

RQ2: Which other features are helpful for retrieval performance, i.e., effectiveness and/or efficiency?

To evaluate, I compare systems using the primary measure nDCG@3, along with additional measures nDCG@5 and nDCG@500 to assess performance beyond the top ranks. The results indicate the following: (1) Document-level evidence significantly improves performance at lower ranks (nDCG@500) but not at the top (nDCG@3 or nDCG@5), with only a modest increase in index size; (2) Document-level indexing increases pipeline

runtime by a factor of six; (3) Combining dense and sparse retrieval yields statistically significant gains in effectiveness over sparse-only retrieval at lower ranks, but not at the top ranks. Overall, these three features are less effective than query rewriting in improving retrieval effectiveness.

These findings are shaped by the TREC CAsT evaluation setup, which relies on general web-scale corpora, and relevance judgments that are prone to pooling bias [27]. Furthermore, limitations in the relevance judgments, annotated at the document level, may affect retrieval effectiveness measurements in unknown ways. While these findings may not generalize directly, the retrieval techniques developed here inform the design of retrieval systems over domain-specific historical document collections, which is the focus of the next chapter.

1.2.2 Need-finding

Need-finding is a foundational process in human-computer interaction that identifies user needs and use contexts, potentially guiding the design of user-centered systems. A publicly available historical document collection centered on one historical figure, Ronald Reagan, was used in this study. I assemble the collection by scraping, parsing, segmenting, and indexing documents across three types, including 8,147 public papers, 2,902 diary entries, and 247 interview transcripts with reporters. Combined with the retrieval techniques developed in the previous chapter, I construct a prototype system that enables conversational search over this collection. The prototype supports single-turn interactions (i.e., without prior conversational context). The first research question explores the system's core utility—supporting information access:

RQ3: Could the envisioned system, as demonstrated by the prototype, return results about

historical figures that LAM experts would perceive as potentially useful?

To evaluate this, I conducted semi-structured contextual interviews with four experts in libraries, archives, and museums (LAM) who could offer insight into this specialized information access problem. Each session began with a brief tutorial and hands-on use of the prototype, followed by a think-aloud inspection of the retrieved results. The conversation then transitioned naturally to the broader implications of deploying such a system in cultural heritage settings:

RQ4: What are the opportunities and challenges for embedding the envisioned system into real information access scenarios, exemplified by that of historical museums?

Participants were prompted with topics such as *"digital archives status quo," "visitor engagement," "curated contents," "visual design," "challenges," "values,"* on which they provided responses. I analyzed interview transcripts using iterative open-coding to identify recurring themes. Overall, participants viewed the prototype as promising and affirmed its value. Their responses inform implications and lessons, including immersive experiences that extend beyond the current text-based prototype, and the importance and current tension of archival assurance. They also helped identify potential target users: museum visitors who are motivated to engage in deeper inquiry or long-term scholarly work.

From the perspective of the framework in Figure 1.1, the prototype supports content retrieval but does not implement any transformation from retrieved content to conversational responses. Thus, participant feedback reflects both near-term reactions to retrieval relevance and longer-term, conceptual perspectives on immersive future systems. Developing an end-to-end system and conducting systematic user studies are recommended directions for future work.

1.2.3 Contextualization

One motivation for rewriting retrieved historical content is to reduce information barriers. For instance, the general web search dataset MS MARCO introduces an answer generation task where systems compose answers from retrieved passages such that the answers make sense when spoken by voice assistants [13]. For my envisioned conversational interaction, if a museum visitor asks *"Any sports experience of you (Ronald Reagan)?"* and the emulated historical figure responds with *"Coach Mac, now in his 90s, is still assisting the football coach at Eureka"*—a sentence from the Reagan collection, retrieved for its its topical relevance to the user utterance—it may be confusing without further context. Names like *"Coach Mac"* and *"Eureka"* require explanation to be meaningful to a general audience. Figure 1.3 shows a 1986 letter from Reagan to his colleague Alonzo Smith, from which this sentence is taken; it includes a reference to Coach Mac and transitions into a discussion of historically Black colleges. Considering this letter, a clearer response might be: *"Coach Mac, who coached me and my teammate Alonzo, now in his 90s, is still assisting the football coach at Eureka, where I graduated in 1932 and remained closely connected throughout my life."* The added clauses, *"who coached me and my teammate Alonzo"* and *"where I graduated in 1932 and maintained a close connection throughout my life"* are two pieces of explanatory information. One draws from the preceding context and one draws from external knowledge (e.g., Wikipedia), helping users better understand the reference. This illustrates a common challenge when using personal historical documents, such as letters or diaries, where references like *"Coach Mac"* and *"Eureka"* may not be self-explanatory. Figure 1.3 shows a 1986 letter from Reagan to his colleague Alonzo Smith that includes this reference to Coach Mac and transitions into a discussion of historically Black colleges. I refer to such potentially confusing segments as *uncontextualized mentions*. The task of adding explanatory information to improve user

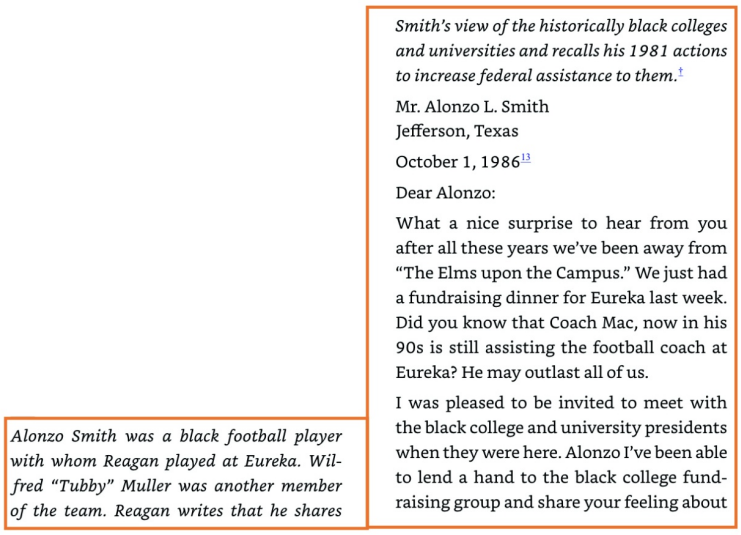


Figure 1.3: A Reagan letter with editorial notes, written to a specific correspondent, illustrating the need for contextualization.

comprehension is known as *elaborative simplification* [150]. In this dissertation, I define the contextualization requirement as generating clause-level *elaboration texts* for *uncontextualized mentions*. These elaboration texts may be derived from two sources: either from preceding context or from external sources such as a knowledge base [176]. Accordingly, a mention is labeled as either Contextual or External, based on its information source.

In the *Need-finding* chapter, I assemble the Reagan Collection, and reference a set of 1,058 Reagan's letters. To investigate the contextualization problem, I process the Reagan collection, where letter sentences are the source texts for contextualization. These letters are typically brief and written to insiders—unlike the general public I envision as the users. To explore the nature of contextualization, I collect human annotations performing contextualization on sentences from letters and measure inter-annotator agreement:

RQ5: How well do people agree when they perform the contextualization task?

These annotations also serve as reference data for evaluating system outputs. I pro-

pose four LLM-based approaches to perform contextualization, each involving one or more passes of few-shot prompting. Each approach identifies mentions that require elaboration in a letter sentence, then generates an elaboration clause for each mention.

- **Approach 1: EntLink** prompts the LLM with a standard named entity recognition recipe to identify named entities, then to generate elaborations using information from Wikipedia.

- **Approach 2: Coreference** prompts the LLM to identify Contextual mentions, then to elaborate using information from the preceding context, which is included in the input.

- **Approach 3: Sequential** combines the second and first approaches in that order, demonstrating a cascading effect.

- **Approach 4: Joint** prompts the LLM to identify both Contextual and External mentions, and to elaborate using either the preceding context, or knowledge embedded in the LLM. Its prompting instructions follow the same structure as the second approach and closely mirror the task instructions given to human annotators, making it a close simulation of how LLMs can perform the task in a human-like manner.

All approaches are implemented using OpenAI's GPT-4o model. For the Joint approach, I also include implementations using Claude 3 Sonnet and Claude 3 Haiku, resulting in a total of six systems.

System outputs are evaluated against human annotations along two dimensions: (1) the accuracy of mention identification, and (2) the quality of the generated elaboration clauses. Additionally, a complementary human evaluation uses 5-point Likert scales to rate the same two dimensions—mention identification and elaboration quality—directly probing how people perceive the effectiveness of such outputs. This evaluation design allows

us to answer:

RQ6: How well can a computer system perform contextualization?

Findings show that system-identified mentions and system-generated elaborations can be effective. System decisions on what to elaborate are often appropriate, and elaborations are generally high quality. The Joint approach achieves over 85% of human-level performance on both dimensions. A drawback of this evaluation design is the time and cost to collect human annotations and run human evaluation. Given this strong performance and the similarity of the LLM prompt to human-readable instructions, I treat output from the Joint approach as *LLM-generated annotations* and explore their use as substitute for human annotations:

RQ7: Can LLM-generated annotations substitute for human annotations in evaluation?

Using LLM-generated annotations yields system rankings similar to those based on human annotations, suggesting that they may in some cases be useful as a basis for recognizing promising systems. While promising, the comparison is limited by the small number of systems being compared, which constrains the strength of this conclusion.

1.2.4 Style Transfer

In addition to contextualization, another rewriting goal is to adapt the text to reflect the speaking style of a historical figure conversing with the general public, such as museum visitors. If the envisioned system responds to a visitor with a letter sentence from the Reagan collection, *"Also much of that aid is to help those countries have a defense*

capability of their own which is less costly than if we had to provide that defense with our own forces," the message may emphasize financial concerns. This can come across as transactional rather than strategic. A more diplomatic phrasing would be, *"And some of that aid, also, is, very practically, military aid to those countries so that they can defend themselves."* The task of rewriting a text to reflect a particular style with distinctive characteristics, while preserving its original content, is known as *text style transfer*. *Style* is defined as the distinctive characteristics of text, including word choice, syntax, metaphors, and discourse features such as stream of thought [82]. In most style transfer problems, both source and target styles are defined through examples rather than fixed rules [82], making few-shot or even zero-shot prompting with LLMs a viable approach [131]. In my problem, personal communication documents such as letters are written, non-conversational texts, therefore substantially diverging from the target style. Therefore, I designate letter sentences to serve as the source style. Sentences from four types of verbal materials—Reagan interviews with students, interviews with reporters, presidential debates, and speeches—serve as target styles. I propose five LLM-based approaches, each following a general prompting schema called SUP-NATINST [167].

- **Approach 1: NOEX** follows the SUP-NATINST schema without any few-shot examples. It specifies the target style descriptively.

- **Approach 2: SUP-NATINST** builds on Approach 1 by incorporating one or a few few-shot examples.

- **Approach 3: PREPAREDDESCR** builds on Approach 2 by specifying the target style in a declarative manner—using style attributes such as "assertive."

- **Approach 4: STYLEREASONING** builds on Approach 3 by incorporating Chain-of-Thought reasoning [168]—the LLM is prompted to first pinpoint segments that have

style misfit, and then rewrite each segment accordingly.

- **Approach 5: ENDGOAL** differs from Approach 4 by reverting to the descriptive specification of the target style in Approach 1.

I select *comparable examples*—pairs of source and target texts that overlap topically (e.g., the same policy discussed in both a Reagan letter and a public address)—as few-shot examples. Approaches 2 through 5 are each implemented in two variants (i.e., two systems), distinguished by the length of their comparable examples: Long examples use source texts with more than 22 tokens, while short examples use 22 tokens or fewer. This results in a total of nine systems.

Evaluation is conducted through human judgments of system outputs and automatic measures assessing (1) target style strength, measured by a trained classifier that estimates the probability a text belongs to the target style—this score is either averaged across texts (Average Probability) or thresholded to compute Accuracy; (2) content preservation, measured by a publicly available Multi-Genre Natural Language Inference (MNLI) entailment model. This leads to the next research question:

RQ8: How well can systems perform the proposed style transfer task?

Across four corpora, results consistently show that prompting without chain-of-thought reasoning and with short comparable examples yields the best overall performance, achieving about 75% of an ideal high baseline. Next, I test the impact of varying the number of examples, and ask:

RQ9: How does the number of prompting examples affect a system's ability to perform style transfer?

To answer ***RQ9***, I run the same automatic evaluation with varying number of prompting examples. Results show that 2-shot prompting performs best for content preservation, while 1-shot prompting yields the best target style strength. Meanwhile, a common limitation of automatic evaluation is the degree to which automatic measures such as Average Probability and MNLI approximate real-world user preferences. This motivates the next question:

RQ10: To what extent can the selected automatic evaluation measures approximate human preferences in style transfer?

I validate automatic measures by assessing their correlation with human ratings. Correlation analysis and manual inspection both show that MNLI correlates moderately well with human judgments of content preservation. In contrast, Average Probability and Accuracy, both derived from the style classifier for measuring target style strength, are overly influenced by common discourse markers in the corpus and fail to capture broader stylistic variation, leading to weak correlation with human ratings of target style strength.

1.3 Dissertation Outline

The remaining chapters of this dissertation are structured as follows: Chapter 2 provides background and reviews related work. Chapter 3 presents foundational retrieval techniques for conversational search, along with empirical results on the TREC CAsT 2021 test collection (a general, large-scale web corpus). These findings serve as inspiration for the envisioned system, whose first stage involves conversational search over smaller-scale, personal document collections. In Chapter 4, I apply retrieval techniques to develop a

single-turn, retrieval-based prototype using the Ronald Reagan collection. This prototype supports a need-finding interview study with experts from libraries, archives, and museums, uncovering potential uses, opportunities, and challenges for the envisioned system. Chapter 5 introduces the task of contextualizing retrieved texts, as the second stage of the envisioned system. I propose prompting approaches using large language models (LLMs) to identify mentions and generate elaboration texts. I then collect human annotations for each step and develop an evaluation methodology to approximate system effectiveness. In Chapter 6, I examine a new text style transfer task as the third stage of the envisioned system: rewriting texts into a more conversational form that uses public-facing language. I propose approaches that prompt LLMs with comparable examples illustrating the style transfer effect while maintaining content alignment, and conduct evaluation accordingly. Finally, Chapter 7 concludes the dissertation by summarizing contributions, reflecting on the limitations of the preceding chapters, and outlining directions for future work.

Chapter 2: Related Work

This dissertation addresses the long-standing vision of virtually emulating specific individuals, commonly referred to as *virtual immortality*. Many memorial museums and institutions in the broader cultural heritage sector are actively curating large collections of historical personal documents for the purpose of historical retelling [40, 169]. These rich resources also present growing opportunities to authentically reanimate a person's life.

To that end, this dissertation proposes a framework for virtual immortality (Section 1.1), built around three components: retrieval, contextualization, and style transfer, applied to content from a *historical document collection*. The framework is informed by emerging technologies, including retrieval techniques for conversational scenarios, large language models (LLMs) with prompting methods that are versatile at empowering text rewriting tasks, such as knowledge-enhanced text generation, elaborative simplification, and text style transfer. The following sections review these topics. Relevant evaluation methods, including statistical tests, are also introduced to lay the groundwork for the evaluation designs presented in Chapters 5 and 6.

2.1 Virtual Immortality

Museum Studies is a broad academic field that concerns museums from different angles, where history museums occupy a significant position. A unique category among history museums is museums that center around a specific historical figure. Museums of historical

figures sometimes co-locate, or share the same administration with the research library (or archives) of historical figures. Cases include the Emmett Till Interpretive Center [53] in Sumner, Mississippi, the Dalí Museum [154] in St. Petersburg, Florida, and the Abraham Lincoln Presidential Museum and Library [1] in Springfield, Illinois. Cases also extend to historical family groups include the Morgan Library and Museum in NYC and the Hagley Museum and Library in Delaware. Public venues exemplified by memorial museums curate collections about historical figures and display themed narratives for the general public.

Traditionally, *reenactments* are activities where professionals in museums or historic sites, either alone or with visitors, dress up in historic outfits, and/or demonstrate historic artifacts [11, 109]. Reenactments achieve a close simulation of past stories in the present. Controversies and budgeting are two realistic concerns with reenactments. More recently, museums that adopt interactive computer technologies have emerged to engage, enrich, and cultivate public interest [16]. The future of museums involves an interlinking between culture and education, and a requirement to creatively engage diverse audiences [163]. Memorial museums can innovate by exhibiting virtual representations of historical figures. For example, the Nixon library had a presidential forum that mimics Nixon to whom visitors could ask more than 280 curated questions [32]. This effort falls into the research area named *virtual immortality*.

Virtual immortality broadly refers to systems designed to enable conversation with representations of specific individuals, whether historical figures or fictional characters [18, 122, 137]. Researchers have hypothesized a two-stage process to achieve virtual immortality: first, creating archival records of a person; second, constructing a live avatar based on those records [18]. Many notable systems in this area are hand-engineered, using curated content to respond to a limited set of anticipated questions rather than accommodating to a wider range of, usually open-ended user questions. Examples include systems that use recordings of actors to portray Charles Darwin [149] or Albert Einstein [107],

a virtual President Nixon [32], interactive storytelling systems featuring Holocaust survivors [7, 9, 159], a representation of the playwright August Strindberg serving as a tour guide about Stockholm [19, 69], and a fictional character (Sergeant Blackwell) who answers military training questions [93].

More recent efforts, however, have focused on automating aspects of this process through advancements in AI. In 2022, Amazon Alexa's head scientist demonstrated a developing feature of its voice assistants, capable of cloning any person's voice, living or deceased, using less than a minute of recorded audio [73]. In one scenario, a child asks *"Alexa, can Grandma finish reading me the Wizard of Oz?"* and Alexa responds in a softer, more humanlike tone that seems to mimick the child's grandmother. The trend of resurrecting the dead with AI has also been examined in a Communications of the ACM article by Kugler [87], published in April 2024, which describes the growing normalization and convenience of recreating individuals through AI avatars. This trend has been largely driven by advances in generative AI, particularly large language models (LLMs) such as ChatGPT, which can generate human-like language. Since this dissertation focuses on text-based applications, a subsequent section will review relevant LLM architectures and prompting techniques.

Kugler highlighted that post-death chatbots present evolving cultural issues, whose implications vary depending on how such technologies are contextualized. He identified emotional pitfalls, including the potential to disrupt the natural grieving process and blur the line between memory and reality; legal liability such as the non-consensual use of posthumous data; and ethical concerns related to personal dignity, raising the notion of *"forced immortality."* One example is *Project December*, a 2020 interactive experience powered by GPT-3, in which users engaged with simulated versions of deceased loved ones as a way to process grief [44]. Public reactions to the system have ranged from direct criticism to mild comparisons with established practices in grief counseling and psychological therapy [60].

The need for a systematic and in-depth ethical inquiry, especially in anticipation of a multimodal realization, is acknowledged but reserved as future work in the development of the envisioned systems in this dissertation (Section 7.3.1).

2.2 Retrieval for Conversational Scenarios

The topic of retrieval in conversational contexts can be understood through two distinct perspectives, depending on the disciplinary focus. The first perspective is retrieval in support of conversational search systems, which differs from traditional web search by emphasizing interactive, multi-turn information-seeking behaviors. This task, commonly referred to as *conversational search* [128], involves interpreting and responding to a user's evolving information need throughout a dialogue. Unlike standard keyword-based ad hoc search, A theoretical framework of conversational search points that user queries in conversational search are context-dependent, they may include omissions, co-references, or ambiguities that require interpretation within a broader conversational history [128]. To benchmark research in this area, the TREC Conversational Assistance Track (CAsT) was conducted annually between 2019 and 2022. The task involves retrieving and ranking relevant passages for each user query within multi-turn sessions, using document collections such as MS MARCO [13] and Wikipedia (from the TREC Complex Answer Retrieval Track). For example, the 2019 CAsT benchmark includes 50 conversational search sessions, each consisting of roughly ten queries, accompanied by relevance judgments. A common retrieval pipeline involves automatic query rewriting, followed by BM25 (a ranking function in information retrieval to assess document relevance to a given query) for initial passage retrieval and a BERT-based re-ranker for final ranking [117]. Over the years, CAsT gradually introduced more realistic task setups. In 2020, the concept of system revealment was introduced [128] , where canonical system responses are included to provide

conversational context that user utterances may reference [42]. In 2021, this was extended by providing a single, manually selected canonical response for every query in the development set. That year also marked our participation in TREC CAsT, during which we developed foundational retrieval techniques based on the field's recent advancements.

The second perspective on retrieval in conversational settings centers on its role in generating natural language responses. This is often referred to as *retrieval-based* [166], or *retrieval-augmented generation* (RAG) [94], where retrieved content is used to ground the dialog system's output in factual sources. While the retrieval perspective in CAsT is focused on modeling and evaluating retrieval relevance, RAG emphasizes the quality of text generation that retrieval enables, especially for downstream applications such as conversational agents. Indeed, researchers in retrieval sometimes see RAG as a long-term goal: transforming retrieved content into responses that are conversationally appropriate and user-friendly. For example, MS MARCO supports a natural language generation task, where systems are required to produce answers suitable for spoken presentation via smart assistants [13]. This second perspective aligns more directly with the goals of Chapter 5 and 6, which involves rewriting historical texts to enhance conversational quality. However, both retrieval-based dialog response generation and retrieval-augmented generation are very broad and rapidly growing areas of research. A comprehensive survey would dilute the relevance to this dissertation's more focused goal. I therefore narrow the scope of related work to three strands that function as procedural steps in this dissertation: entity linking, knowledge-enhanced text generation, and text style transfer.

2.3 Large Language Models and Prompting Methods

Large language models (LLMs) have transformed the landscape of natural language processing (NLP). These models are built on the Transformer architecture, introduced by

Vaswani et al. [164], which has become the backbone for numerous NLP tasks. Among the most influential Transformer-based models is the Generative Pre-trained Transformer (GPT) family. Notably, GPT-3 [25] introduced few-shot and zero-shot learning via in-context learning—a paradigm where models are guided using natural language prompts and examples without requiring task-specific fine-tuning. In this context, crafting input prompts that specify a task, optionally including input-output example demonstrations to instruct the model on how to respond is prominent. One increasingly popular prompting strategy is chain-of-thought prompting [168], which encourages the model to generate intermediate reasoning steps before producing a final answer. This approach has been shown to improve performance on tasks that require multi-step reasoning. Prompting has proven particularly effective for novel or low-resourced tasks. Benchmark studies, such as the SUPER-NATURALINSTRUCTIONS project [12, 167], have contributed large-scale collection of tasks (ranging from 176 to over 1,600) with corresponding instructions. These benchmarks adopt a unified schema, typically consisting of a natural language task definition, positive and/or negative examples (input-output pairs), and short explanations or rationales. In this dissertation, I solve both text rewriting tasks, contextualization and style transfer, using LLMs without further fine-tuning. My focus is on prompt engineering, refining instructions and curating examples from training data to support effective few-shot learning. For example, a chain-of-thought prompting strategy is adopted in one proposed system in Chapter 6 to guide GPT-based LLMs through style transfer tasks: pinpointing text segments that do not fit to the target style, and revising the segments for the target style.

Later models such as GPT-3.5, ChatGPT, and GPT-4 further enhance LLM capabilities by improving alignment with human preferences, increasing factual accuracy, and enabling more effective conversational responses. Among them, ChatGPT is specifically optimized for dialogue generation, making it a candidate for conversational rewriting tasks. How-

ever, despite these improvements, a key limitation remains: the issue of hallucination [77], where models generate content that is fluent but factually incorrect or unverifiable. This is particularly problematic in applications on historical texts, where fidelity to source material and traceability are essential. Accordingly, this dissertation draws both inspiration and caution from the evolution of LLMs.

2.4 Entity Linking

The contextualization stage in the envisioned system in this dissertation elicits the need to first identify *uncontextualized mentions*—text spans that may lack sufficient information for interpretation by modern audiences—followed by elaborating on them. This task is closely related to research in *Entity Linking* [143], which typically follows a three-step process:

1. Identifying named entities in text;
2. Distinguishing between similarly named entities,
3. Linking each entity to a unique entry from a knowledge base (KB).

Wikipedia is a common choice of KB in Entity Linking tasks, and the associated task of *Wikification* involves identifying important words or phrases in a document and linking them to appropriate Wikipedia articles [112]. Entity Linking systems often classify mentions into three types:

- **Named mention**: specific references to a full name or partial name (e.g., *Lincoln* for *Abraham Lincoln*),
- **Nominal mention**: noun phrases that refer to an entity without using a full or partial name (e.g., *the president*),

- **Pronominal mention**: references using pronouns (e.g., *he*, *they*).

The selection of which entities to process in Entity Linking can vary. Some systems rely on predefined categories (e.g., names, organizations, locations), while others prioritize based on application-specific importance—for example, focusing on key concepts in educational content [112]. Like Wikification, while *uncontextualized mentions* as introduced in Chapter 5 may correspond to any of the three types above, only the most important ones are of interest. Unlike traditional Entity Linking, which outputs a link to a knowledge base, my contextualization process generates clause-level elaboration texts, which are better suited for conversational interaction than hyperlinks.

2.5 Knowledge-enhanced Text Generation

After identifying mentions, contextualization generates elaborative texts, but *what information should be included in such elaborations?* The field of *Knowledge-enhanced Text Generation* (KETG) [176] offers useful guidance, particularly through its emphasis on knowledge selection, which involves identifying relevant information from various sources to enhance generation. The KETG framework improves text generation by integrating both internal knowledge (information embedded in the input text) and external knowledge (drawn from sources such as knowledge bases). In this dissertation, I assume that a response by the historical figure may draw on various sources, including contemporaneous news articles, encyclopedia entries (e.g., Wikipedia), local textual context, and neighboring content from the same document collection. These assumptions are subsequently refined into a formal problem formulation (Chapter 5, Section 5.1). While KETG research explores a broad ensemble of methods for incorporating structured and unstructured knowledge, some of which include network structure, dictionaries and tables, my approach focuses specifically on integrating unstructured textual knowledge.

Elaborative simplification [150] is a specific knowledge-enhanced text generation task that mimics the work of human editors: composing documents that are both simplified and elaborative, with the original context serving as the primary source of knowledge. Srikanth et al. [150] characterized this process as a sentence-level transformation, where content additions are made to enhance contextual specificity through the insertion of definitions, explanations, or clarifications. Their approach uses GPT-2, a pretrained large language model fine-tuned on the elaborative simplification task dataset. The model was chosen for its ability to draw on world and commonsense knowledge, considered competitive at the time of development, while also attending the specific context provided in the input. The work in Chapter 5 differs from elaborative simplification in granularity. Rather than rewriting at the sentence level, my problem formulation focuses on rewriting at smaller textual units, and performs clause-level transformations. This finer-grained rewriting increase the likelihood of producing elaborations that are both contextually appropriate and semantically coherent.

2.6 Text Style Transfer

Text style transfer is the task to rewrite a piece of text to reflect a different style while preserving its original content. The general goal is to modify the form or stylistic attributes within the text, such as sentiment, formality, or persona characteristics, to meet specific requirements of an application [58]. Early work on style transfer relied on hand-crafted rules to perform paraphrasing and simplification, which could induce stylistic effects [31]. Later, statistical approaches emerged, such as models guided by word co-occurrence graphs built from large corpora in the target style [15, 79]. More recent work has adopted Transformer-based models, treating style transfer as a sequence-to-sequence text generation task [103]. Other approaches formulate it as a text-infilling task, leveraging Masked Language Mod-

els (MLMs) such as BERT or RoBERTa, which are fine-tuned to rewrite selected spans in alignment with a target style [104, 105]. GPT-based large language models (LLMs) have demonstrated strong performance in both standard style transfer tasks (e.g., sentiment transfer) and more open-ended cases, where styles are defined with natural language descriptors like *"melodramatic"* or *"uses metaphor"* [131]. Style transfer problems include formality transfer [129], gender obfuscation [130], and emulating highly distinctive voices such as characters from *Star Trek* [79].

One closely related application to Chapter 6 is author imitation [72, 80, 152, 173], where the goal is to transform content into the distinctive style of a particular author. While my dissertation shares the overarching goal of modeling conversational interactions in the voice of a historical figure, it differs from typical author imitation in an important way: in my case, the source texts were also written by the same author. That is, the original language already reflects the individual's voice in one context. The text style the transformation targets a different communicative style within the same author's voice. This intra-author variation arises naturally due to different communicative goals across document types (e.g., public speech vs. private letters).

Readers might also compare my work to persona-based dialogue systems [96, 108, 147, 177, 179], where the goal is to generate responses consistent with a predefined, often artificial persona (e.g., *"I had spag bol for dinner (dinner: spag bol"*). While there are conceptual similarities, such as evaluating style transfer strength with a classifier, akin to dialogue consistency, my task is distinct in that it does not involve generating simple, chit-chat-style responses from scratch. Instead, Chapter 6 is a self-contained style transfer problem over historical texts: transforming rich content from a historical document collection to match a desired conversational style.

2.7 Evaluation Methodology

The evaluation methodology underlying this dissertation is multi-faceted and domain-specific. In information retrieval (IR), standard evaluation measures include normalized Discounted Cumulative Gain (nDCG) [78], which compares a system's ranking to an ideal ordering with all relevant results at the top, and Mean Average Precision (MAP) [28], which computes the mean of average precision scores across all queries. Both measures are used in Chapter 3 to evaluate system output against ranked relevance judgments.

In contrast, natural language generation (NLG), relevant to Chapter 5 and 6, relies on a combination of human and automatic evaluation. Human evaluation in NLG typically follows either pointwise or pairwise approaches [95].

- **Pointwise judgments** assign scores to individual outputs, often on ordinal scales such as a 5-point Likert scale. These scores may be numeric or defined with task-specific descriptors [5]. However, pointwise ratings can be cognitively demanding and suffer from inconsistency, particularly for complex tasks [50].

- **Pairwise preferences**, where evaluators choose between two outputs, are often found to yield higher inter-annotator agreement and better reflect actual human preference [20].

Listwise ranking, also sometimes used in IR, lets evaluators rank a set of system outputs at once and has advantages such as reduced cognitive overload and resistance to cyclic inconsistencies. In Chapter 6, I adopt a hybrid evaluation design using both pointwise and pairwise formats to assess style transfer quality.

Automatic evaluation in NLG falls into two broad categories: reference-based and reference-free evaluation [153].

- **Reference-based evaluation** compares system output (the *hypothesis*) to one or more reference texts provided in a benchmark. Common reference-based measures include BLEU [121], ROUGE [100], and METEOR [14], which primarily assess n-gram overlap. In style transfer tasks, these metrics approximate qualities like content preservation [82].

- **Reference-free evaluation** scores solely based on the hypothesis. One such example is using a style classifier to compute style transfer accuracy, i.e., the percentage of outputs predicted as being in the target style.

More sophisticated automatic measures based on pre-trained neural models and of both categories are increasingly used. For example, the reference-based BERTScore [178] leverages contextual embeddings to compute semantic similarity at the token level. The reference-based MNLI model, fine-tuned on the natural language inference (NLI) dataset, assesses the semantic relationship between the original and rewritten texts [22]. In addition, the reference-free, fine-tuned text classifiers based on RoBERTa-large can be used to estimate style strength for text style transfer problems.

Beyond reporting mean measure values, this dissertation employs statistical analysis to validate findings. This includes visualizing 95% bootstrapped confidence intervals for the automatic measures, and use of a bootstrap hypothesis test for more rigorously distinguishing between system pairs when one's confidence interval overlaps with another's measure estimate (Appendix E).

Correlation analysis is a common form of meta-evaluation that can characterize evaluation quality [63]. For instance, I assess correlation between sentence-level human judgments and automatic measure values (Chapters 5 and 6). I also assess correlation between system-level measure scores derived from different sets of human annotations (Chapter 5). The following correlation measures are widely reported in the NLP and IR literature. Spear-

man's ρ [148] and Kendall's τ [84] both assess rank correlation [61]. Kendall's τ measures monotonic relationship between two set of values through the count of concordant vs. discordant pairs. Pearson correlation [123] measures linear correlation between scores rather than ranks, and Pearson rank [64] is a head-weighted, score-sensitive variant of Pearson correlation, emphasizes agreement on score magnitude in higher-ranked outputs. Spearman's ρ is Pearson correlation calculated with the rank of values instead of the actual values [140]. Accordingly, I chose the ones that fit the characteristics of the evaluation requirements.

2.8 Summary

This chapter surveys the literature that informs the dissertation's three core components: retrieval, contextualization, and style transfer. It begins by grounding these components in the broader vision of virtual immortality The chapter reviews retrieval for conversational search, and discusses the growing role of large language models (LLMs) and prompting techniques in supporting flexible, low-resource generation tasks. It then examines prior research in entity linking and knowledge-enhanced text generation, including tasks such as elaborative simplification. This is followed by a review of text style transfer. Finally, the chapter introduces a range of evaluation methodologies. Together, these strands of research frame the dissertation's contributions and clarify its position in the existing literature.

Chapter 3: Retrieval

A conversational agent has the capacity to satisfy a user's information need, expressed and interpreted through turn-taking conversational interaction. This process is known as *conversational information seeking* [39, 41] To accomplish this, the agent performs *conversational search*, an information retrieval task situated within a conversational context. This chapter presents a technical contribution to conversational search, as part of the TREC CAsT 2021 submission. The Conversational Assistance Track from the Text Retrieval Conference (TREC CAsT) is a shared task held by NIST annually between 2019 and 2022 to study Conversational Information Seeking (CIS). The developed technique is tested in this chapter with a general corpus, but is also applicable as the first step for conversational information seeking scenario that requires retrieval over personal document collections. I name this chapter as *Retrieval* instead of *Conversational Search* to reflect the nature of its role as the first stage component within my overall framework. Chapter 4 draws on these techniques to develop a single-turn, retrieval-based prototype (see Section 4.2) for need-finding in support of this dissertation.

At a conversational turn, the user leads the conversational flow by implicitly introducing a topic, continuing on the same topic, or shifting from one topic to another. Interpreting the user utterance in the context of conversational interaction and deciphering the information need is natural for humans, but challenging for machines. A process for machines to encapsulate prior conversational context and current user utterance into a reformulated query is known as query rewriting [42]. Query rewriting is a valuable step, as the reformulated

query provides a better representation of the information need than the original conversation, and this can make retrieval more effective. In 2020, the best result of retrieval systems submitted to TREC CAsT with automatic query rewriting was quite close to the best result of retrieval systems with manual query rewriting [42]. To connect with this dissertation, the overall framework in Figure 1.1 proposes a retrieval component that includes query rewriting.

This chapter describes mainly about four systems that I developed and submitted to TREC CAsT 2021.[1] For 2021, the organizers specified a baseline system that is the best retrieval system from the prior year (TREC CAsT 2020) to which new systems were compared. That system had a three-stage pipelined design (detailed in Section 3.2.1). The first stage is an automatic query re-writer, implemented using a T5-base model trained on the CANARD conversational question rewriting dataset.[2] The second stage is a BM25 retriever performing keyword-based matching. The third stage is a pointwise neural passage re-ranker implemented using a mono-T5 model trained on the MS MARCO passage retrieval dataset [118]. Our submitted systems implemented the design of the baseline's first-stage question rewriter (detailed in Section 3.2.1). Section 3.3.2 measures the effect of query rewriting. We experimented with a different second-stage full-collection retriever from the baseline design. We also implemented a pointwise re-ranker using monoBERT and consistently used it for the third stage reranking. In keeping with the track guidelines, we also measure the relative effectiveness of our second-stage full-collection retrieval pass using the end-to-end ranking quality of the full three-stage pipeline as an extrinsic measure of retrieval effectiveness (as detailed in Section 3.3.2). We explored two approaches to improve the second (full-collection retrieval) stage. The first is augmenting content representation

[1] This chapter is based on a 2021 technical report [126]. The full bibliographic citation is: Xin Qian and Douglas W. Oard. Full-Collection Search with Passage and Document Evidence. In *Text REtrieval Conference (TREC), Conversational Assistance Track 2021*. Throughout this chapter, I use "we" and "our" to reflect the collaborative nature of this work.

[2] `https://github.com/daltonj/treccastweb/tree/master/2021/baselines`

with additional document-level evidence. The second is dense retrieval performed over the full collection through sharding. Sharding trades off between retrieval effectiveness and efficiency with a distributed design to perform sparse and dense retrieval over heterogeneous document corpus. Our examination can be understood in terms of two research questions:

- ***RQ1:*** Does query rewriting provide the retrieval system with more effective queries?
- ***RQ2:*** Which other features are helpful for retrieval performance, i.e., effectiveness and/or efficiency?

The remainder of this chapter describes the task and techniques. It starts with the problem settings of TREC CAsT, followed by an overview of the participating systems, and concludes with post-hoc evaluation results using the test collection.

3.1 Task Formulation

In the submission category *automatic rewrite*, each instance of retrieval is one turn in the conversational interaction. It is structured with one raw utterance U_k, the conversational context prior to the current k-th utterance consisting of raw utterances $U' = \{U_1, U_2, \ldots, U_{k-1}\}$, and one canonical system response to each of those utterances denoted $R' = \{R_1, R_2, \ldots, R_{k-1}\}$. For each raw utterance U_k, the task is to retrieve a list of passages $P_k = \{p_{k_1}, p_{k_2}, ...\}$ as relevant responses to the utterance. While an initial retrieval pass (i.e., the second stage of our three-stage pipeline) can use traditional IR techniques for efficiency, a competitive system will presumably do another retrieval pass on the most highly ranked initial results, using a more time-consuming but also more effective pointwise re-ranker $\mathbb{M}$ that scores the probability of a passage p being relevant in the context of the current utterance U_k, denoted as $\mathbb{M}(rel = 1|p, U_k, U', R')$. The collection to be retrieved from is in general heterogeneous, constructed from several document collections

$C = \{C_1, C_2, ...\}$. For CAsT 2021, there are three such collections: an English Wikipedia dump based on the knowledge-intensive language tasks (KILT) benchmark [124], version 1 of the MS MARCO document collection, and the Washington Post (WaPo) collection, where WaPo was a new addition that year. Table 3.1 summarizes the statistics of the three parts of the CAsT 2021 collection. Passage splitting of documents is performed using a script provided by the TREC CAsT organizers, which employs the SpaCy sentence segmentation tool, followed by chunking into non-overlapping passages up to a fixed size.

Table 3.1: Statistics of the CAsT 2021 collection.

	Wikipedia/KILT	MS MARCO	WaPo
# Documents	5.0M	3.2M	0.7M
# Passages	20.0M	22.0M	3.8M
Avg # Passages per Document	4.0	7.0	5.3

3.2 Key Techniques

All of our systems follow a three-stage, pipelined baseline in the track guidelines:[3] (1) query rewriting using T5-base; (2) document retrieval using BM25 followed by segmentation to passages; (3) passage re-ranking using monoT5. They were built from combinations of the following features, which are also illustrated in Figure 3.1.

- Run #1: Baseline BM25 implementation with passage index;
- Run #2: BM25 with document index, augmented with additional document-level evidence;
- Run #3: Reciprocal rank fusion of Run #1, Run #2, and the organizer's baseline;
- Run #4: Fused dense and sparse retrieval with sharding.

[3] `https://github.com/daltonj/treccastweb/tree/master/2021/baselines`

3.2.1 A Three-stage Piepline: Query rewriting, retrieval, and re-ranking

For 2021, the organizers provided the design of a baseline system structured as a three-stage pipeline.[4] For the first stage, the CAsT baseline architecture used a T5-base model [101] to generate automatic query rewrites.[5] All of our systems implement the same idea—when given a query, along with the conversational context that is a list of canonical responses, we rewrite the query with the same pre-trained T5-base question rewriter,[6] publicly available on Huggingface.[7] This T5 rewriter is trained on the CANARD dataset [51], which includes 40K questions, to train models for question-in-context rewriting. Each instance in the CANARD dataset contains a question together with prior questions and answers, and a human-created rewritten question. As input to the rewriter, we provide the current utterance (user question), all previous utterances, and (at most) three most recent canonical system responses. We concatenate these inputs using the special separator token (|||), and the model then generates a rewritten query.

With the rewritten query, our simplest systems then perform full-collection retrieval using BM25 as the second stage, returning the top-1000 indexed units. We have two options for the indexed units, corresponding to two granularity:

- Run #1: **Passages** are segmented from documents as pre-processing, using the same passage chunker as in the organizers' baseline, then indexed, and retrieved using BM25 with default Anserini parameters ($k_1 = 0.82$, $b = 0.68$, see Figure 3.1).
- Run #2: **Full Documents** (without segmentation) are indexed, and retrieved using BM25 with organizers' baseline parameters ($k_1 = 4.46$, $b = 0.82$, see Figure 3.1).

[4] However, the implementation was not released.

[5] For clarification, query rewriting is used interchangeably with question reformulation throughout the dissertation.

[6] Note, however, that in section 3.3.3 we show that our rewriter actually generates different rewrites!

[7] `https://huggingface.co/castorini/t5-base-canard`

The retrieved documents are then chunked into passages in the same way as in Run #1. Note that because 1000 documents are retrieved, this results in more than 1000 passages.

Our third stage re-ranks the list of passage retrieval generated by the second stage using a pointwise monoBERT-large model that was pre-trained on MS MARCO. We used monoBERT-large rather than a monoT5 model used in the organizers' baseline because our pilot experiments showed limited improvements from using monoT5-base rather than monoBERT-large. In retrospect, a better choice might have been either (1) the monoT5-large model, which was found to outperform monoBERT-large by the winning team at CAsT 2020 [42], or (2) the duoBERT-large two-stage pipeline for leveraging both pointwise and pairwise training that yielded strong results for another high-scoring team at CAsT 2020 [65].

3.2.2 Augmentation with Document-Level Evidence

In Run #2 and Run #4 we experiment with adding additional document-level features to the second-stage (full-collection) retrieval process. Each collection includes a URL and some form of title for each document. In the Washington Post collection, the title is the headline of the news story. In the Wikipedia dump (KILT), the title is the HTML title field for the Wikipedia page. For the MS MARCO collection, we used the title field formatted in the raw collection tsv file. Table 3.2 lists three examples on the two fields. We tokenized both titles and URLs using Lucene's default English analyzer.

The University of Waterloo (in the 2022 TREC Deep Learning Track) reported substantial improvements (around 20 points absolute on recall@100 on dev set queries) from the use of additional content (which in their case also included section headers; we did not use the section headers in the augmented MS MARCO collection). This report suggests

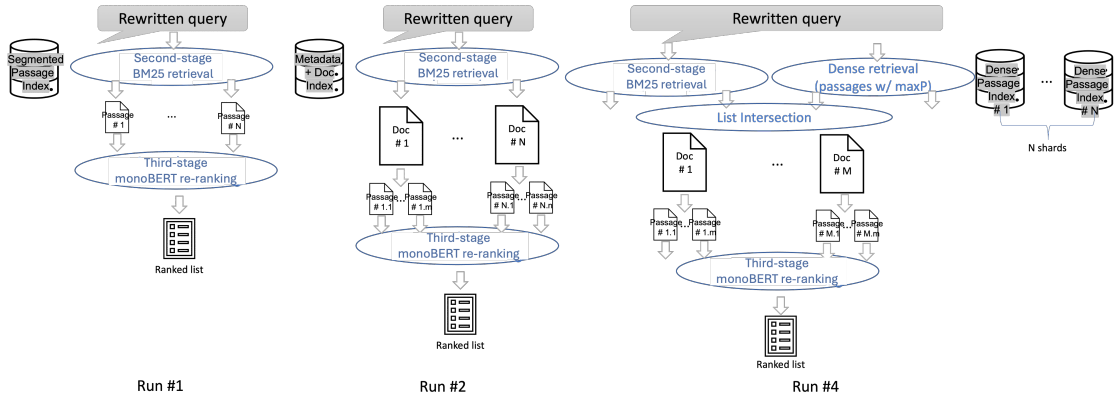


Figure 3.1: Submitted runs to TREC CAsT 2021. Compared with the baseline pipeline of query rewriting (T5-base), document retrieval (BM25), and then passage re-ranking (monoBERT-large), our Run #1 modifies the second-stage indexing unit to be passages rather than documents. Run #2 indexes documents in the second stage, but augments each document with document-level evidence. Run #3, our best run (not shown here), fuses the final output of Runs #1, #2, and the organizer-provided baseline. Run #4 fuses TCT-ColBERT (a distilled ColBERT model) passage retrieval with BM25 document retrieval as the second stage, using sharding to accommodate the large collection size.

leveraging additional document-level evidence for all three CAsT collections. This inspiration motivated a post hoc experiment in which we also used this additional document-level evidence during (third-stage) re-ranking. This additional evidence might help in two ways: it might add terms not present in the text, or it might serve to reinforce (i.e., upweight) important terms present in both the text and the additional evidence.

Collection	URL	Title/headline
Wikipedia/KILT	`https://en.wikipedia.org/w/index.php?title=Origin%20of%20the%20domestic%20dog&oldid=908221312`	Origin of the domestic dog
MS Marco	`https://www.sciencedaily.com/releases/2014/01/140107102634.htm`	Cancer Statistics 2014: Death rates continue to drop
WaPo	`https://www.washingtonpost.com/sports/colleges/danny-coale-jarrett-boykin-are-a-perfect-1-2-punch-for-virginia-tech/2011/12/31/gIQAAaW4SP_story.html`	Danny Coale, Jarrett Boykin are a perfect 1-2 punch for Virginia Tech

Table 3.2: Examples of additional document-level evidence.

3.2.3 Dense Retrieval and Sharding

Neural ranking methods using dense representations have been shown to achieve better recall than traditional ranking methods that rely on sparse representations, such as the BM25 model that we used in Run #1 and Run #2 [83]. Recently, fairly effective neural bi-encoders that are sufficiently efficient for use with moderately large collections have been introduced [85]. Scaling such approaches up to collections of the size of CAsT 2021 still requires some parallelism, however. We therefore combined sharding with an efficient neural bi-encoder. The key idea in this approach is to perform shallow-depth retrieval on shards, using an efficient approximate nearest neighbor on dense representations that are computed separately for each query and each document. The hope is that this would improve recall over that achieved using sparse methods, although achieving that benefit depends on the relevant documents being reasonably well distributed across the shards. Sharding also reduces GPU memory requirements, making it possible to fit each shard into our compute infrastructure's 200GB memory limit.

A dense encoder based on a Transformer model has length limitations on its input, so we first divide documents into passages as in Run #1. We then augment each passage with the additional document-level evidence for the document from which the passage was extracted. Next, we divide the passage collection into 200 shards of equal size, with 17 shards from the Washington Post collection, 100 shards from MS MARCO, and 82 shards from Wikipedia/KILT. When performing this sharding, we respect passage boundaries, but not document boundaries, and we assign passages to shards in order, without randomization.

For the specific dense encoder, we used the TCT-ColBERT model, which is publicly available on Huggingface.[8] This is a distilled version of ColBERT that replaces ColBERT's effective MaxSim operation in the model architecture with a less expensive dot product.

[8] `https://huggingface.co/castorini/tct_colbert-v2-hnp-msmarco-r2`

TCT-ColBERT improves the ColBERT model's efficiency while retaining most of the effectiveness. After encoding the passages, ANN search is performed with FAISS, the dense vector similarity search library. This process produces a score for each passage. We then calculate each document score using maxP (i.e., the largest of its passage scores). Independently, BM25 (on document terms only, with no augmentation) also produces a score for each document. We combine the two scores using weighted CombSUM, giving 100% of the weight to the Augmented TCT-ColBERT MaxP score and 10% of the weight to the BM25 score, as recommended in the TCT-ColBERT documentation.[9]

Unlike our other submitted runs, in Run #4 we perform third-stage re-ranking on a per-shard basis, for all passages from the top-scored documents in each shard. To obtain a final top-1000 ranked list over the full collection, we take the union of the top-scoring 50 passages from each of the 200 shards (totaling 10,000 passages), by doing CombMAX fusion of ranked lists using the polyfuse library.[10] The procedure is illustrated in Figure 3.2. Third stage re-ranking was performed on shards rather than the union of the collection for reasons of implementation convenience, and we expect the effect to be small in practice. Because all submitted runs use the same pointwise monoBERT-large re-ranker, and re-ranking is performed independently on each shard before taking the union, the final ranking—at least up to rank 50—is guaranteed to be identical to that obtained by re-ranking the union of all shard results directly. However, if fewer than 50 passages are re-ranked per shard, this may lead to the omission of passages that would otherwise receive high ranks.

[9]No normalization is performed before we apply weighted CombSUM; in future work we plan to more fully explore the design space for fusion techniques.

[10]`https://github.com/rmit-ir/polyfuse`

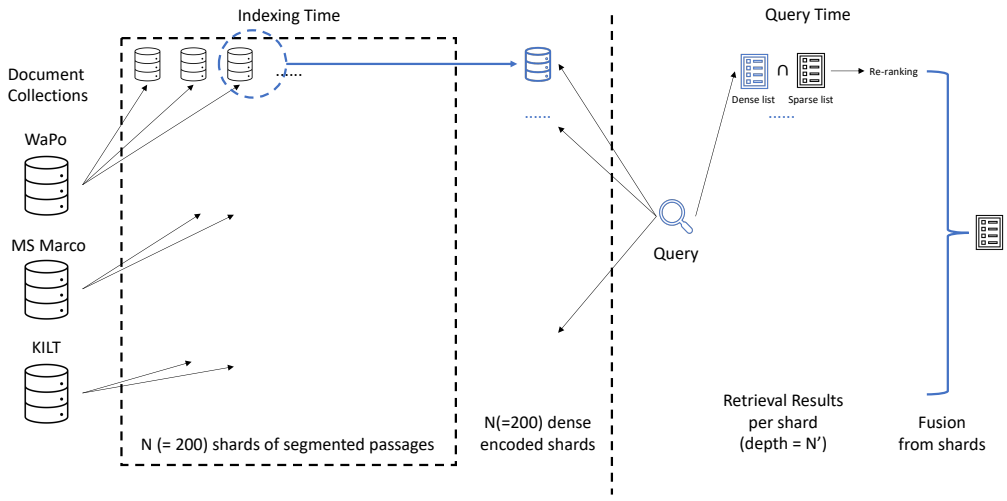


Figure 3.2: Illustrating Run #4, combining dense with sparse retrieval with sharding.

3.2.4 Fusion

Fusing the results from two or more runs is often an effective technique to improve ranking quality [98]. We tried two fusion techniques:

Run #3: To establish a high baseline, we combine the results from Run #1, Run #2, and the organizer's baseline. We use reciprocal rank fusion (RRF) with $k = 60$, which has been shown to be fairly robust [10].

Run #4: As described in Section 3.2.3, sharding produces multiple ranked lists, one per shard. We do weighted CombSUM fusion within each shard to combine the results from dense retrieval (using TCT-ColBERT) and sparse retrieval (using BM25). The recombination of results from multiple shards is also a fusion operation [165], although a simple one using CombMAX because the shards are logically disjoint.

3.3 Evaluation and Analysis

In this section, we evaluate three post hoc runs (that we didn't submit to TREC CAsT) following the same protocol used by the track organizers and discuss results. It is followed by a case study on the first-stage query rewriting.

3.3.1 Evaluation Design

We submitted a ranked list of passages for each run. The organizers notified all teams about inconsistencies in passage definitions that made passage-based evaluation impractical in 2021 [43]. Instead they chose to perform relevance judgments and evaluation at document scale. This was done by mapping ranked lists from passages to documents using MaxP (i.e., by replacing passage identifiers with the document identifier from that passage and then deduplicating the resulting document ranking by removing lower-ranked duplicates). This resulted in shallower judgment pools, and it required some changes to the relevance judgment guidelines to avoid penalizing the presence of extraneous information. We have implemented these changes to evaluate two post hoc runs in which (first-stage) query rewriting is omitted. A third post hoc run augments the passage representation used for (third-stage) re-ranking with additional document-scale evidence.

The primary evaluation measure is normalized Discounted Cumulative Gain at position 3 (nDCG@3), which focuses only on the top 3 results, and which gives decreasing weight to results in positions 2 or 3. nDCG is common in IR. nDCG@5, nDCG@500 and Mean Average Precision at 500 (MAP@500) are also reported to characterize results even lower in the ranked list. All evaluation measures are reported as averages over the computed measure for each query (i.e., for each evaluated conversational turn).

3.3.2 Evaluation Results

Table 3.3 summarizes the results for our submitted runs. They are reported as they were received from the track organizers. We also include best and median from the TREC website, which were shared internally among participants and organizers. These results are averaged over judged topics, as well as results from the organizers' baseline. '*' indicates a statistically significant improvement over the organizers' baseline using a two-sided paired t-test at $p < 0.05$ with Bonferroni correction. Similarly, '†' indicates a statistically significant improvement over Run #1, and '$\diamond$' over Run #2. Results for post hoc runs without the query formulation stage are also reported as contrastive conditions, using local automatic evaluation based on the TREC CAsT 2021 test collection.

Effect of first-stage query rewriting: The improvements from query rewriting shown in Table 3.3 are quite substantial. Comparing Run #1 or #2, with and without (wo) rewriting, illustrates this simple ablation study. Query rewriting brings very substantial absolute improvements of 0.16 or 0.21 in nDCG@3 for Run #1 or Run #2, respectively, with similar benefits observed for other measures.

Benefit of augmentation with document-level evidence: Table 3.4 shows a benefit from augmenting passages with terms from the title and URL of the document from which the passage was drawn. In this case, the additional evidence was used in both the second-stage retrieval and the third-stage re-ranking. The 0.029 absolute improvement in nDCG@3 is not statistically significant, but the comparable improvements in nDCG@5 and nDCG@500 are both statistically significant. One caveat to this analysis is that computing evaluation measures on documents rather than passages, as has been done this year, may favor the addition of document-level evidence. Note that we cannot tell with this study design whether

Table 3.3: Results of submitted and post hoc runs. '*' indicates significant improvement over the organizers' baseline, '†' over Run #1, '◇' over Run #2. Results without query rewriting were scored locally.

	nDCG@3	nDCG@5	nDCG@500	MAP@500
Best	0.8067	0.7716	0.7668	0.5483
Median	0.3820	0.3872	0.4499	0.2431
Baseline	**0.4357**	0.4265	0.5036	0.3135
Run #1	0.3875	0.3764	0.4625	0.2619
wo rewriting	0.2216	0.2100	0.2613	0.1435
Run #2	0.3985	0.3904	0.4784	$0.2811^{\dagger}$
wo rewriting	0.1888	0.1878	0.2375	0.1257
Run #3	$0.4252^{\dagger}$	$\mathbf{0.4275}^{\dagger\diamond}$	$\mathbf{0.5388}^{*\dagger\diamond}$	$\mathbf{0.3221}^{\dagger\diamond}$
Run #4	0.3768	0.3744	$\mathbf{0.5116}^{\dagger\diamond}$	$0.2841^{\dagger}$

the benefit results from the effects of augmentation on the second-stage retrieval, the third-stage re-ranking, or both. Comparing to Run #2, which used augmentation only in second-stage retrieval, is confounded by the fact that Run #2 indexed documents rather than passages for stage two. Note also that because the same document-level evidence is indexed repeatedly for each passage from the same document, there is some additional storage overhead for the index, (in this case, a relative increase of 9%).

Benefit of fusion: Comparing Run #3 to Runs #1, #2, and the organizer's baseline, we see that the fused run (Run #3) statistically significantly outperforms every individual run

Table 3.4: Benefit of retrieval from a non-augmented passage index (Run #1) vs. a passage index augmented with document title and URL terms (scored locally). '‡' indicates a statistically significant improvement over Run #1.

	Index size	nDCG@3	nDCG@5	nDCG@500	MAP@500
Original index (Run #1)	66G	0.3875	0.3764	0.4625	0.2619
Augmented index + re-ranking	72G	**0.4162**	$\mathbf{0.4100}^{\ddagger}$	$\mathbf{0.4815}^{\ddagger}$	**0.2770**

in the combination by nDCG@500, with an absolute improvement of 0.035 over the best of the three constituent runs. No statistically significant improvement over the best of the three constituent runs was observed for nDCG@3 or nDCG@5, however, indicating that the benefit observed in nDCG@500 occurs later in the ranked list.

Benefit of combining dense and sparse retrieval: As Table 3.3 shows, Run #4, which combines dense and sparse retrieval, achieves a statistically significant improvement of 0.033 over Run #2, our most closely comparable sparse-only run, by nDCG@500. However, we note no improvement from Run #2 to Run #4 in nDCG@3 or nDCG@5, indicating that the benefits we observe in nDCG@500 are coming later in the ranked list. Note also that as we have shown with Run #3, fusion can be useful even when both runs use sparse retrieval, and our study design is not able to separate that effect from specific benefits that may be accruing from using dense retrieval in the set of runs being fused.

Efficiency: The end-to-end throughput for all three stages of the Run #1 (passage-as-indexing-unit) and Run #2 (document-as-indexing-unit) pipelines are 47.3 and 304.6 seconds per query, respectively. As Table 3.2 shows, the average number of passages per topic varies between 4 and 7, and this factor of 6 or so in end-to-end processing time is thus dominated by the fact that the document-as-indexing-unit pipeline results in the generation of more passages that require re-ranking.

3.3.3 Case Study

Table 3.5 shows some examples of automatic query rewriting and a comparison with the manually rewritten queries. While the manual rewrites demonstrate desired rewriting effects, the automatic rewrites achieve some useful improvements:

Turn	Original	Manual	Official	Ours
1	I just had a breast biopsy for cancer. What are the most common types?	I just had a breast biopsy for cancer. What are the most common types of **breast cancer**?	What are the most common types of cancer **in regards to breast biopsy?**	What are the most common types of **cancer**?
2	Once it breaks out, how likely is **it** to spread?	Once it breaks out, how likely is **lobular carcinoma breast cancer** to spread?	Once the cancer breaks out, how likely is it to spread?	
3	How deadly is **it**?	How deadly is lobular carcinoma in situ?	How deadly is **LCIS?**	How deadly is **lobular carcinoma** in situ?
4	**What? No, I want to know about** the deadliness of lobular carcinoma in situ.		What is the deadliness of lobular carcinoma in situ?	I want to know about the deadliness of lobular carcinoma in situ.
5	**Wow, that's better than I thought.** What are common treatments?	**Wow, that's better than I thought.** What are common treatments for lobular carcinoma in situ?	What are common treatments for lobular carcinoma in situ?	
6	How does it behave differently from PLCIS?	How does LCIS behave differently from PLCIS?	How does LCIS behave differently from **ALH** and PLCIS?	How does lobular neoplasia behave differently from PLCIS?
7	What makes lobular cancer distinct?		What makes lobular cancer distinct from PLCIS?	What makes lobular cancer distinct?
8	For the first stage, what are the alternatives to surgery?	For the first stage, what are the alternatives to surgery for invasive lobular cancer?	For the first stage of IdC, what are the alternatives to surgery?	For the first stage of **lobular carcinoma**, what are the alternatives to surgery?
9	No, I meant for lobular.	**No, I meant -** what are the alternatives to surgery for stage 1 invasive lobular cancer	**No, I meant for lobular carcinoma.**	
10	Does freezing work?	Does freezing tumors work **as an alternative to surgery for stage 1 invasive lobular cancer?**	Does freezing work for breast cancer?	

Table 3.5: Original questions for one example topic, compared with the manually rewritten "silver standard," the automatically rewritten question from the organizers' baseline, and our automatically rewritten question. We note that in seven of the ten turns, our automatic rewrite differs from the automatic rewrite used in the organizers' baseline, despite our effort to closely replicate the usage of a T5-base model from their description. IdC = Invasive ductal carcinoma.

Utterance Simplification: Query rewriting simplifies colloquial utterances into a single-sentence, interrogative sentence, as in turns 1 and 4, making terms re-weighted.

Pronoun Resolution: Query rewriting expands pronouns such as replacing *it* with *lobular carcinoma breast cancer*, or expanding the abbreviation *LCIS*. However, this effect only appears from turn 3, but not earlier, possibly due to our context concatenation approach in which inference is on at most three previous canonical responses.

Clause Addition: Query rewriting adds clauses based on the context to complement the original utterance, where the meaning of the whole query is substantiated. Examples include turns 1 and 8, where utterances become *the most common types* ***of cancer***, and *the first stage* ***of lobular carcinoma***.

3.4 Summary

This chapter lays the foundation for the retrieval component within the framework by presenting four systems submitted to TREC CAsT 2021 and three post hoc runs that we scored locally. The four submitted runs all employ automatic query rewriting and explore three key ideas: (1) indexing either document-scale or passage-scale content (Run #1 and #2), (2) using sharding to scale dense retrieval to large collections (Run #4), and (3) combining results from different methods using result fusion (Run #3 and #4). We used the three post hoc runs to explore questions in the effect of specific techniques, including query rewriting, document-level evidence augmentation, result fusion, and the combination of dense and sparse retrieval. Results show that retrieval performance benefits significantly from automatic query rewriting, and that passage-level retrieval can be improved

using document-level cues such as titles or URLs. Furthermore, result fusion proves useful for integrating different retrieval strategies, including the combination of dense and sparse approaches. A case study based on our system outputs and the TREC CAsT 2021 corpus illustrates how rewriting can improve conversational search through effects such as utterance simplification, pronoun resolution, and clause addition.

In summary, this chapter provides a comprehensive overview of retrieval-related techniques relevant to the framework. While some of the challenges addressed in TREC CAsT 2021, such as operating at web scale, differ from those in this dissertation, some would be directly useful for supporting conversational interaction with historical document collections. Several of them inform the prototype in the next chapter (Chapter 4) that supported a need-finding study.

Chapter 4: Need-finding Study

The envisioned conversational system for interacting with historical figures first motivates a need-finding study to explore how such systems might actually be used. Specifically, I assembled a text collection centered on one historical figure, Ronald Reagan, with whom participants in this need-finding study would be reasonably familiar with. Then, I built a single-turn, retrieval-based prototype that interacts with users using content retrieved from that collection. The prototype incorporates retrieval techniques introduced in Chapter 3, reflecting the first core capability of the envisioned system. From the perspective of cultural heritage research, the system functions as an interdisciplinary boundary object [151], enabling an interview study in which four academic experts from the domains of libraries, archives, and museums responded to our vision, prompted in part by this prototype. A qualitative analysis summarizes the findings of this study.[1]

4.1 President Reagan Collection

In the United States, presidential libraries typically co-locate a historical museum built around the legacy of a presidential administration with the National Archives and Records Administration (NARA) staff that manages the records of that administration. Materials from these presidential libraries are naturally of interest to scholars (e.g., [110, 135]), but

[1]This chapter is based on work presented at the 2021 iConference Doctoral Colloquium [127]: Xin Qian, Douglas W. Oard, and Joel Chan. Conversational Interaction with Historical Figures: What's it good for? In *iConference'21*. Throughout this chapter, I use "we" and "our" to reflect the collaborative nature of this work.

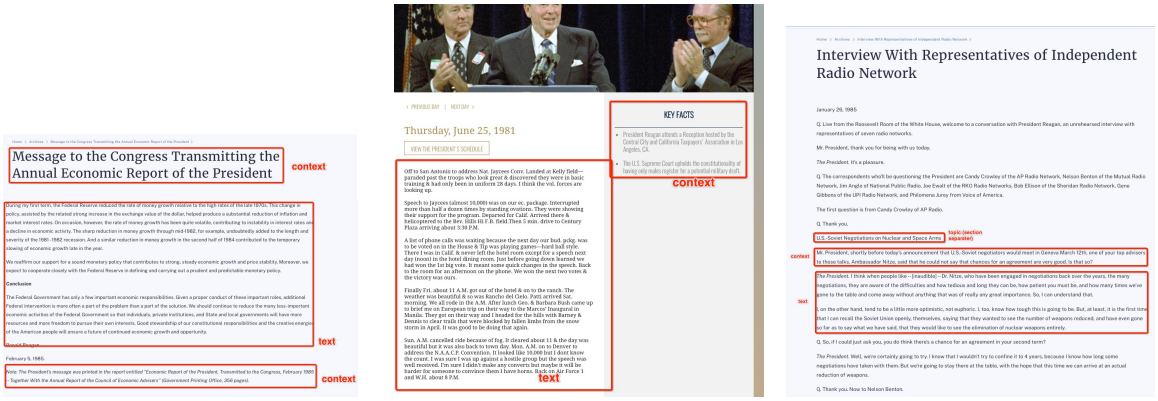


Figure 4.1: Example public paper, diary entry, and interview. The texts use fonts that are not legible for print readers. Each document type has a distinctive structure, but we process them uniformly by concatenating the document body.

for our present purpose it is the museum function of a presidential library that most interests us. The libraries of more recent presidents hold larger quantities of digital and digitized records, but because of access restrictions for more recent materials, it is the records of administrations further in the past that are able to provide access to the largest fractions of their holdings. There is thus currently a sweet spot that starts roughly with the Carter administration and continues through roughly the Clinton administration from which substantial digital or digitized materials are now available. Among that window, we chose to work with records from the Reagan administration. Among the many presidential paper collections of considerable size, Ronald Reagan's legacy has particularly attracted multiple waves of scholarly interest, with his presidential papers growing to a substantial volume by the 2000s. Bates attributes Reagan's rhetoric to a consistent worldview that is sincere and

Table 4.1: Statistics of the Reagan collection.

Document type	**Count**
Public papers	8,147
Diary entries	2,902
Interview with reporters	247

Table 4.2: Some questions that could be asked about the Reagan collection.

Question type	Example
Domestic policy	*What do you think employment levels will be next year?*
International policy	*How will the United States deal with Cuba?*
News event	*What do you think of Doonesbury's critiques of you?*
Personal affairs	*How would you like to be remembered?*

inspiring [17].

The Ronald Reagan Presidential Library and Museum website[2] contains records of the Reagan Administration from 1981 to 1989, such as speeches and reports, records of the President's daily activities, and donated personal collections. Additionally, there are a number of secondary sources on President Reagan that draw on these materials, including a biography [23] and an annotated collection of Reagan's letters [144] that can be used for research purposes (without public redistribution) under the fair use provisions of U.S. copyright law. From these web and published materials, we assembled a moderately sized, diverse text collection that offers insights into President Reagan's thoughts and statements. Although different document types have distinct structures, we process them to impose some uniformity by flattening the document body into a single field containing a list of texts, while placing surrounding context—such as titles, notes, and key facts—into separate fields. Table 4.1 gives statistics of the collection that we assembled. Each webpage is a single document. The collection includes 8,147 public papers, 2,902 diary entries, and 247 interview transcripts with reporters.

4.2 A Single-turn, Retrieval-based Prototype

The prototype aims to retrieve content relevant to user questions for a single conversation turn. Given a user question, the system uses two sequential components, the first

[2] `https://www.reaganlibrary.gov/`

of which seeks to find passages (i.e., segmented parts of documents) that contain an answer, and the second of which seeks to extract the best answer from a retrieved passage. Figure 4.2 illustrates this framework. A user question can be on any aspect of Reagan's experiences or attitudes that are documented in the collection (see Table 4.2 for examples).

4.2.1 Retrieval

The retrieval component follows the pipelined design introduced in Chapter 3, particularly Section 3.2.1. It first applies non-neural relevance matching to retrieve an initial set of highly ranked passages ("retrieval"), which are then re-ranked using a neural text matching model ("re-ranking") [119]. Automatic query rewriting was not incorporated into this retrieval component, as the current prototype only processes single-turn user questions and lacks access to conversational context. To accommodate the input length limitations of the computationally intensive reranker, all documents in the collection are segmented into overlapping 300-word passages with a stride of 150 words. These passages are indexed using ElasticSearch v.6.5.4 for efficient term-based retrieval. The top 10 passages retrieved by ElasticSearch are then reranked by using monoBERT [117], a Transformer-based BERT model [45] fine-tuned on the MS MARCO passage retrieval task. monoBERT performs pointwise classification of query-passage relevance. We use the BERT-base ver-

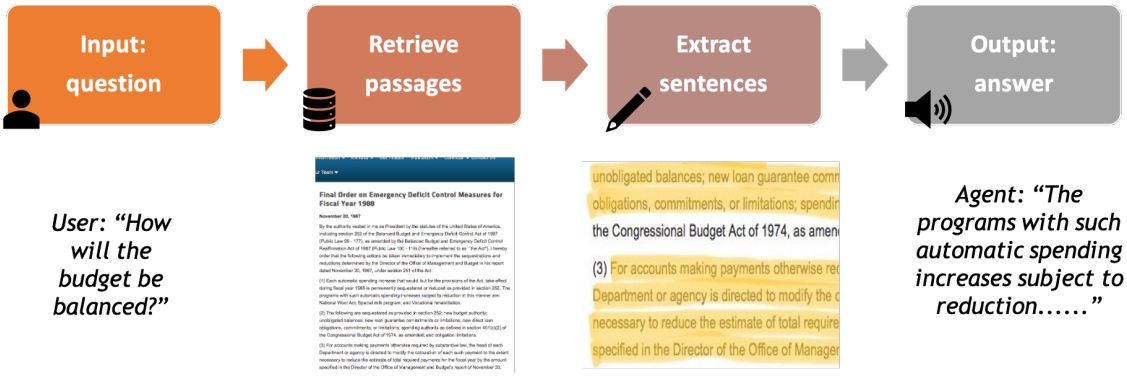


Figure 4.2: Two sequential components—retrieval and extraction—constitute the prototype.

sion released by the original authors,[3] without domain-specific fine-tuning. Each input to the reranker is a concatenation of the user's question and a candidate passage. The reranker cross-encodes this pair and outputs a matching score, which is used to sort the passages such that those most likely to contain an answer are ranked highest.

4.2.2 Extraction

The next component extracts relevant sentences from retrieved passages. We model our approach on techniques used in the Machine Reading Comprehension (MRC) task in open-domain question answering [34], which seeks to extract relevant spans from a specified text as answers to an input question. While the more general MRC setting allows spans of arbitrary length (but usually just several words) as answers, our current implementation extracts spans at the sentence level, as an answer sentence selection task [174]. Since the participants in our study could examine sentences that precede and follow the selected sentence, we felt that a sentence selection task would suffice to support our study. For an answer sentence selection task, WikiQA [174] provides a suitable data collection to train a model for this component. It includes 3,047 questions from Wikipedia-related user questions sampled from Bing query logs, paired with a total of 29,258 sentences from summary paragraphs in Wikipedia pages. We use spaCy [74] v2.2.4, specifically its Sentencizer module, to identify the sentences in each passage. The extraction model is again a neural, Transformer-based BERT text matching model. We use the Huggingface's transformers library[4] v2.4.1, fine-tune it as a classification task ourselves, with the default training procedure and hyperparameters. For fine-tuning, the extraction model cross-encodes the user's question with each sentence from a passage, and then predicts a label. At inference time, the model works with the user's question and each sentence from retrieved passages of the

[3] `https://github.com/nyu-dl/dl4marco-bert`
[4] `https://huggingface.co/transformers/`

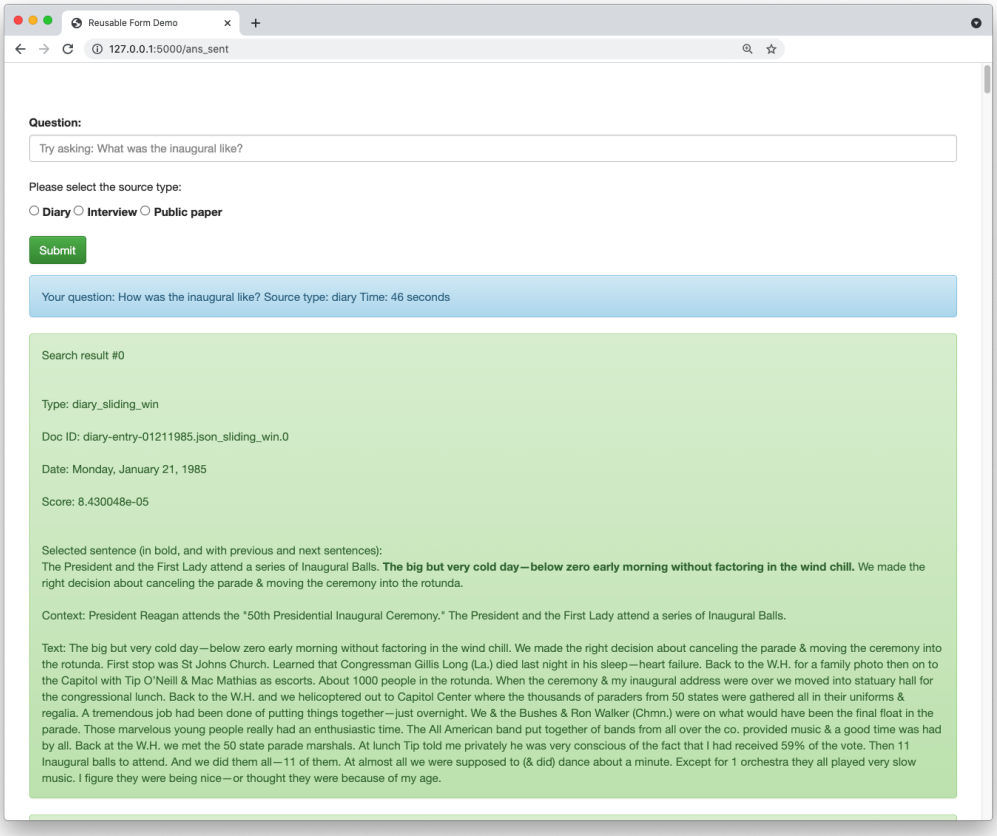


Figure 4.3: A web interface for the prototype. This screenshot shows the output of the final stage, answer sentence selection, on diaries.

Presidential Reagan collection.

4.2.3 Web Interface

The final component of our prototype is a web interface (see Figure 4.3) that can display outputs from the first step alone, the first and second steps, or all three: ElasticSearch term-matching retrieval, neural passage reranking, and answer sentence selection. Table 4.3 summarizes the typical time between issuing the question and the availability of each element of the displayed response. We focus here on effectiveness rather than efficiency; in a deployed system, these components could be optimized for efficient response, with sub-second latency.

4.3 Expert Interviews

Historical museum exhibits impart a presentation of history to visitors to fulfill a role of public education, to which selected artifacts from the collection of historical figures contribute. We chose a presidential library as a putative setting for our envisioned system. This setting occupies a middle ground in a continuum of scenarios for accessing the legacy of historical figures, ranging from formal learning, to everyday entertainment. How can systems of the type we envision support information access in this setting? The system also sits at an intersection of computer science and museum studies. This position requires a substantial translation between participants from both fields, only after which can a shared vision emerge [151]. To facilitate this translation, we used our prototype as a boundary object to support in-depth, semi-structured, contextual interviews [88, 48] with academic experts in libraries, archives, and museums (LAM). Our research questions were:

- ***RQ3:*** Could the envisioned system, as demonstrated by the prototype, return results about historical figures that LAM experts would perceive as potentially useful?
- ***RQ4:*** What are the opportunities and challenges for embedding the envisioned system into real information access scenarios, exemplified by that of historical museums?

Institutional Review Board approval was obtained in advance of our study (see Appendix A).

Table 4.3: Empirical differences (format and response time) across the three elements of the prototype: non-neural matching, neural reranking, and answer sentence selection.

	Passage retrieval		**Answer sentence selection**
	Non-neural term-matching	**Neural reranking**	
Format	Passage of up to 300 words		Single sentence
Response time	∼0 seconds	10–15 seconds	∼60 seconds

Table 4.4: Biographical sketch and research focus (rephrased for anonymity) for each participant.

ID	Biographical sketch	Research focus
P1	Female, PhD., English literature	Designing digital systems for trans-media storytelling in the virtual world
P2	Male, J.D.	Public access to archival materials, including presidential library materials
P3	Male, Ph.D., Information Studies; Graduate certificate in Museum Studies	Understanding, access and use of cultural heritage collections
P4	Female, PhD., Museum Anthropology	How technologies could enable heritage institutions to share knowledge with communities

4.3.1 Participants

Recruitment of interview participants was based on convenience sampling from public research universities in the US. We chose four academic experts in libraries, archives, and museums. Participants' ages ranged from 18 to 66. Two are female, while two are male. Three identified as Caucasian, one as Asian. Table 4.4 describes their backgrounds. Despite the sample's small size, the sample has diversity regarding individual expertise.

4.3.2 Interview Protocol

Each participant scheduled a one-on-one interview with me, conducted remotely via Zoom in March and April 2021. Interviews were audio- and screen-recorded. Each interview lasted approximately 45 minutes, with no monetary compensation.

The interview began with a tutorial video about the prototype system. The participant then tried the live system by verbally posing questions, which I (the interviewer) typed on their behalf. The participant and I inspected the results from each stage together as they appeared. Participants were asked to think aloud and share their reactions to the results.

After using the prototype, the participant and I began a semi-structured interview conversation. The conversation began with a general question about their impression towards the system: *"What kind of challenges and/or value do you see in a system like this?"* To

foster a natural conversational flow, the interviewer often prompted participants to elaborate on their answers from perspectives related to their own expertise. In addition to questions tailored to participants' experience, the interview also included a small set of more general questions that could be asked wherever appropriate, often as follow-ups to capture participants' passing thoughts in earlier responses. For example, *"How do museums engage visitors, of whom a large body only made a single visit, in a longer connection to the museum's subject?"* could be asked if the participant mentioned museum visitors; *"Text answers can be in the form of long-form narratives. What strategies do museums use to efficiently and effectively present them,"* could be asked if the conversation reached a visual aspect of museum exhibitions; or *"Museum visitors receive information from curated contents on display, how do museums decide those contents?"* could be asked if topics related to curation arose. These procedures were documented in an interview guide.

4.3.3 Analysis Approach

Overall, the analysis approach was based on iterative open-coding. The data for this was four interviews of a total duration of 3 hours and 7 minutes, auto-transcribed by Zoom. An initial pass of memoing happened during and right after each interview, where the interviewer took margin notes and then created a bullet-point summary on the printed interview guide. The notes include questions answered versus skipped, terms mentioned in participants' responses to describe or pinpoint references, and general observations. Answered interview questions from early interviews partially informed the types of questions answerable and interesting for later interviews, especially for experts with similar research directions (e.g., P3 to P4).

The four interview transcripts were imported to NVIVO v.1.4.1, then each was open-coded in separate sessions and in reverse order of interviewing (i.e., P4 to P1), to identify

themes. In these open-coding sessions, I tagged related text spans with initial codes. Operations included creating new codes or grouping spans into existing codes. Some initial codes adopted the exact wording as noted in memos. Grouping into existing codes often involved cross-checks with the code's existing text spans, and slight updates to the code. I decided span boundaries usually as turn-taking points, or at transition phrases, such as *"and I think [PAUSE]"* or *"if that makes sense"*. Some initial codes from these sessions included *"controversy in archives: ownership"* and *"archives need language (later revised as, cross-lingual) support"*. Each open-coding session was considered completed when I perceived both a reasonable coverage of coded spans and an exhaustion of discovering existing or new codes throughout an entire transcript.

A more focused coding session was then conducted on the tagged initial codes. This was an iterative process that aimed towards highest-level implications, engaging with the raw text spans as "data", in the context of each participants' background and interests. It yielded a nested structure from the initial codes, where top-level codes are close to the implications we summarize below in Section 4.3.6. A more refined sort-through adjusted that structure into a temporal or spatial logic sensible for presentation [24]. For example, P2 described a process for the inquiring users, while P3 and P4 thought through the question of who are the users. I subsequently made the former code subsidiary to the later code.

The qualitative interviewing methodology requires describing my position as the research investigator, i.e., reflecting the researcher as the "instrument" for producing knowledge from personal narratives [24]. While I had expertise in system design and development and has been influenced by views of technologies as socially constructed [125], I lacked systematic training in museum studies. A syllabus in museum studies[5] also informed my analysis.

[5]HIST 691: Museum Studies by Dr. Spencer R. Crew from George Mason University.

4.3.4 Replicability

We acknowledge an inherent complexity in our interview study protocol, where the study both uses a computational prototype (an instrument novel in design), and recruits academic experts with specific expertise. Given these considerations, conceptual replication [3] would be appropriate. To support conceptual replication, we describe the prototype configuration (Section 4.2) and summarize domain experts' backgrounds and demographics (Section 4.3.1), while keeping confidential actual participant identities pursuant to our IRB-approved study protocol.

4.3.5 Results from Prototype Exploration

On average, participants spent 11.3 minutes interacting with the system, and issued 1.5 questions. Participants asked questions impromptu (i.e., from the top of their minds) based on their past knowledge. Participants rarely commented directly about result quality being good, bad, relevant, or irrelevant. Instead, they often dug directly into result contents and spoke about insights. For example, upon seeing results that matched their expectation, P3 showed a contented smile and added that, *"Yeah, so say no to drugs was a popular slogan during the Reagan administration. So it was the Reagan project."* Subsequent conversations are also positive towards the robustness of results. Half of the participants proactively connected or compared-and-contrasted the system with mainstream search engines (P1: *"If I were querying Google, I would expect to get many news reports. What I might be able to extract from this system that I could not from Google is searching certain genres or types of information. And the addition of diaries could be very valuable..."*; P2: *"Google and others have done smaller-scale projects... It seems that your project as the digitization initiatives scale up, would become much more interesting for researchers, but you have enough."*)

Despite the promising results, one noteworthy instance of poor results came in response

to one question by P2. The question involved a named entity with a common English first name. Results mentioned a different person from Reagan's cabinet. P2 pinpointed those results as *"false positives"*. The error suggests that the system should adequately handle named entities due to the specificity of social circles in the materials of historical figures.

Overall, participants expressed a generally positive impression towards the system. P1 affirmed the value of a system that *"speaks from primary source information."* P4 associated the system's functionality with a recent, massive full-text search tool that digitized books (which had been discontinued at the time of writing this dissertation). All participants articulated potential use cases for the system, including the system being a display for users to wrap up with summative questions at the end of an exhibition (P1), being an automated desk assistant in a museum library for locating materials in binder books (P2 and P3) and being an interface that indigenous leaders could use to distill cultural values from indigenous archive collections (P4). We unify these cases into themes to be taken as implications, and characterize the target user population in Section 4.3.6.

4.3.6 Design Implications

Analysis of the interview study identified three design implications.

Implication 1: Creating immersive experiences Selecting museums as the venue underscores the importance of user experience design. P3 reflected that any possible means of visual design accompanying the textual content would enhance the system's value as a museum exhibit. P3 further suggested incorporating *"pictures, videos, maps, rather than just text,"* implying the multi-sensory requirement for effective interaction.

Some recent literature supplements P3's suggestion and highlights the importance of museums exhibiting continuity of time and place [29, 133] to contextualize the history,

which we summarize as "in-the-moment." That phrasing differentiates this idea from well-implemented "of-the-moment" displays, such as the Colonial Williamsburg Visitor Center for the eve of the American Revolution. Featuring an "in-the-moment" experience might be appropriately sensible for a museum whose subject is a historical figure. For example, visitors to the US Holocaust Memorial Museum in Washington, DC receive "identification cards" of Holocaust experiencers, to align their visit with experiencers' life narratives in sequence. If museum text materials have metadata annotations describing time and place, that could help instantiate several instances of our envisioned system for distribution across the physical museum layout. P1, an expert in digital humanities, directly pictured an "in-the-moment" system competency as if the retrieval system had provided time-travel immersion. Reading from retrieved texts, they reported that the system restored intensive memory, as if bringing them *"right back in a very emotional way"* to their earlier memories of the Reagan administration's silence on the AIDS epidemic.

At the same time, P1 raised a thoughtful concern about whether our application might trigger the *uncanny valley* effect—the discomfort users feel when a humanoid system appears almost, but not quite, human [157]. As P1 described it, *"I think there's an interesting dimension where it's almost like you're summoning the dead back to life. And that also creates some weirdness–there's a term uncanny valley in robotics and AI."* Since our prototype focuses exclusively on text and does not involve realistic visuals, I noted that such concerns are better addressed in future iterations.

Implication 2: Considering archival assurance *Assurance* refers to steps that prove item value or physical existence, and records item information [66]. Reacting to the results of a query on Reagan's Berlin Wall address, P3 critiqued the contents for being potentially unverifiable by the user of such a system since *"no visitors in the museum will be actually in that Berlin Wall address."*

The diversity of issues that might arise precludes any universal solution. For example, arranging manuscripts into a public-facing collection sometimes involves more sensitive factors than the obvious format conversion operations [67]. Developing a conversational system on top of those public-facing collections without awareness of these factors could be problematic. P4 provided an example of this, noting that indigenous materials could have been illegally confiscated from tribal groups and removed from their original cultural contexts [71], and that *"the records are usually described by predominantly white institutions and white systems and often colonial systems, i.e., systems of power that do not tend to highlight those stories or those historical figures"*.

On a related point, gaps in the records that are available for digital processing could also be a concern. Digitization remains an ongoing strategic challenge for archives [115]. Nowadays, public access to archives in a National Archives and Records Administration (NARA) research room still relies heavily on physical notebooks and printed finding aids, with only limited integration of computer databases (P2). Consistently, P4 emphasized that a major hurdle lies in digitizing archival text in any form, noting that *"very few archival collections are digitized . . . at all, let alone transcribed or . . . text searchable in any way. "*

Meanwhile, significant attention has been paid to the processing of personal papers and manuscripts [67], a responsibility increasingly shared by actors beyond professional archivists. The U.S. presidential library system, comprising at least fourteen repositories, has been a site of sustained scholarly editing. Administered by a federal agency, the system is charged with preserving and providing public access to presidential records—an endeavor shaped by agendas ranging from reducing physical storage needs to digitizing millions of textual and non-textual materials [38], and more recently, debating how to manage born-digital records since the Obama administration [54]. Public participation has also expanded through cultural heritage crowdsourcing, which has grown since online crowdsourcing gained prominence in the early 2000s [162, 161]. From 1987 and 2018, the *Ein-*

stein Papers project[6] released multiple volumes of Einstein's writings and correspondence, supported by a range of contributors including science editors and historians of physics. Similarly, *Shakespeare's World* invites volunteers to transcribe digitized manuscripts from the early modern period housed at the Folger Shakespeare Library [162]. Finally, advances in optical character recognition (OCR) and handwritten text recognition (HTR) have increasingly supported digitization efforts. As noted by a community manager for the *By the People* transcription program at the Library of Congress, such tools can make digitizing archival records not only more efficient and scalable, but also more enjoyable for participants [4].

Implication 3: Engaging inquiring visitors into community-based scholar roles Many museums include a reading room or reading room that provides access to related materials [68, 26]). Visitors who read in these libraries represent a specific population willing to invest extra time for deeper inquiry. For convenience, we describe them as "reading room" scholars.

Our envisioned system provides a progression point that could entice single-visit visitors to become reading room scholars. P1 also imagined visitors to *"come across the system at the end of the exhibition... primed to ask, well-formulated queries based on the knowledge acquired."* P1 further suggested that museum curators might well want to collaborate closely with the system designers to foster just this sort of synergy. While the system could provide a successful initial interaction of single-visit visitors with the archival content, a growing sense of curiosity or resonance could move some of the people towards long-term, scholarly work. For example, P2 pointed out that *"having a narrative"* would be beneficial to the younger population, who learn history through mass media instead of books with details, and who *"frankly are not well versed in history, of what happened in the Cuban*

[6] `https://www.einstein.caltech.edu/`

Missile Crisis or the assassination of Martin Luther King."

Connecting Implication 3 with Implication 2, we might imagine reading room scholars from the public contributing to archival assurance. P4 acknowledged the system being intriguing for *"genealogists as a huge portion of users, trying to reconstruct their family histories, e.g., Black or Latin community members."* For example, one recent change at the Smithsonian's National Anthropological Archives has been welcoming indigenous researchers to correct metadata and description in archives when they came across errors in the reading room [26].

4.4 Summary

This chapter investigated the utility of the envisioned system by assembling a historical document collection, developing a retrieval-based prototype, and drawing insights from experts in libraries, archives, and museums. These insights help us envision how a system that allows museum visitors to converse with a representation of an actual historical figure might be used in practice, and what challenges could arise in that context. Our study yielded implications related to immersive experiences, archival assurance, and supporting a continuum of engagement that may help some visitors connect more deeply with the content over time. While this interview study provides preliminary insight into how the envisioned system might be used, a gap remains between the full vision and the current prototype. To address this, the next two chapters (5 and 6) focus on two text rewriting tasks as the core of our continued technical development.

Chapter 5: Rewriting for Contextualization

My envisioned system works with a historical document collection whose documents are written well before the anticipated conversational interaction. There is disconnected context between the original and the contemporary, in terms of time, place, people involved, and other relevant factors. Directly using retrieved historical content as a system response can create an information barrier. For my envisioned conversational interaction, if a museum visitor asks *"Any sports experience for you (Ronald Reagan)?"* and the emulated historical figure responds with *"Coach Mac, now in his 90s, is still assisting the football coach at Eureka"*—a sentence from the Reagan collection, retrieved for its its topical relevance to the user utterance—names like *"Coach Mac"* and *"Eureka"* may be confusing without further context. A clearer response might be: *"Coach Mac, who coached me and my teammate Alonzo, now in his 90s, is still assisting the football coach at Eureka, where I graduated in 1932 and remained closely connected throughout my life."* I define the process of rewriting retrieved historical content to enhance user comprehension as *contextualization*.

Throughout this chapter, user comprehension is defined as the level of understanding expected from the general public with at least a high school education. After completing this dissertation, I became aware of other standards for public information accessibility—such as the federal Plain Writing Act of 2010,[1] and a state Plain Writing Equity Standard[2]

[1] `https://www.plainlanguage.gov/law/`

[2] `https://hub.innovation.ca.gov/content-design/plain-language-equity-standard/`

that recommends writing at an 8th-grade reading level or lower. Designing tasks for more specific subgroups (e.g., those reading at or above the 8th-grade level but below the high school graduation level) could yield valuable insights but may also introduce additional challenges, including the need for research protocols involving minors, verifying participants' reading comprehension levels, or or confining rather than enabling more innovative uses of cultural heritage data.

Contextualization can be helpful in several settings. First, retrieval-based conversational agents rewrite retrieved content into responses that can be understood when read aloud by a smart speaker, without requiring any additional context [13]. Second, in the completed need-finding study (Chapter 4), one interview participant highlighted the value of a conversational system that can properly introduce context to help museum visitors recall what is in the exhibition. Third, mention resolution can play a role in processing cultural heritage data, as recognized in the HIPE workshop (Named Entity Recognition (NER) and Linking on Historical Documents).[3] Fourth, there has been some work on how to provide contextual information to support user comprehension of machine responses that consist of existing content in the form of *linking dialogues*, which serve to resolve the incoherence between how users phrase their questions, and differently worded but relevant existing content [62]. Examples of linking dialogues include an introduction to the upcoming speaker (in the case of pre-recorded videos), or a statement that establishes contextual elements not made explicit in the existing content. However, the constrained format of linking dialogues underscore the need to explore more direct forms of contextualization.

The main objective of this chapter, situated in the full envisioned system, is to develop effective systems to perform contextualization. The first step towards this objective is to formulate the task. My task formulation is informed by two key observations that help

[3] `https://hipe-eval.github.io/HIPE-2022/`. Running twice in 2019 and 2022, it has the objectives to make named entity processing systems more "robust, adaptable, and transferable" on historical materials.

define the specific requirements for contextualizing historical documents. Throughout this chapter, I focus on working with sentence-level retrieved content. The first observation is that each sentence is rooted in an original context, such as the surrounding text within a paragraph. The second observation is that confusing elements in retrieved content, such as *"Coach Mac"* or *"Eureka"* from the earlier example, are concrete text references, often pointing to people, places, events, or time periods. I refer to these as *uncontextualized mentions* and use *mentions* as an abbreviation in this chapter. A formal definition of *uncontextualized mentions* is provided in Section 5.1. A key requirement for contextualization is resolving these mentions.

Resolving uncontextualized mentions shares similarities with the task of entity linking (as discussed in Chapter 2), though a key difference exists. In terms of similarity, uncontextualized mentions can correspond to the three common types of entities addressed in entity linking, as outlined in Section 2.4: (1) Named mention, such as *the Meharry Medical School*; (2) Nominal mention, such as *the situation*; (3) Pronominal mention, such as *its* in *its debt problem*. Contextualization, like the Wikification task in entity linking, also focuses on identifying and resolving only important mentions, involving some selection mechanism to determine which mentions warrant further information [112]. The key difference between contextualization and entity linking lies in their output formats. Entity linking outputs a mapping, typically a link to a unique entity in a knowledge base. This format is well suited for web-based information access. In contrast, contextualization produces textual output, such as the clearer sentence in the earlier example, designed to integrate seamlessly into a conversational interaction. I refer to this output as *elaboration text*, and use *elaboration* both as an abbreviation and as a term for the process of generating it throughout this chapter.

Prior work on *elaborative simplification* [150] defines elaboration text as *"text in the form of definitions, explanations, or clarifications to improve readability by providing read-*

ers with necessary additional context." To better fit the needs of contextualization, I refine this definition: in this dissertation, an elaboration text is a clause that can be inserted into the original sentence to supply contextual information. The key requirement for contextualization is thus generating elaboration texts for uncontextualized mentions. While I work with sentence-level retrieved content that may contain multiple mentions, the focus is on each individual mention and its corresponding elaboration text, rather than the interplay between multiple mentions and elaboration texts within a sentence. This focus shapes both the experimental setup and the interpretation of results.

What information can an elaboration text include? A prior work, *Knowledge-enhanced Text Generation* (KETG) [176], suggests that two sources of information can be selectively integrated into the original sentence: context from preceding sentences within the same passage, or content from external sources. For nominal or pronominal mentions, the elaboration text typically draws from the preceding context, such as earlier sentences in the same paragraph of the original document. In contrast, for named mentions, elaboration generally requires information external to the document, often sourced from a knowledge base. Based on the origin of the information used to elaborate a mention, I distinguish between two types of uncontextualized mentions as below.

- **External mentions**, whose elaboration text relies on external resources such as knowledge bases.
- **Contextual mentions**, whose elaboration text relies on the preceding context of the sentence-level retrieved content.

This dichotomy assumes a clear separation between the two information sources and motivates system development in Section 5.2. However, some mentions may require elaboration that draws on both sources, as evidenced during human performance of the contextualization task (See Section 5.5.3 on "Nested Cases").

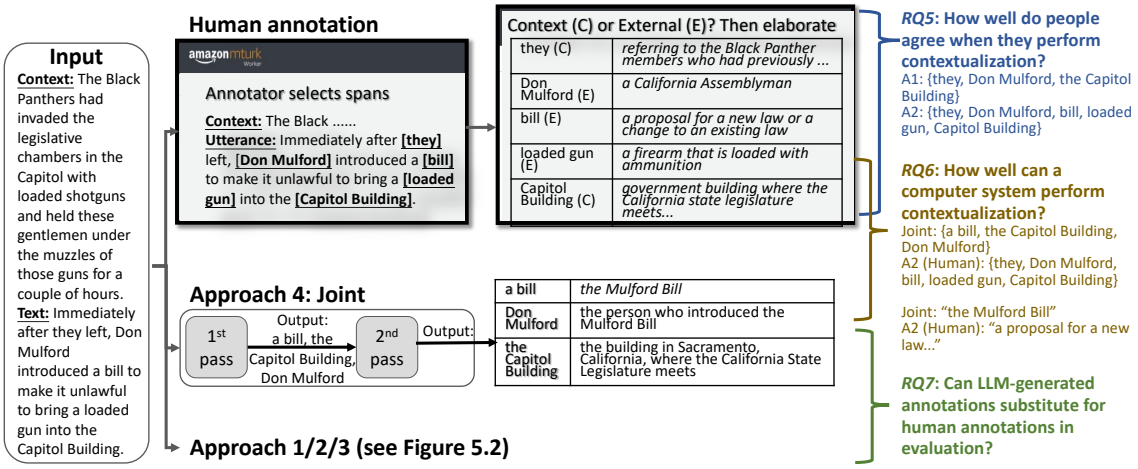


Figure 5.1: Illustrating the contextualization task, including the input, human annotation, proposed approaches, and research questions (RQs).

Related work has also designed categories for resolving important spans in text. The PLABA 2024 workshop [160] focuses on resolving complex terms in biomedical abstracts, and defines five distinct strategies for resolving them (substitution, explanation, generalization, exemplification, omission),[4] similar in spirit to my two-way categorization of mentions based on their respective elaboration strategies.

Thus far, the requirement of contextualization has been elaboration of uncontextualized mentions. In the next two sections, I formalize the overall process of elaborating mentions and propose four proposed approaches (implemented as six systems in total). Each involves one or more passes of few-shot prompting. The remainder of this chapter is structured by three key research questions:

- ***RQ5***: How well do people agree when they perform the contextualization task?
- ***RQ6***: How well can a computer system perform contextualization?
- ***RQ7***: Can LLM-generated annotations substitute for human annotations in evaluation?

[4] `https://bionlp.nlm.nih.gov/plaba2024/`.

Figure 5.1 illustrates the relevant parts of these RQs while Figure 5.2 provides more details on the proposed approaches, i.e., computer systems that perform contextualization. To answer *RQ5*, I design a human annotation study involving two locally recruited annotators using the Amazon Mechanical Turk platform. Different annotators exhibit individual differences, such as identifying overlapping but slightly different mentions. They show moderate agreement on whether a mention should be elaborated using external knowledge or the preceding context. To answer *RQ6*, I develop four approaches and conduct a human evaluation, which shows that most approaches are effective at identifying mentions and generating elaboration text. To complement the human evaluation, I also conduct automatic evaluation, using human annotations as the reference. One of the four approaches demonstrates competitive performance. Despite their strengths, all approaches tend to overlook contextual mentions that are crucial for conveying the intended meaning yet more challenging to capture, highlighting a key limitation. *RQ7* examines whether large language models (LLM) can serve as substitutes for human annotations for evaluation. Preliminary results suggest a potentially positive outcome, as LLM-generated annotations yield system rankings that are consistent with those produced using human annotations as reference.

The key contributions of this chapter are as follows:

- The introduction of a contextualization task aimed at improving the comprehension of historical texts in modern conversational interactions, elicited as requirements for elaborating uncontextualized mentions;
- The development of a human annotation protocol, including a tool implemented using Amazon Mechanical Turk, along with the resulting analysis of human performance on contextualization;
- A multi-faceted evaluation methodology consisting of human evaluation that provides ratings, automatic evaluation based on human annotations, and automatic eval-

uation based on LLM-generated annotations;

- A competitive LLM-based approach, referred to as the "Joint" approach, validated through both human ratings and automatic evaluation.

5.1 Problem Formulation

Let u denote the retrieved content, intended for use as an utterance in a conversational interaction. Since I focus on the sentence level, let u specifically represent a sentence. For example, it could be the output from the prototype retrieval system in Chapter 4, which considers it relevant to the current conversation. The original context that precedes u in the historical document is denoted as C, such as the preceding sentences from the same paragraph, in which case $C = \{c_1, c_2, ...\}$. The goal is to contextualize u into another piece of text u', where mentions are elaborated.

Further, I assume that u contains some set of n uncontextualized mentions whose elaboration would most effectively improve comprehension in modern conversational interactions. This set of mentions is denoted as $M = \{m_1, ..., m_n\}$. A model f identifies this set of mentions based only on u.

$$M = \{m_1, m_2, \ldots, m_n\}, \text{ where } M = f(u) \tag{5.1}$$

Each mention m_i is then elaborated by a second model g, which generates an elaboration text e_i with information that is not from u for the purpose of contextualization, where e_i is a clause that can be added to u. The information used by function g comes from either external resources such as a knowledge base KB, or the preceding context C of the retrieved content.

$$e_i = g(u, m_i, KB \cup C) \text{ for } i = 1, \ldots, n \tag{5.2}$$

The knowledge base, consisting of numerous entries, is represented as: $KB = \{kb_1, kb_2, ...\}$.

The set of mentions $M = \{m_1, ..., m_n\}$, and their corresponding elaboration texts are the critical building blocks for contextualization. The resulting contextualized sentence, denoted as u', for a specific mention m_i, is constructed by inserting the elaboration text e_i immediately after the first occurrence of m_i in u, with commas delimiting the insertion. For example, the elaboration *"who coached me and my teammate Alonzo"* is inserted directly after the mention *"Coach Mac"* in the earlier example. My analysis is conducted primarily at the level of individual mentions m_i and their corresponding contextualized sentence u'.

Finally, the overall goal of contextualizing u is to resolve multiple uncontextualized mentions by inserting elaboration texts for each. Such a fully contextualized sentence, the utterance for conversational interaction, denoted as u_n, can be obtained by sequentially inserting the elaboration texts for all the mentions into u using a function i. Although I do not implement this overall contextualization step, it can be formally defined as follows:

$$u_i = \begin{cases} i(u, m_1, e_1) & \text{if } i = 1 \\ i(u_{i-1}, m_i, e_i) & \text{if } i > 1 \end{cases} \tag{5.3}$$

5.2 Prompting LLMs for Contextualization

Four of my contextualization approaches utilize pre-trained LLMs to first identify mentions to be elaborated and three of the four use LLMs to generate elaboration texts. They all perform one to more passes of few-shot prompting to achieve this process. When there is more than one pass, it is a staged process. These approaches share similarities, so Figure 5.2 provides a general overview. To save space, the lower left legend shows some shared details for all gray boxes. The elements in the prompts were selected based on the generalized prompting instructions from Wang et al. [167], which is also referenced in the

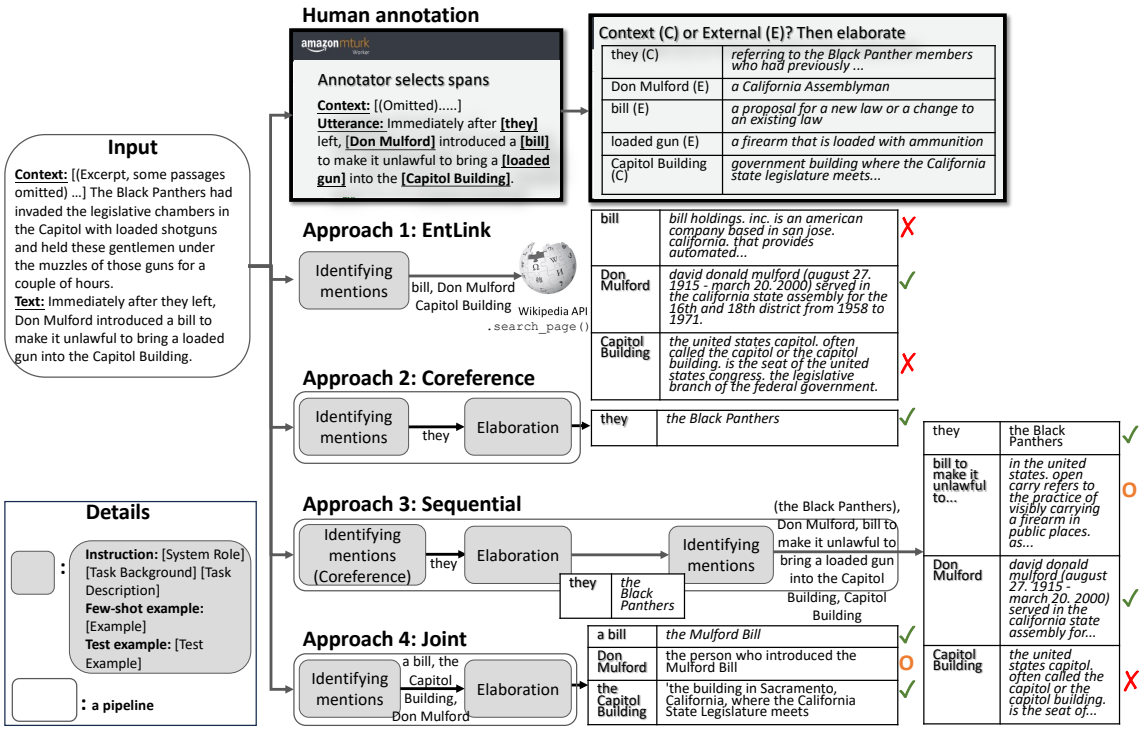


Figure 5.2: Illustrating the proposed approaches. Each uses one or several passes of few-shot prompting to identify mentions and then elaborate on them. ✓, x, and o show examples of elaborations as correct, incorrect, or not particularly useful or relevant, respectively, based on human evaluation. The elaboration for *Capitol Building* (bottom right) is incorrect, as it should refer to the one in California, based on the context.

style transfer chapter. For each approach, and specifically within each pass of few-shot prompting (each gray box), an instruction provides a plain-text definition of the task, followed by one or more few-shot examples. These examples contain elements such as the text, mentions, context, and explanation, etc. presented as desirable input and output. Table 5.1 organizes the elements whose variable contents are detailed in Tables 5.2 and 5.3. Each prompt ends with a field for supplying the actual input. To maintain consistency, the wording of each element reuses phrasing from the human annotation design in Section 5.5, which was iteratively refined for human readability. I now outline the unique features of each approach.

Approach 1: NER and Wikipedia elaboration ("EntLink") The first approach is a simple baseline that closely follows a standard entity linking process (precisely, "Wikifica-

tion" process [112], where Wikipedia is the knowledge base). It then treats the identified entities as the set of mentions and generates elaboration text for each, using only the corresponding definitions from the knowledge base. Aligning with the problem formulation, this baseline is designed with the simplifying assumption that mentions $M = \{m_1, m_2, ..., m_n\}$ are all External mentions, and no Contextual mentions are considered. Unlike Wikification, however, the inclusion of mentions does not consider their importance; any entity identified through named entity recognition (NER) is elaborated.

OpenAI defines an LLM use case of NER, for identifying some user-specified named entity types in a given text.[5] I use that use case to implement the model f that identifies the set of mentions. The prompting instruction focuses solely on entity linking, without any reference to the contextualization task. Then, I iterate over all mentions to generate elaboration texts using the following strategy (referred as the "Wikipedia elaboration strategy"), implemented as the model g. For each identified mention m_i (that is an entity), I use the Wikipedia Python library to retrieve the most relevant Wikipedia page kb_i based on the entity's textual form. I use the spaCy library to segment the page into individual sentences, and take the first sentence with the assumption that it contains the most essential definition of the mention m_i. For this approach, e_i is the first sentence from kb_i. In Figure 5.2, this approach performs only one pass of few-shot prompting.

Approach 2: Coreference tagging and contextual elaboration ("Coreference") The second approach, another baseline, aims to maximally differ from an NER process in Approach 1. Recall that Approach 1 assumes that mentions $M = \{m_1, m_2, ..., m_n\}$ are all External mentions. My second approach therefore performs contextualization by focusing on the text u and the preceding context C, with no explicit reliance on external resources. It performs two passes of few-shot prompting. In the first pass, the prompting instruction

[5]The entity types are person, organizations, geopolitical and other locations, events, work of art, and law. Examples for each from the official documentation is Johannes Gutenberg, The Beatles, Germany, Renaissance, tort liability.

introduces the background of contextualization as conversational interaction with general museum visitors, defines what constitutes a Contextual mention (see Table 5.1[6]), and then uses several examples to illustrate several Contextual mentions. The prompting instruction then takes as input the text u to be contextualized, and identifies a set of Contextual mentions. That set of mentions is processed in a second pass of few-shot prompting. The prompting instruction describes the task of elaboration for that mention, and explicitly requires the use of contextual information. It includes two examples from the training split to demonstrate valid elaboration texts, and concludes by taking the actual context, text, and mention as input to produce an elaboration text e_i.

Approach 3: Sequential combination ("Sequential") My third approach applies the first two approaches in a sequence and therefore involves three passes of few-shot prompting. It begins with coreference tagging and contextual elaboration. Then, it inserts those elaboration texts back to the original text u, creating a partially contextualized text. A third pass of few-shot prompting, i.e., instructions to perform NER, followed by the Approach 1's Wikipedia elaboration strategy, is applied to this partially contextualized text. The decision to perform coreference tagging and contextual elaboration (Approach 2 "Coreference") before NER and Wikipedia elaboration (Approach 1 "EntLink") is a deliberate choice. However, it raises a question: which sequence yields a better combined effect on contextualization—Coreference first or EntLink first? The design choice in Approach 3 is guided by two considerations: (1) Deciding which mentions to tag during coreference tagging as in Approach 2 is generally more intuitive than setting user-specified named entity types as in Approach 1. (2) In the event that a mention qualifies as both an External and a Contextual mention, I expect that the preceding context is more likely provide accurate information for elaboration, whereas Wikipedia-based elaborations may be less reliable due to potential NER errors or the possibility that the first sentence of a Wikipedia entry is not

[6]The prompting instruction refers to it as a Coreference mention.

# Pass	**Prompting Elements**	**EntLink**	**Coreference**	**Sequential**	**Joint**
First pass	System Role	You are an expert in Natural Language Processing.	`Whom_talking_to` (Texts in typewriter font represent variables, whose full contents are listed in the next two tables. This convention applies throughout this table.)		
	Task Background	N/A	`Two_types_mentions`		
	Task Description	Your task is to identify **common Named Entities (NER) in a given text**. The possible common Named Entities (NER) types are exclusively: *person, organizations, geopolitical entities, other locations, events, work of art, and law*.	Your task is to identify **particularly the coreference mentions** from the original text that need further explanation for a general audience (with high school education). `Task_elab_choice_delimiter`		Your task is to identify **all the mentions** from the original text that need further explanation for a general audience (with high school education). `Task_elab_choice_delimiter`
	Few-shot Examples	`Example_ner`	`Example_coreference`		`Example_joint`
Second pass	System Role	N/A	`Whom_talking_to`		
	Task Background	N/A	`Two_types_mentions`		
	Task Description	N/A	Your task is to provide an explanation **to the coreference mention from** the original utterance. If this mention is a coreference mention, your explanation shall use information from context. Conversely, if this mention is an external mention, your explanation shall use world knowledge. `Task_elab_text_delimiter`		Your task is to provide an explanation **to the mention from** the original utterance. If this mention is a coreference mention, your explanation shall use information from context. Conversely, if this mention is an external mention, your explanation shall use world knowledge. `Task_elab_text_delimiter`
	Few-shot Examples	N/A	`Example_coreference_elab`		`Example_joint_elab`
Third pass	System Role	N/A	N/A	You are an expert in Natural Language Processing.	N/A
	Task Background	N/A	N/A	N/A	N/A
	Task Description	N/A	N/A	Your task is to identify common Named Entities (NER) in a given text. The possible common Named Entities (NER) types are exclusively: *person, organizations, geopolitical entities, other locations, events, work of art, and law*.	N/A
	Few-shot Examples	N/A	N/A	`Example_ner`	N/A

Table 5.1: Prompting elements for each approach. Texts in typewriter font represent variables containing long texts, whose full contents are in the next two tables (5.2 and 5.3). Bold text highlights the differences in the same element across approaches.

Variable	Content
`Whom_talking_to`	Imagine that you are a conversational agent emulating the historical figure of Ronald Reagan, the 40-th President of the United States. You are talking to a general audience from nowadays (the 2020s), specifically, a young adult at the high school education level. This person may ask for clarification on certain parts of what you said. You will explain these parts further based on some conversation context and the knowledge of a general audience at the high school education level.
`Two_types_mentions`	These parts of the conversation are called mentions, which come in two types: coreference mentions, explained through other expressions from or based on conversation context that refers to the same person or thing, and external mentions, explained with world knowledge. Mentions are minimum units that need explanation. A coreference mention is an expression that refers to an earlier mentioned person or thing within a conversation context. A coreference mention can be a pronoun, a possessive determiner, a name, a description, or any expression sharing a common referent. An external mention shall be explained with world knowledge, whose meaning remains constant regardless of the conversation context. External mentions often involve specific knowledge about the world, such as historical events, or cultural references that are not defined solely by the context of the conversation. Additionally, if a mention contains complex vocabulary that may be difficult for a general audience (with high school education) to understand, it is more likely to be an external mention. Conversely, if a mention needs explanation even if the vocabulary is simple for a high school education level, the mention is more likely to be a coreference mention.
`Task_elab_choice_delimiter`	You should answer with the exact text from the original text. Do not write sentences. If there are several mentions in your answer, only list these mentions as-is, format them as a list, separate them with commas, and do not include spaces, periods or duplicates. For example: mention1,mention2,mention3 If you think that there is no mention that needs further explanation for a general audience (with high school education), you should write "There are no mentions." instead. In other words, if there are no such parts, do not write anything else, but only "There are no mentions."
`Example_ner`	Text: "In Germany, in 1440, goldsmith Johannes Gutenberg invented the movable-type printing press. Modelled on the design of the existing screw presses, a single Renaissance movable-type printing press could produce up to 3,600 pages per workday." Output: {"gpe": ["Germany"], "date": ["1440"], "person": ["Johannes Gutenberg"], "product": ["movable-type printing press"], "event": ["Renaissance"], "quantity": ["3,600 pages"], "time": ["workday"]}
`Example_coreference`	Original text: The old memories of "gunboat diplomacy" are still with them. The answer is: Them Original text: There can be no question that we must not minimize in any way the threat to the free world by the Soviet Union. The answer is: the free world Original text: The most obvious is that the program was created with no provision for a means test. The answer is: the program Original Text: Certainly he was most helpful in our request that HEW permit us to require welfare recipients to perform useful community service in return for their grants. The answer is: he,our Original text: For many years virtually since the inception of public schools nonsectarian prayers were offered in schools. The answer is: There are no mentions.

Table 5.2: Long text contents of the variables from Table 5.1.

Variable	Content
Example_joint	Original text: The old memories of "gunboat diplomacy" are still with them. The answer is: "gunboat diplomacy",them Original text: There can be no question that we must not minimize in any way the threat to the free world by the Soviet Union. The answer is: the free world,the Soviet Union Original text: The most obvious is that the program was created with no provision for a means test. The answer is: the program,means test Original text: Certainly he was most helpful in our request that HEW permit us to require welfare recipients to perform useful community service in return for their grants. The answer is: he,our,HEW Original text: For many years virtually since the inception of public schools nonsectarian prayers were offered in schools. The answer is: There are no mentions.
Task_elab_text_delimiter	You should write your explanation text as a clause that can be inserted to the original text. Do not write a full sentence. In other words, only return the clause without any beginning or ending punctuation.
Example_coreference_elab	Context: With our attention to Iran and the Middle East, most Americans are unaware of what has been going on in the Caribbean islands. Original text: We are being ringed in there by islands, which, one after the other, have come under the influence of the Soviet Union by way of Castro. Mention: there The answer is: the Caribbean islands Context: I said during the campaign that we would do nothing to hurt those presently dependent on Social Security checks that we would not pull the rug out from under those people so dependent. Original text: I did say that I would try to restore the integrity to the program. Mention: the program The answer is: the Social Security program
Example_joint_elab	Context: I did say that I would try to restore the integrity to the program. Original text: I did say that I would try to restore the integrity to the program. Mention: the program Mention type: coreference What is your explanation to this mention? The answer is: the Social Security program Context: Thank you too for your prayers; you will be in mine. Original text: I promise too that we will continue trying to help peace come to El Salvador. Mention: El Salvador Mention type: external What is your explanation to this mention? The answer is: a country in Central America which is bordered on the northeast by Honduras, on the northwest by Guatemala, and on the south by the Pacific Ocean

Table 5.3: Long text contents of the variables from Table 5.1 (continued).

a good elaboration for the mention. For example, the Wikipedia content associated with a Wikipedia entity called *"Open carry in the United States"* shown in Figure 5.2 does not directly explain not the NER-identified mention in *u "bill to make it unlawful to bring a loaded gun into the Capitol Building."*

Approach 4: Joint contextualization ("Joint") Approaches 1 and 2 are specifically designed for the contextualization task. Approach 3 combines the two by applying them in sequence. Approach 4 differs from and improves upon Approach 1 by replacing the NER process and Wikipedia-based elaboration with prompts specifically tailored to contextualization. Recall that in Approach 2, the prompting instruction defines what constitutes a Contextual mention and that is followed by another instruction to generate an elaboration text based on the preceding context. Approach 4 extends Approach 2 by instructing the LLM to identify Contextual and External mentions, with prompting examples aligned to that instruction. It is followed by another instruction to elaborate these mentions. As a result, the prompting instruction of Approach 4 ("Joint") is detailed and explanatory, closely resembling an annotation guideline for human annotators to perform contextualization.

Implementation Details All four approaches use the the OpenAI GPT-4o model (version 2024-05-13). For the Joint approach, I also implement it using the Claude 3 Sonnet and Haiku models by Anthropic.[7] Different implementations of the same approach are referred to as different systems in the experiments. The model version used for each Joint system is indicated in all result tables.

[7]Anthropic is an AI research company that, in March 2024, claimed that one of its Claude 3 models outperformed both GPT-4 and Gemini in mathematical problem solving, coding, general knowledge and other areas.

5.3 Corpora

This chapter empirically focuses on applying contextualization to letters. The Reagan collection described in Chapter 4 serves as the primary source for evaluating contextualization. An updated portion of the collection—letters—was drawn from the book *Reagan: A Life in Letters* (more details in Section 6.3), which includes approximately 1,058 letters written by Ronald Reagan over his lifetime. Letters stand out as written monologues, in contrast to spoken-language transcripts such as public papers and interviews. They are hypothetically valuable for contextualization due to their brevity and the fact that they were originally addressed to someone other than the intended user of our system.

In more details, I extract individual letters by segmenting the entire text from the book using salutations as starting points and various forms of Reagan's signature as ending boundaries. After obtaining individual letters, I further segment the letters into paragraphs, where a paragraph is usually cohesive on one topic and has one or more sentences. I use the spaCy library to segment paragraphs into sentences, where a sentence is the input text u for contextualization. Each sentence can be traced back to its context, which in my experiments are preceding sentences from the same paragraph. The segmented letter paragraphs, and letter sentences are later updated into the Reagan collection from Chapter 4. The data are split at the letter level, with all corresponding sentences grouped into training, validation, and test sets containing 1,768, 395, and 360 sentences, respectively. I use the training split to manually select prompting examples for the proposed approaches, and apply proposed approaches to the validation and test splits.

5.4 Human Evaluation

At the highest level, the evaluation of contextualization is divided into human evaluation and automatic evaluation. This section presents the human evaluation design and results. Human evaluation provides direct insights into how people perceive the effectiveness of contextualization (***RQ6***), i.e., *"a direct evaluation of the utility of proposed approaches"* according to Reiter [132]. It is conducted directly on system outputs. A human provides preferences based on contextualization requirements, divided into two aspects: identifying mentions that should be elaborated, and generating elaborative texts for those mentions.

5.4.1 Design

One evaluator (referred to as A1) was recruited to provide their judgments to a random sample of contextualization output. A1 is a native English-speaking undergraduate majoring in Information Science at a U.S. university, aged between 18 and 25. Initially, A1 and a second annotator with the same background (referred to as A2) were recruited to perform a human annotation task (see procedures in Section 5.5). Based on A1's demonstrated efficiency and ability to produce sufficient and meaningful work, A1 were subsequently engaged in the human evaluation task described in this section. The results of human evaluation and human annotation are presented in reverse chronological order to align with the writing structure, starting with human evaluation, which serves as a reference point for the overall automatic evaluation.

The test split of the corpus contains 360 sentences. I randomly sample 50 mentions identified by each system from those sentences, and collect their corresponding elaboration text. As a result, the selected mentions are not paired across systems—they represent independent data points for statistical analysis. A total of 300 mentions and elaboration texts, along with the original texts and preceding contexts, are shuffled and displayed as 300 rows

(a) Labels for the 5-point rating scale used for mention identification. When training the evaluator, the wording of these labels slightly differed in verb choice—for example, *"Explaining this mention is critically important"* was used, which conveys the same intent as *"Identifying this mention is critically important for contextualization,"* the phrasing adopted here for clarity.

Value	Label
5	Identifying this mention is **critically important** for contextualization.
4	Identifying this mention is **highly desirable** for contextualization.
3	Identifying this mention **has some added value** for contextualization.
2	Identifying this mention **does not add value** for contextualization.
1	Identifying this mention is **counterproductive** for contextualization.

(b) Labels for the 5-point rating scale used to evaluate elaboration texts. When training the evaluator, the wording such as "Explanation to this mention is excellent" was used.

Value	Label
5	The elaboration text for this mention is **excellent**.
4	The elaboration text for this mention is **correct and of good quality but has minor flaws**, such as being slightly too long or too short, not perfectly formatted, or containing incorrectly encoded non-English characters.
3	The elaboration text for this mention is **correct but not particularly useful or relevant**.
2	The elaboration text for this mention may **contain incorrect information**.
1	The elaboration text for this mention is **highly inaccurate to the point of being misleading**.

Table 5.4: Likert scales for the two aspects: mention identification and elaboration texts.

in an offline Excel spreadsheet for human evaluation, where the name of the system that identify each mention is shielded. The evaluator provided preferences based on contextualization requirements, divided into two aspects, using a 5-point Likert scale: mention

Table 5.5: The median and mean of ratings, and mean of binary values for each system, evaluated on a random sample of 50 positive mentions from the test split.

System	Identifying Mentions			Elaboration Texts		
	Median	Mean	Mean (Binary)	Median	Mean	Mean (Binary)
Joint (Sonnet)	**5.0**	4.14	**0.88**	**4.0**	**3.96**	**0.96**
Joint (Haiku)	4.0	3.82	0.82	**4.0**	3.92	0.94
Joint (GPT)	4.0	3.82	**0.88**	**4.0**	3.64	0.90
EntLink	3.0	3.36	0.80	3.0	2.80	0.68
Coreference	**5.0**	**4.22**	0.84	**4.0**	3.66	0.94
Sequential	4.0	3.82	0.84	**4.0**	3.40	0.86

identification and elaboration quality. Table 5.4 defines the scales and their labels.[8]

5.4.2 Results

Table 5.5 reports the median and mean ratings assigned by the human evaluator A1 for each system and for each aspect. The median ratings are reported to align with the nature of ordinal rating scales, which do not assume evenly spaced numerical intervals. However, median ratings often result in ties and provide limited differentiation among systems—though they still help identify underperformers, such as EntLink. The following analysis is based on mean ratings.

For identifying mentions, all mean ratings fall within the range 3.3 to 4.2, indicating that the mentions are generally perceived as having some added value. The mean ratings rank the systems as follows: Coreference first; Joint (Sonnet) second; a tie for third place among Joint (Haiku), Joint (GPT), and Sequential; and finally EntLink.

Across all mentions identified, the human evaluator A1 assigned the following number of mentions to each point on the scale, from lowest to highest: 0 as counterproductive, 47 as having no value, 64 as having some added value, 72 as highly desirable, and 117 as critically important. This means that the evaluator considered 84% of mentions identified by any system as having some added value. Table 5.6 shows a few example mentions identified by various systems, corresponding to each value assigned by the evaluator. The evaluator demonstrated a strong preference for Contextual Mentions to be rated as critically important—examples include *"those," "They," "his,"* and *"the same decision."* By contrast, External Mentions, such as *"The Constitution," "Bercy Sports Stadium,"* and *"Senator Chavez"* are generally rated as having some added value.[9]

For elaboration texts, all average ratings, except that of the EntLink system, fall within

[8]The scales are ordinal, not interval [5].

[9]This classification of Contextual and External Mentions was by me based on the definitions the two.

(a) Mentions that are sampled from each level of the human evaluator A1's ratings.

Value	Example mentions
5	those, they, they, here, the woman, us, He's, They, They, bugging
4	Our house, people like yourselves, draft48, USSR, Senate committee, Angola, Reverend Billy Graham, Pentecostals
3	president, Bercy Sports Stadium, the ranch, Washington, law, Los Angeles, Senator Chavez, Washington, Mexican workers, The Constitution
2	fire protection, his, country, prestigious, last New Year's, that office, friends, both of us, nuclear missiles, the atrocities
1	N/A

(b) Elaboration texts sampled from each level of A1's ratings.

Value	Example elaborations (Bulleted, with the corresponding mention in parentheses)
5	- (these gentlemen) the members of the legislative chambers, - (maturing) becoming older and developing both physically and mentally, as we are discussing in this birthday context, - (studio cooperation) access to and assistance from Hollywood studios for reporting and publicity purposes
4	- (the convention) the Republican National Convention, a formal gathering where the Republican Party officially nominates its candidate for president and establishes its platform for the upcoming election - (Israel) a country in the Middle East, located on the eastern shore of the Mediterranean Sea, known for its historical and religious significance
3	- (Ireland) a country in Europe located to the northwest of continental Europe, occupying most of the island of Ireland - (Dublin) dublin is the capital city of ireland. - (Cuban military) the cuban revolutionary armed forces (spanish: fuerzas armadas revolucionarias; far) are the military forces of cuba.
2	- (his own country) the country of the person being referred to in the conversation, which in this context would likely be the United States - (those on our own side) the American soldiers and their allies - (our people) the American citizens who are capable of self-government and whose needs and wants should be prioritized
1	- (friends) friends is an american television sitcom created by david crane and marta kauffman. which aired on nbc from september 22. 1994. to may 6. 2004. lasting ten seasons. - (him) him (sometimes stylized as h.i.m.) was a finnish gothic rock band from helsinki. - (he) he... is a 2011 bhojpuri film directed by mangesh joshi starring madan deodhar. rajendra gupta and shashank shende.

Table 5.6: Examples sampled from each value of the human evaluator's ratings.

the range of 3.4 to 4.0, indicating generally accurate elaborations. Four systems average above 3.6. The average ratings rank the systems as follows: Joint (Sonnet), Joint (Haiku), Coreference, Joint (GPT), Sequential, and EntLink.

Across all elaboration texts, the evaluator rated 23 of them as highly inaccurate, 13 as potentially containing incorrect information, 66 as not particularly useful, 168 as in high quality while having minor flaws, and 30 as excellent. This means that the evaluator considered 66% of elaboration texts as having high quality. Table 5.6 shows a few elaboration texts generated by various systems, corresponding to each label point as perceived by the evaluator. The evaluator demonstrated a preference for medium-length elaboration texts: those rated as excellent (5) average 75 characters, while those rated as having minor flaws (4) average 102 characters.

Research suggests that 5-point human ratings are subject to cognitive uncertainty [50], so I also convert the ratings for identifying mentions into binary values. Mentions rated 3 or higher are assigned binary value 1, signifying at least some added value, while those rated below 3 are assigned binary value 0, indicating a lack of added value. Similarly, I convert the ratings for elaboration into binary values: mentions rated 3 or higher are assigned binary value 1, signifying a correct elaboration, while those rated below 3 are assigned binary value 0, indicating incorrect elaboration. The mean of binary values are also reported in Table 5.5.

The binary values for identified mentions averages between 0.8 and 0.9, indicating limited variance across systems. The binary values also yield a different ranking: a tie for first place between Joint (Sonnet) and Joint (GPT); then a tie for third place between Coreference and Sequential; followed by Joint (Haiku); and finally EntLink. The rank shift for Approach 2 (Coreference) between the mean and binary value rankings may suggest Coreference's unique strength in identifying mentions that A1 considers critically important.

All binary values for elaboration texts (except that of EntLink) average over 0.85, indi-

cating both high overall accuracy and limited variance across systems in elaboration correctness. The EntLink system is particularly prone to elaborating with incorrect information, as it often links mentions to incorrect knowledge base entities that share the same surface form, and struggles to accurately elaborate Contextual mentions. For example, the pronoun *"he"* should be a Contextual mention but it is mistakenly elaborated as a film titled *"He... (2011)."*

To summarize the findings for ***RQ6***, human evaluation results indicate that five out of six systems generally perform contextualization fairly well: they usually identify mentions with at least some added value and they generate elaboration texts that tends toward good quality.

5.5 Human Annotation

Human annotations play two important roles in contextualization. First, researchers collect human annotations from experts in pilot studies, such as auxiliary labels describing textual characteristics, to better understand the task [150]. Human annotations offer insights into the relevant "linguistic knowledge" [150], as reflected by the degree of agreement among annotators. The degree of agreement is the focus of ***RQ5***. Second, many text rewriting tasks can benefit from some available human-written ground-truth texts that work as "references" for conducting automatic evaluation (***RQ6***) [35, 150]. The references are often a naturally-emerging human-written corpus. For example, in the open-ended story generation task, Chiang et al. [35] worked with a dataset that consists of stories that Reddit users wrote spontaneously to a sub-reddit named *WritingPrompts*. In the absence of a naturally occurring human-written corpus about contextualization, creating one becomes important. Regarding terminology, the term *human annotations* is often used broadly to refer to human-written ground-truth outputs in various forms, including text [156]. In the

context of contextualization, identifying mentions, as done by the proposed approaches, has a non-textual form of output, whereas elaboration texts are textual output. Therefore, this section describes the human annotation process, including identifying mentions and elaboration, and present the results, which constitute a contribution of this chapter.

5.5.1 Design

In this human annotation process, local (not crowdsourced) annotators performed contextualization by identifying mentions and writing elaboration texts, using a user interface (UI) on Amazon Mechanical Turk (MTurk) Worker Sandbox.

Tools and Annotators: Regarding annotator recruitment, previous studies [150] have cautioned against relying on the general MTurk worker pool for cognitively demanding tasks, highlighting risks such as high performance variability, poor calibration, and the potential for misleading conclusions. An alternative approach is to use the Amazon Mechanical Turk Worker Sandbox, a simulated environment with most features of the MTurk marketplace but designed for internal development where locally selected workers can perform the task. The third approach is Turkle [92][10] is a Django-based web application that is an open-source alternative to MTurk. I recruited annotators through community channels and instructed them to perform the annotation tasks in the Worker Sandbox. However, the ideal annotators for the contextualization task would be individuals with historical expertise or deep familiarity with the Reagan collection. The annotation team consists of two annotators who are native English-speaking undergraduate majoring in Information Science at a U.S. university, aged between 18 and 25. They responded to a recruitment post and completed a pre-screening and training session remotely via Zoom. The first annotator (A1) subsequently served as the evaluator for the human evaluation as described in

[10] `https://turkle.readthedocs.io`.

Section 5.4.1. The second annotator (A2) was recruited only for this section.

Training of Annotators: During the training session, the annotators were introduced to an instruction manual and example annotations, and they also received a copy of the training manual for subsequent reference. The full training manual is in Appendix D. It outlines the contextualization task, then provides a concise set of decision rules in natural language, while still encouraging annotators to exercise discretion on a case-by-case basis when working on each example. Each text that needs contextualization is shown to an annotator as a Human Intelligence Task (HIT) in MTurk, i.e., a single, self-contained task to be completed online. The annotation interface for a HIT is shown in Figure 5.3. In brief, the annotators were instructed to identify mentions based on what they expect a general audience with at least high school education would not understand. The annotators first read the test, and selected text spans as their choice of the set of mentions for contextualization, by dragging their mouse. For each such mention, a corresponding group of questions appeared below the text. Within each question group, the annotators classified the corresponding mention as either External or Contextual, and then typed an elaboration text.

Main Annotation Process: Annotators completed HITs constructed from texts in the validation and test splits of the contextualization corpora (Section 5.3). Throughout the annotation process, each annotator was assigned HITs in sequential batches of increasing size (e.g., 25, 50, 100), continuing until all HITs are completed. Annotators completed batches asynchronously and were compensated at a rate of $20 per hour for their focused time. After annotators completed a batch, I reviewed their annotations and asked questions about their thought process to ensure a basic but sufficient level of adherence to the instructions.

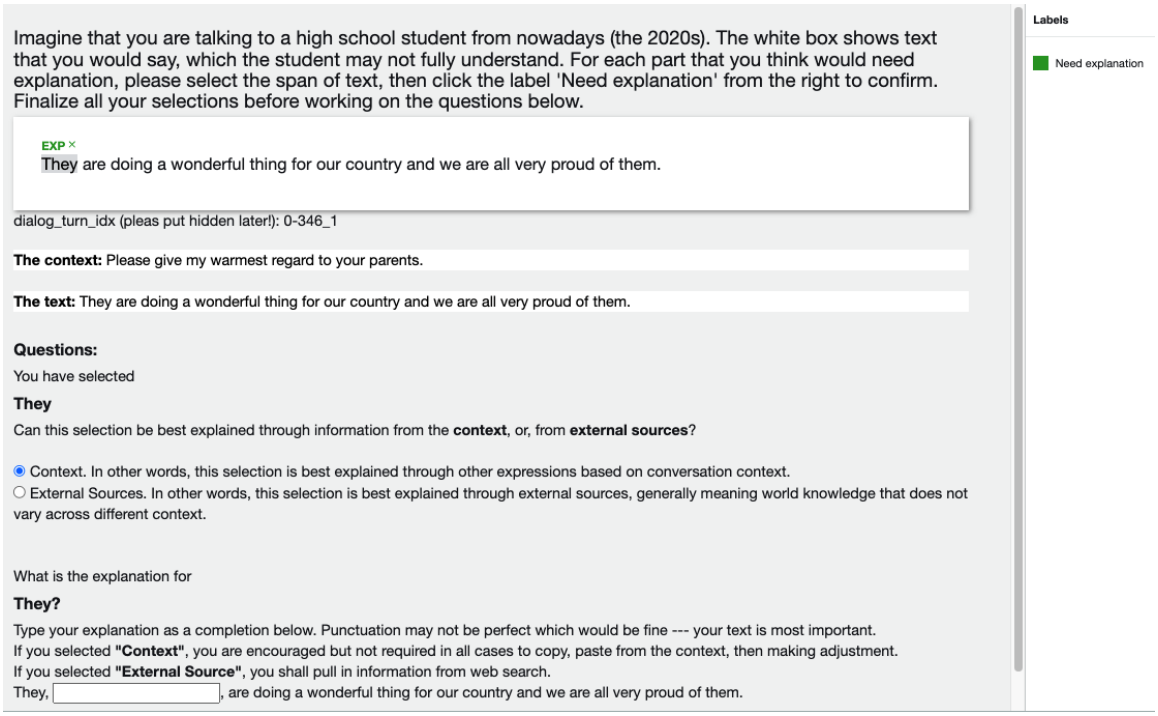


Figure 5.3: Human annotation interface on Amazon Mechanical Turk (MTurk). The annotators first identified a set of mentions for contextualization, then wrote elaboration texts to contextualize each mention. The wording used in the interface differs slightly from that presented in the dissertation.

5.5.2 Fuzzy Matching Aligns Differing Human Annotations

When multiple annotators have worked on HITs based on the same text, their annotations often differ, allowing analysis of individual differences. These differences are welcomed: the task requires careful reading, thoughtful reasoning, and often yields diverse interpretations of the text. Doing analysis on individual differences fulfills the first role of human annotations: capturing relevant "linguistic knowledge" (***RQ5***). Regarding the second role—supporting automatic evaluation (***RQ6***)—other NLP tasks that collect human annotations often aggregate numerical judgments from multiple annotators to mitigate the effect of individual subjectivity [150]. Since annotations for contextualization are structured texts rather than numerical values, there is no standard aggregation method to process

them. I therefore choose one annotator's annotations as the primary annotations for all texts in the automatic evaluation (Section 5.6).

What are the differences specifically? Different annotators may identify overlapping sets of mentions, but these sets often differ in minor details. For example, in the running example shown in Figure 5.1, one annotator selected the spans: *"they," "Don Mulford," "bill," "loaded gun,"* and *"Capitol Building."* whereas another annotator selected *"they," "Don Mulford,"* and *"the Capitol Building." "Capitol Building"* and *"the Capitol Building"* do not exactly match, but refer to essentially the same mention. I treat such cases as reflecting the annotators' shared intent and thus as agreement.

To measure agreement without overly penalizing minor variations from exact match, I use a fuzzy matching strategy that aligns mentions based on the following two criteria: (1) they differ only by the presence of *"the;"* or (2) they share overlapping non-stop words and non-punctuation tokens, after splitting the text by whitespace. When two mentions are aligned, both are normalized to the shorter form. Fuzzy match also helps increase the number of matched elaboration texts. For instance, one annotator identified *"mandated"* with the elaboration text *"required or enforced by law or authority"* while another annotator identified *"mandated prayer"* with the elaboration text *"a government-enforced religious practice."* The similarity between those elaboration texts should be measured rather than completely ignored due to minor variations in the underlying mentions.

However, there are still corner cases where fuzzy matching fails, resulting in false matches. For example, the matching may be suboptimal depending on how two sets of mentions are ordered. If the set {..., *House Ways and Means Committee,* ...} is aligned with the set {..., *House, Ways and Means Committee,* ...}, a longer mention like *"House Ways and Means Committee"* might be matched with *"House"*, missing a more appropriate pairing such as *"Ways and Means Committee"* if based on the comparative length. A potential improvement to fuzzy matching is to apply a more fine-grained threshold by Dice's

coefficient [46], which measures string similarity by computing the character-level n-gram overlap between two strings.

5.5.3 Results

Amount of Work, Timeline, and Compensation: The two annotators, identified as A1 and A2, each annotated the same set of 395 HITs (one sentence per HIT) from the validation split and 360 HITs from the test split. A1 identified 801 mentions in the validation split and 510 in the test splits, while A2 identified 1,042 and 845 mentions, respectively. The annotation was conducted between late October 2024 and February 2025, with a total compensation ranging from $600 to $1000 for approximately 30 to 50 hours of work.

Inter-annotator Agreement: Inter-annotator agreement measures agreement in decisions by the two annotators throughout the annotation process. When two annotators identify different sets of mentions, inter-annotator agreement cannot be reliably calculated at the mention-level due to the unknown probability of negative decisions (i.e., choosing not to identify a particular mention). Recall that in the annotation interface (see Figure 5.3), a pop-up question prompts them to decide whether the mention is External or Contextual. I apply the fuzzy matching strategy to the two sets of mentions to obtain a joint set of aligned mentions, and compute Cohen's kappa [36] on the External vs. Contextual labels over the joint set. Cohen's kappa between annotators A1 and A2 is 0.641 based on 642 aligned mentions from all HITs in the validation split, and 0.707 based on 364 aligned mentions from the test split. The value 0.641 is often interpreted as indicating moderate to substantial agreement [8, 90]. The value 0.707 is generally regarded as substantial agreement. In response to ***RQ5***, human annotators demonstrated a strong shared understanding of the contextualization task: they identified overlapping mentions, and consistently assigned the same category (i.e., elaboration type) to those mentions. I defer the assessment of agree-

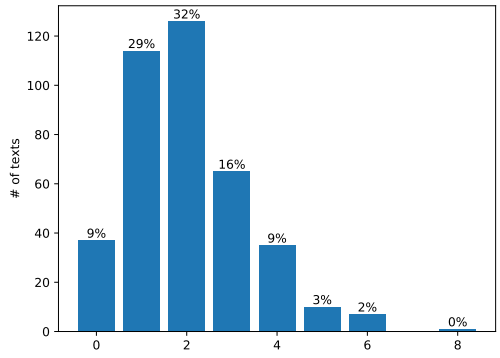
ment on elaboration texts until the appropriate evaluation measures are introduced in the automatic evaluation section.

Nested Cases: During intermediate feedback conversations with the annotators, a small proportion of mentions (in fewer than 10% of texts) emerge as *nested cases*, whose category falls between External and Contextual—often best understood as a mixture of both. These mentions would negatively impact the inter-annotator agreement as reported. While these cases can generally be elaborated with contextual information, a more informative elaboration often requires additional information from external sources. For example, in one text *"They would be executed if they tried to live as citizens under the Sandinistas."*, the mention *"They"* can be elaborated as *"some of the Contras."* if it's a Contextual mention, but the elaboration text is not fully informative. Information from external sources can further clarify that *the Contras* means *"various right-wing rebel groups active from 1979 to 1990..."*

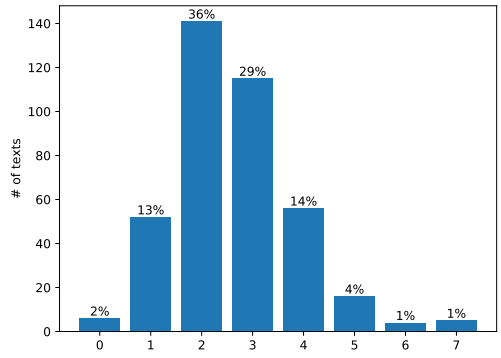
Individual Differences: Figure 5.4 illustrates the distribution of mentions per text and the number of characters per mention for both annotators. On the validation split, the average number of mentions by A1 is 2.03; for A2, the average is 2.63. While A1 and A2 differ in the number of mentions they extract per text, they have similar character lengths per mention. The average length of mentions by A1 is 9.95 characters, with an average word count of 1.49 words. For A2, the average length is 10.86 characters, with an average word count of 1.69 words. According to A1 and A2, 9% and 2% texts, respectively, contain no mentions requiring contextualization. These are typically simple, self-contained statements on some issue or event.[11]

To better understand the mentions identified by A1 and A2, I divide the 395 texts from the validation split into three groups based on the relative difference in the number of

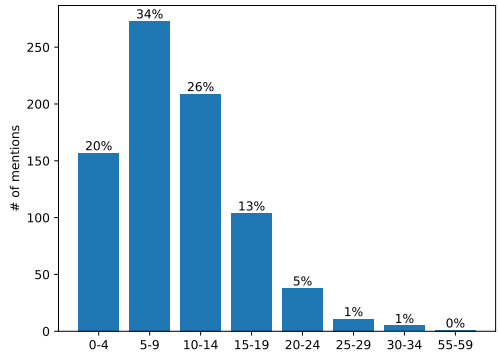
[11] Such as one statement on journalism practice, *"Would it have been fair to let one reporter have an added question without giving others the same opportunity?"* or another statement on prayer freedom, *"Anyone was free to pray if he or she so desired."*



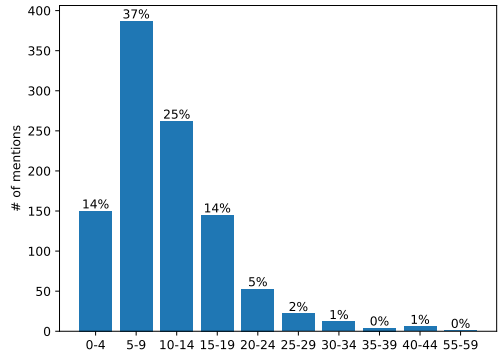
(a) # of mentions per text by A1.



(b) # of mentions per text by A2.



(c) # of characters per mention by A1.



(d) # of characters per mention by A2.

Figure 5.4: Histograms showing mentions per text and character count per mention for A1 and A2 on the validation split.

mentions identified (0-20%, 40-60%, 80-100%). Due to the skewed distribution, these ranges roughly form three categories, described as: A1 identifies at least as many mentions as A2; A2 identifies as many or slightly more than A1; and A2 identifies at least two more mentions than A1. I then randomly sample 5 texts (sentences) from the first and last categories, and compare the mentions extracted in Table 5.7.

For those texts, there is a high overlap in mentions extracted, while two annotators show different characteristics. In Table 5.7(a), unique mentions by A1 include *that situation, falsehood, criticism, advocating, mandated,* and *economic reforms,* with elabora-

Utterance	Mentions and Elaborations by A1	Mentions and Elaborations by A2
Finally; as for Lebanon, that situation is only one facet of the whole Middle East problem.	Lebanon: a country in the Levant region of West Asia **that situation**: the specific issues or conflict occurring in Lebanon at the time Middle East problem: the complex and ongoing political, social, and military issues in the Middle East region	Lebanon: country in the Middle East **facet**: or aspect Middle East problem: referring to the broader set of issues in the Middle East, which could include political instability, territorial disputes, religious conflicts
And of course as always there are some deluded souls who believe the falsehoods of the lynchers and go along with the mob.	**deluded**: misled or deceived into believing something that is not true **falsehoods**: lies or statements that are not true lynchers: people attacking someone's reputation or character **mob**: group of people acting emotionally and often irrationally	**deluded souls**: individuals who are misled or deceived by false information or manipulations, often lacking critical judgment or awareness lynchers: people who engage in a mob or vigilante act of punishment, often unjust or illegal **the mob**: the group of people who collectively follow or support an action or belief, often in an unthinking or chaotic manner, typically referring to a crowd that seeks justice outside of legal or proper channels.
There have been criticisms from time to time by some in that process but still we have a good relationship.	**criticisms**: expressions of disapproval or pointing out faults that process: meaning the Contadora process	that process: the Contadora process, a diplomatic effort aimed at achieving peace and resolving conflicts in Central America, particularly involving countries like Nicaragua, El Salvador, and Honduras
Its opponents made the argument that we were advocating mandated prayer.	Its: refers to the school prayer amendment **advocating**: publicly supporting or arguing in favor of something **mandated**: required or enforced by law or authority	Its: the school prayer amendment **mandated prayer**: a government-enforced religious practice
At the same time it is the only institution which can impose economic reforms on extravagant countries as a condition for borrowing.	**it**: the International Monetary Fund (IMF) **institution**: meaning established organization or foundation **impose**: meaning force or place something on others **economic reforms**: changes made to improve the economy, which can include policies or regulations affecting financial systems, government spending, or taxation **extravagant**: meaning characterized by excessive spending or wastefulness condition for borrowing: a requirement that must be met before a loan can be obtained	**the only institution**: referring to the IMF **impose economic reforms**: requiring changes in a country's economic policies **extravagant countries**: nations with excessive or unsustainable spending condition for borrowing: meaning that these reforms are mandatory for the country to receive loans from the IMF

(a) Texts where A1 extracts at least as many as A2.

It was added by a Democratic member of the Democratic-controlled House Ways and Means Committee, and again was part of the price we had to pay to get our economic program adopted.	It: the 70% income tax rate on unearned or investment income, which was reduced to 50% Democratic: pertaining to the Democratic Party, a political party in the United States **Democratic-controlled**: refers to a situation where the Democratic Party holds the majority of control in a particular legislative body **House Ways and Means Committee**: a committee in the U.S. House of Representatives responsible for handling issues related to taxation and government spending adopted: accepted, approved, or put into effect	It: the reduction of the 70% income tax rate on unearned or investment income to 50% Democratic: members of the Democratic Party **House**: the House of Representatives, one of the two chambers of the United States Congress **Ways and Means Committee**: a committee in the House of Representatives responsible for overseeing taxation and other revenue-related matters **the price we had to pay**: the compromises made by the administration to secure approval of its economic program **economic program**: the overall set of policies or reforms aimed at improving the economy adopted: meaning the economic program was officially accepted and enacted by Congress
Tell her also that a lot of the gloomy talk on the TV news is just that talk.	her: referring to Chrissy	her: Chrissy **gloomy**: or negative **that talk**: implying that the negative commentary on the news is just talk, not necessarily reflective of the real situation
A second incident that now will be taken care of by the equal access bill had to do with a group of students holding a Bible discussion at recess on the school grounds.	equal access bill: a law or proposed law intended to ensure that all groups have the same rights to use public spaces or resources	**second incident**: or another related situation equal access bill: the legislation ensuring student groups have equal rights to meet on school grounds **Bible discussion**: a conversation about religious topics from the Bible
He said: Things were different then; the press was on our side.	He: Dick Nixon	He: referring to Richard Nixon, the 37th President of the U.S. **Things**: between the media's role during WWII and in the context of Grenada **the press**: referring to the news media
Our party should strive to change this one is not an Irish-American for example but is instead an American of Irish descent.	descent: meaning ancestry or lineage	**party**: referring to the Republican Party **this**: the practice of hyphenating identity American of Irish descent: emphasizing a unified American identity over divided labels based on heritage

(b) Texts where A2 extracts at least two more mentions than A1.

Table 5.7: Randomly sampled texts (sentences) illustrate individual differences in mention identification (highlighted in bold). Mentions are shown in their original form, rather than their normalized form after fuzzy matching.

tions primarily drawn from external sources. In particular, mentions such as *falsehood, criticism, advocating*, are in my opinion less critical, as they serve mainly as vocabulary explanations. In Table 5.7(b), about half of the unique mentions by A2 are Contextual, including *the price we had to pay, that talk, Bible discussion, Things,* and *this*. A1 and A2's elaborations sometimes reflect differing interpretations of the same mention. For instance, A1 provides valid vocabulary-based elaborations for *impose* and *adopted*, while A2 offers more contextual interpretations—*impose* as *"requiring changes in a country's economic policies"* and *adopted* as *"the economic program was officially accepted and enacted by Congress"*. The former elaborations more briefly support textual comprehension, while the latter more holistically capture the intended meaning: both are valid, yet they reflect different tendencies in performing contextualization (addressing ***RQ5***).

I then divide each annotator's mentions from the validation split into three equally sized tertiles based on mention character length (0-33%, 34-66%, 67-100%), labeled as "Short," "Medium," and "Long." Based on prior observation that A2 identified more Contextual mentions and provided such elaborations, I further pay attention to a long-tail range (90-100%) based on mention character length, and label it as the "Longest." I then randomly sample 10 mentions from each range. The sampled mentions are shown in Table 5.8. Mentions sampled from Short, Medium, and Long are generally similar between A1 and A2. Mentions sampled from the Longest show that A2 tended to identify as mentions lengthy phrases that include pronouns, such as *"does have a plan for all of us," "from both of us warmest regard,"* and *"sends her love and both of us do,"* whereas A1 seldom identified such phrase as mentions. This observation reinforces that A2 tended to comprehend meanings more holistically.

In summary, annotations by A2 are on average denser, with more mentions per sentence and slightly longer in mention character length. A2 also tended to read between the lines, engaging in deeper comprehension, which resulted in elaboration texts that convey implied

Range	Mentions identified by A1	Mentions identified by A2
Short (0-33% length)	*They, Chinese, boyish, program, plants, Contras, there, it, record, them*	*ranch, It, Nancy, it, record, he, we, staff, Defense, his*
Medium (34-66% length)	*inevitably, The Soviets, convention, abandoned, legislation, delegates, subversion, D'Aubuisson, engraving, San Diego*	*Minnesota, moving ahead, that call, compromised, well-to-do, guerrillas, disarmament, campaign, Aerospace, call for*
Long (67-100% length)	*crocodile tears, surrender-or-die ultimatum, administration, Social Security, to the contrary, social gospel, employer and employee tax, constitutional, 25 cents out of each dollar, contribution*	*attorney general, march of empire, surrender-or-die ultimatum, Republican primary, Greater Glory Hath No Man, private property, domestic side, more than 15 years old, Social Security recipients, vicious act of terrorism*
Longest (90-100% length)	*The school authorities, Pediatric AIDS Foundation, 25 cents out of each dollar, Korean airline massacre, state and federal levels, essential social programs, separation of church and state, autonomous corporation, Kuchel, Christopher and Knight, Commodity Credit Corporation*	*heavy print and headlines, surrender-or-die ultimatum, not interfere in the practice of religion, does have a plan for all of us, from both of us warmest regard, loan and scholarship programs, essential social programs, sends her love and both of us do, wrong method had been tried, profane and obscene language*

Table 5.8: Randomly sampled mentions identified by A1 and A2, divided into three tertiles based on mention character length. Additionally, the Longest group includes mentions in the 90-100% character length range.

meaning beyond what is explicitly stated. Since there is no standard aggregation strategy for multiple human annotations, I treat A2's annotations as a primary human reference, and primarily analyze results based on A2's annotations during the automatic evaluation (*RQ6*).

5.6 Automatic Evaluation

Automatic evaluation offers a complementary perspective to human evaluation for addressing ***RQ6***, especially as future approaches may not undergo human evaluation. It uses human annotations as the ground truth to compute evaluation measures where system performance can be compared. The next subsection introduces evaluation measures targeting two aspects of contextualization (see Table 5.9)—identifying mentions, and elaboration.

Table 5.9: Proposed evaluation measures for contextualization.

Aspects	**Automatic Evaluation Measures**
Identifying Mentions	TP, FP, FN, Precision, Recall, F_1, F_2 (corpus-level)
Elaboration	MNLI (comparing full sentences with inserted elaboration texts) BLEU-4 (comparing elaboration texts directly)

Corresponding results are in Section 5.6.2.

While human evaluation directly reflects human perception, automatic evaluation relies on measures that approximate the intended effect. Validation studies assess the validity of these measures by examining their correlation with human evaluation results [132]. In other words, the results presented in Section 5.6.2 can be considered valid answers to ***RQ6*** only if the selected measures are supported by validation studies. Therefore, Section 5.6.3 presents validation results.

5.6.1 Design

Measures for Identifying Mentions: To evaluate the differences between mentions identified by systems and by human annotations, one way is to examine three raw counts from the confusion matrix. Mentions identified by a proposed approach are categorized as True Positives (TP), False Positives (FP), and False Negatives (FN) based on comparison with the ground truth set of human-identified mentions, as described in Section 5.5. When determining whether a system-identified mention match with a human-identified mention, the fuzzy matching strategy is applied.

Alternative measures not based on the confusion matrix, such as mean difference, mean absolute error (MAE), and Cohen's kappa, have been used to analyze the differences between LLM and human labels for tasks like search preferences [155]. However, these measures are not well-suited for the contextualization task. Cohen's kappa is inappropriate due to the unknown probability of negative decisions (i.e., choosing not to identify a given mention, see Section 5.5). Reporting accuracy alongside mean absolute error (MAE) is appropriate for search preference or other classification tasks, where each example has both machine-predicted and human-annotated labels, allowing for a complete confusion matrix. By contrast, contextualization lacks a clear notion of True Negatives (TN), as text spans

not identified by either humans or machines cannot be meaningfully counted, making such measures less suitable.

To continue, the raw counts in each category (TP, FP, and FN) enable the calculation of precision and recall, which are then used to derive the F_1 and F_2 scores [134]. These F-scores serve as single-valued for easier interpretation: higher scores generally indicate a better balance between precision and recall in identifying mentions. F_1 equally weighs precision and recall, whereas F_2 places greater emphasis on recall. For contextualization, F_1 rewards a balanced amount of elaboration that is both accurate and informative, while F_2 prioritizes identifying all mentions that requires elaboration. I consider F_1 the primary measure, as systems with high F_2 scores may still produce excessive false positives (FP), leading to unnecessarily lengthy texts due to over-elaboration. Take the last text in Table 5.7b as an example. Suppose the ground truth identifies *"party," "this,"* and *"descent"*[12] as mentions. In this case, the annotations by A1 and A2—treated as if they were system outputs being evaluated—would achieve F_1 scores 0.5 and 1, respectively. While both A1 and A2 achieve perfect precision (based on fuzzy matching), A2 outperforms A1 by identifying more TP mentions, resulting in higher recall and thus a better F_1 score. Although this example illustrates the calculation at the sentence level, I use F-scores as corpus-level measures. Note that both human annotators identified relatively few mentions per text (typically three or fewer), which makes sentence-level F_1 scoring less appropriate.[13] My focus is on individual mentions despite high variance across texts; therefore, precision and recall are computed as single corpus-level values, which are then used to calculate the F-measure.

Statistical Tests for Identifying Mentions: The F_1 and F_2 scores each introduce a ranking among my six systems, which include three LLM implementations for Approach

[12]Fuzzy matching would align *"descent"* with *"American of Irish descent."*

[13]This is because the harmonic mean used in F-score calculations is particularly sensitive to low values; if either precision or recall is zero for a given text, the F_1 score becomes zero. Therefore, it is preferable to compute precision and recall using micro-averaging at the corpus level rather than macro-averaging across individual texts.

4 (Joint). Do system comparisons based on single-valued F-scores (F_1 and F_2), computed on a single sample, reflect real differences? To answer this, doing bootstrapping [146] gives 95% bootstrapped confidence intervals for the F-measure values. These confidence intervals are then visualized using error bars for each system placed side by side, allowing for visually comparing any pair of ranks. For example, in the case of F_1, a total of 1,000 samples are generated by repeatedly sampling with replacement from the 395 validation split texts, creating 1,000 resampled sets, each containing 395 texts and yielding an F_1 score for each system. Based on these 1,000 F_1 scores, a 95% confidence interval is derived for each system. If neither confidence interval of one system overlaps the point estimate of another system, this serves as a strong—sufficient but not necessary—indicator that the F_1 scores of the two approaches are statistically distinguishable.

For the system pairs with confidence intervals that overlap, statistical significance is more rigorously determined using a bootstrap hypothesis test, whose details are in Appendix. E. Specifically, for each such pair of systems, I obtain a two-sided p-value under the null hypothesis. The significance level (α) of 0.05 is adjusted using the Bonferroni Correction [21], which divides α by the number of tests, to control for false positives when conducting multiple tests simultaneously. A difference between the F-scores of the two approaches is deemed statistically significant when the p-value is less than the significance level (α), after Bonferroni correction.

Classifying Mentions as External or Contextual: During experiments, I also let the LLM-based approaches to attribute the mentions identified into the two categories ("External" and "Contextual," essentially, a classification task). This serves as an intermediate output that the LLMs can provide to assist my analysis of their behaviors. The prompt to attribute mention types is in Table 5.10, and the prompt examples are manually picked from the training split. By grouping mentions based on the attribution labels by systems, I later examine how these approaches realize the design goal in Table 5.14.

Prompting Elements	**Attributing Mention Types**
System Role	`Whom_talking_to` (Texts in typewriter font represent variables listed in Table 5.2.)
Task Background	`Two_types_mentions`
Task Description	Your task is to first read the context, the original text, and a particular mention, then answer the multiple-choice question below, on whether a particular mention from the original text is a coreference or external mention. You should answer with only one word. Your answer should be either "coreference" or "external". Do not write sentences. Do not provide any explanation.
Few-shot Examples	Context: You have entered into the most meaningful relationship there is in all human life. Original text: It can be whatever you decide to make it. Mention: It Question: Which of the following option is the type of this mention: It? - coreference - external The answer is: coreference Context: Nicaragua now has a military greater than that of all Central American nations combined. Original text: The government has declared its revolution is not bound by any national boundaries Mention: The government Question: Which of the following option is the type of this mention: The government? - coreference - external The answer is: coreference Context: Thank you very much for your letter. Original text: I'm sure the drumbeat of false propaganda must be even more frustrating to you and your family than it is to me. Mention: false propaganda Question: Which of the following option is the type of this mention: false propaganda? - coreference - external The answer is: external Context: I'm at 190 pounds which was my pre-operation weight and feel just fine. Original text: I will inform our White House Dr. Burton Smith of your letter and the information about hydrazine sulfate. Mention: hydrazine sulfate Question: Which of the following option is the type of this mention: hydrazine sulfate? - coreference - external The answer is: external

Table 5.10: Prompt elements for categorizing mentions as External or Contextual. The wording of these prompts differs slightly from the wording in this dissertation.

Measures for Elaboration Quality: For elaboration, quality is measured by comparing it to human annotations, which are assumed to be high-quality. The measures used are BLEU and MNLI. For simplicity, I assume a single ground-truth human elaboration per mention, which means results based on A1 and A2's annotations are separately calculated. The scores are computed at the corpus level, i.e., micro-averaged over all elaboration texts, each corresponding to an individual mention, rather than at the sentence level. The scores can only be computed for mentions that were true positives (TP) from mention identification because otherwise there is no ground-truth elaboration. Scores does not differentiate External or Contextual mentions, since most approaches (except Joint) do not explicitly rely on category information to elaborate. Since the number of true positive (TP) mentions varies across different approaches, I also report the average BLEU and MNLI scores over all human-identified mentions under the column "All Relevant (TP+FN)", for which the total number of mentions remains consistent across approaches.[14] The values reported under "All Relevant (TP+FN)" reflect the cascading effect of both identifying mentions and elaboration: if a system fails to identify one mention, no elaboration is produced, resulting in a score of zero. BLEU and MNLI scores computed using A2's annotations for the "All Relevant (TP+FN)" condition serve my primary evaluation measures for elaboration quality, while the scores for the "TP (only)" condition help examine the choice of BLEU and MNLI as evaluation measures.

BLEU (BiLingual Evaluation Understudy) [121] measures the precision of word-level n-grams, i.e., between the system and ground-truth sets of words in the texts. In contextualization, an elaboration text is a clause, while all other parts of the original text remain unchanged. There would therefore be significant word overlap between the original text and the contextualized text with the elaboration inserted. So instead, I calculate the BLEU score between only the system and ground-truth elaboration texts, i.e., clauses. BLEU-

[14]The values under "All Relevant" are derived by multiplying the values under "TP" by Recall= $\frac{TP}{TP+FN}$.

4 is a widely used implementation of BLEU that considers up to 4-gram overlaps. The weight distribution reflects how different n-gram matches contribute to the overall score, considering both lexical choices and syntactic structure. I compute the BLEU scores using the NLTK Python library, applying the default uniform n-gram weight distribution of $(0.25, 0.25, 0.25, 0.25)$ for BLEU-4. Additionally, I use Smoothing Method 3, also known as NIST geometric sequence smoothing [33]. This method mitigates harsh penalties in sentence-level BLEU when unseen n-grams occur, and has been shown to correlate better with human judgment than unsmoothed scores. It works by replacing zero n-gram match counts with small adjusted values derived from a geometric sequence based on n. I do not limit the evaluation to BLEU alone because of its well-known limitations, such as its wide variability in correlation strength and potential biases when evaluating neural network outputs [132].

MNLI models the relationship between the ground truth (as premise) and the system output (as hypothesis) in terms of entailment, contradiction, or neutrality. The MNLI score is obtained from the HuggingFace RoBERTa-large model[15], which is a pre-trained transformer-based neural network model, fine-tuned on the MNLI (Multi-Genre Natural Language Inference) dataset. Unlike BLEU, I calculate the MNLI score between two full contextualized sentences. A contextualized sentence is the original text with one elaboration text for one mention inserted. The ground truth is the original text with one human-written elaboration text inserted, while the system output is the original text with one system elaboration text inserted. I calculate the MNLI score in both directions—ground truth to system output, and system output to ground truth—and then report the average of the two. Calculating bi-directional MNLI aims to indicate mutual entailment as a proxy for semantic equivalence [6]. An MNLI score that is close to 1 indicates that a system elaboration text is essentially a paraphrase to the ground-truth elaboration text.

[15] `https://huggingface.co/FacebookAI/roberta-large-mnli`

I also apply the procedure of plotting bootstrapped confidence intervals for both BLEU and MNLI measures as part of the statistical analysis. In addition to reporting BLEU and MNLI values, I examine mentions and elaborations by their contents, and try to understand how those measures reflect elaboration quality. Prior work cautions against interpreting BLEU scores at the sentence level due to their generally low correlation with human judgments; however, reporting corpus-level BLEU, as done here, remains a common practice.

Evaluation Criteria Considered but Excluded: The above constitutes the complete set of measurements for automatic evaluation. Finally, it is worth noting some excluded measures. For text rewriting tasks such as text style transfer [82], common evaluation criteria often include the preservation of original semantics and the natural language quality of the rewritten text. However, both were omitted in the final selection of evaluation measures and were omitted from being reported in the experiments for this dissertation, except a discussion of the design as below. The preservation of original semantics, also referred to as semantic consistency, can be evaluated through textual entailment between the original and contextualized texts. Rather than modeling it explicitly, this aspect is reflected in the reported MNLI scores and the analysis of mentions and elaborations. For example, inserting a human-written elaboration clause into the original text is assumed to preserve meaning perfectly, given that annotators performed the contextualization with high quality. Furthermore, previous human evaluations (Section 5.4.1) specifies that elaborations containing incorrect information, such as elaborating a different mention with the same surface form, are considered low quality, highlighting a key instance of semantic inconsistency. Regarding the natural language quality of the contextualized text, this criterion evaluates whether the output forms a natural and fluent English sentence. Because elaboration texts are, by design, clauses rather than complete sentences, they do not achieve full fluency on their own. Given that large language models (LLMs) are well-known for generating natural and fluent text [25], rendering elaboration texts into fully fluent contextualized sentences is left

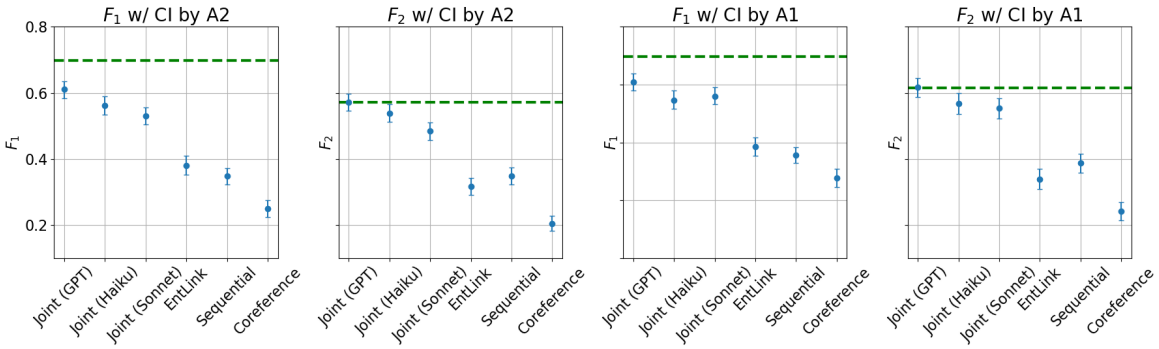


Figure 5.5: F-scores and 95% confidence intervals, on the validation split, based on annotations from A2 (primary) and A1. Reference human baselines are shown as green horizontal lines, based on evaluating A1's identified mentions against A2's annotations (in the first and second subplots), and A2's mentions against A1's annotations (in the third and fourth).

Table 5.11: Results on mention identification by all approaches on the validation split. Slashes separate the values by different annotators, A1 and A2.

Model Family	System	TP	FP	FN	Precision	Recall	F_1	F_2
Human (reference)		642/642	400/159	159/400	0.616/0.801	0.801/0.616	0.697/0.697	0.756/0.646
OpenAI	Joint (GPT)	497/569	334/258	303/469	**0.598/0.688**	**0.621/0.548**	**0.609/0.610**	**0.616/0.571**
Claude 3	Joint (Haiku)	466/546	440/359	335/495	0.514/0.603	0.582/0.524	0.546/0.561	0.567/0.539
	Joint (Sonnet)	439/479	326/284	362/563	0.574/0.628	0.548/0.460	0.561/0.531	0.553/0.486
OpenAI	EntLink	252/296	259/215	549/746	0.493/0.579	0.315/0.284	0.384/0.381	0.339/0.316
	Sequential	330/364	724/685	471/678	0.313/0.347	0.412/0.349	0.356/0.348	0.388/0.349
	Coreference	179/191	309/297	622/851	0.367/0.391	0.223/0.183	0.278/0.250	0.242/0.205

for future work.

5.6.2 Results

Identifying Mentions: Figure 5.5 visually presents the performance of all systems and human reference baselines on the validation split, with each dot's vertical position indicating its F-score point estimate for mention identification. In response to ***RQ6***, the systems are ordered in descending order of F_1 based on A2's annotations, with Joint (GPT) consistently outperforming all others. The top three positions are occupied by the three Joint systems across all comparisons. Table 5.11 reports the same evaluation measure values for identifying mentions but in a tabular view. The numbers in each cell separated by

slashes represent the values based on annotators A1 and A2. The first row is a reference human baseline (the green dashed horizontal line in Figure 5.5), where A2 and A1's identified mentions are evaluated by treating one set of annotations (A1's/A2's) as the ground truth and the other as the system output. The next six rows are in descending order of F_1 performance based on A2's annotations, From the F_1 and F_2 scores, the six systems yield four rankings. For discussion purposes, measure values based on A2's annotations are referenced, as A2 identifies more mentions per text and exhibits a stronger ability to read between the lines and do deeper comprehension (as discussed in Section 5.5.3). The top-performing one, Joint (GPT), achieves an F_1 score of 0.610 and an F_2 score of 0.571. Compared to the human reference (with A1's identified mentions evaluated against A2 as the ground truth yielding F_1 of 0.697 and F_2 of 0.646), Joint (GPT) achieves 87.5% of human performance by F_1 and 88.4% by F_2. In all four rankings by both annotators, the top three positions are consistently held by the three Joint systems, each achieving over 75% of human performance. From the fourth to the sixth rank, EntLink ranks fourth by F_1 and fifth by F_2, while Sequential ranks fourth for by F_2 and fifth by F_1. Coreference is consistently in the sixth (last) rank across all four rankings, with its measure values never exceeding 40% of human performance.

Additionally, I find that systems rankings based on A1 and A2's annotations as the ground truth are highly correlated, demonstrating a strong inter-annotator agreement in response to (***RQ5***)—not in the content of their annotations, as reported by Cohen's kappa on External vs. Contextual classification labels (Section 5.5.3), but this time based on extrinsic measures reflecting the use of those annotations for evaluation. Kendall's tau, a non-parametric rank correlation coefficient for measuring ordinal association between two variables [84], is 0.867 (p-value: 0.017) for F_1 rankings, and 1.000 (p-value: 0.003) for F_2 rankings.

When keeping the annotator constant and switching between F_1 and F_2, all three Joint

systems consistently maintain their advantage. This suggests that the Joint systems are well-balanced in identifying mentions. Among the three other systems, a consistent rank swap occurs between EntLink and Sequential, revealing that Sequential identifies more desired mentions than EntLink. EntLink is disadvantaged by low recall despite having a competitive precision, which is presumably tied to its design goal, focusing on named entities but not contextual mentions. I examine design goal in Table 5.14.

Figure 5.5 visualizes bootstrapped 95% confidence intervals for F_1 and F_2. The three Joint systems consistently show statistically significant advantages over the other three systems. The disadvantage of Coreference compared to all other systems is also consistently statistically significant. Next, I examine these overlapping confidence interval (CI) bars with the more rigorous bootstrap hypothesis test. Each sub-plot in Figure 5.5 has three pairs of overlapping CIs. The adjusted significance level $\alpha' = 0.0167$ (i.e., $0.05/3$), is used to account for multiple comparisons based on Bonferroni Correction. The tests confirm further that Joint (GPT) has a statistically significant advantage over both Joint (Haiku) and Joint (Sonnet) by both F-scores and with either annotator's annotations. Joint (Haiku) significantly outperforms Joint (Sonnet) only with A2 annotations for both F-scores. EntLink significantly outperforms Sequential in F_1 scores for both annotation sets, while Sequential has the advantage in F_2 with both annotations.

Table 5.12 and Figure 5.6 present the same experiments conducted on the test split.[16] Figure 5.6 shows the F-scores and confidence intervals (CIs), followed by bootstrap hypothesis testing to assess statistical significance. Across both the validation and test splits, Joint (GPT) shows a statistically significant advantage over the other two Joint systems in F_1 scores based on A2's annotations (the primary ground truth). Additionally, all three Joint systems significantly outperform the other systems under A2's annotations. In comparison with the human reference (in the first row), the best system in each column has 95.9% to

[16]Results on the test split are consistently presented at the bottom of the pages in this dissertation.

110.0% of human performance. These findings provide a strong answer to ***RQ6***, demonstrating the effectiveness of these systems in performing contextualization. The human reference F_1 and F_2 scores range from 72.4% to 83.5% compared to those same scores on the validation split. This suggests increasing individual differences between the two annotators over time, addressing ***RQ5***. The agreement between annotators on system rankings remains high. Kendall's tau between rankings using each annotator's annotations is 0.733 (not significant, p-value: 0.056) for F_1, and 0.828 (significant, p-value: 0.022) for F_2.

Next, to understand how system-identified mentions differ from human annotations, and to compare each system's strengths, I analyze the content of true positives (TP), false positives (FP), and false negatives (FN). I examine the 10 validation texts analyzed during human annotations (Table 5.7), which were selected as two sets earlier; one with at least

Table 5.12: Results on mention identification by all systems on the test split. Slash separate the values by annotators, A1 and A2.

Model Family	System	TP	FP	FN	Precision	Recall	F_1	F_2
Human (reference)		364/364	481/146	146/481	0.431/0.714	0.714/0.431	0.537/0.537	0.631/0.468
OpenAI	Joint (GPT)	335/423	392/304	175/422	0.461/**0.582**	**0.657/0.501**	0.542/**0.538**	**0.605/0.515**
Claude 3	Joint (Haiku)	299/406	512/404	211/439	0.369/0.501	0.586/0.480	0.453/0.491	0.524/0.484
	Joint (Sonnet)	326/385	353/296	184/460	0.480/0.565	0.639/0.456	**0.548**/0.505	0.599/0.474
OpenAI	EntLink	231/244	221/208	279/601	**0.511**/0.540	0.453/0.289	0.480/0.376	0.463/0.318
	Coreference	165/180	333/318	345/665	0.331/0.361	0.324/0.213	0.327/0.268	0.325/0.232
	Sequential	324/325	728/716	186/520	0.308/0.312	0.635/0.385	0.415/0.345	0.524/0.368

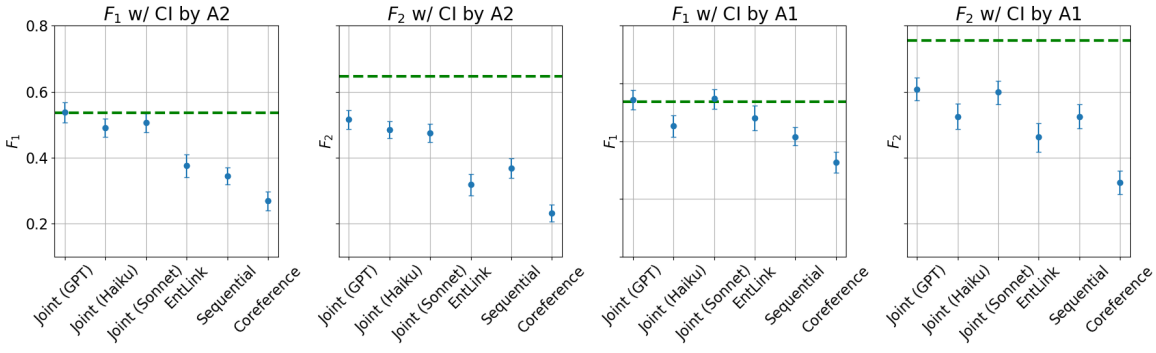


Figure 5.6: F-scores and 95% confidence intervals, on the test split, based on annotations from A2 (primary) and A1, following the format of Table 5.5.

as many mentions from A1 as from A2, and the other with more mentions from A2 than from A1. Table 5.13 presents four of those texts; the remaining six are omitted due to repetitive patterns. For the same text, each system identifies a slightly different set of mentions, alongside those identified by annotator A2. True positives (TP) are listed without parentheses, while false positives (FP) are enclosed in parentheses. The number of mentions identified, particularly the proportion of true positives (TPs), clearly demonstrates the strength of Joint (GPT). EntLink does not include contextual spans or colloquial terms that aren't recognized as proper names, which aligns with its design. Coreference provides a strong complement to EntLink, often identifying mentions that are distinct from those identified by EntLink, which is also expected by design. Additionally, Coreference extracts some valid mentions in my opinion, such as *"that situation,"* that are overlooked by A2. Its strength lies in precision, while its weakness is limited recall.

The following is a summary of six recurring patterns identified, ordered by decreasing prevalence, each of which may affect the TP, FP, and FN counts when evaluated using human annotations. Without reiterating this evaluation condition, I note the direction of impact for each pattern.

- **Challenging Long Contextual Mentions**: Based on prior analysis, A2 (Human) annotations tend to more holistically capture the intended meaning of the texts compared to A1, placing greater emphasis on contextual cues rather than external or surface-level interpretations. For instance, in the second example, the phrase *"the only institution"*—referring specifically to the IMF—is more aligned with the context than a generic explanation of *"institution"* as any organization. Among the systems, EntLink is the only one that consistently favors shorter, less contextualized variants of mentions, which are sometimes less valuable. While these may still count as true positives (TPs) under fuzzy matching, they should have been penalized as a

Utterance	A2 (Human)	EntLink	Coref.	Sequential	Joint(Sonnet)	Joint(Haiku)	Joint(GPT)
Finally; as for Lebanon, that situation is only one facet of the whole Middle East problem.	Lebanon, facet, Middle East problem	Middle East, Lebanon	– (that situation)	Middle East, Lebanon, (that situation)	Middle East problem, Lebanon, (that situation)	Lebanon, Middle East problem	Lebanon, Middle East problem
At the same time it is the only institution which can impose economic reforms on extravagant countries as a condition for borrowing.	the only institution, impose economic reforms, extravagant countries, condition for borrowing	institution	– (it)	– (IMF, it)	extravagant countries, economic reforms, institution, (it)	borrowing, extravagant countries	extravagant countries, the only institution, economic reforms
It was added by a Democratic member of the Democratic-controlled House Ways and Means Committee, and again was part of the price we had to pay to get our economic program adopted.	It, Democratic, House, Ways and Means Committee, the price we had to pay, economic program, adopted	House, Democratic	It, (our)	House, It, Democratic, (our, 70 percent income tax rate)	House, economic program	House, economic program	It, Democratic, House, economic program
A second incident that now will be taken care of by the equal access bill had to do with a group of students holding a Bible discussion at recess on the school grounds.	second incident, equal access bill, Bible discussion	incident, equal access bill, Bible, (school)	second incident	equal access bill, second incident, Bible, (school, incident)	equal access bill	equal access bill, Bible discussion, (recess, school grounds)	equal access bill

Table 5.13: Mentions identified by different systems are shown under true positive (TP) and false positive (FP) categories. A dash (–) indicates that a system identified no true positive (TP) mentions in a text. FP mentions are shown in parentheses. False negatives (FN) are not listed but can be inferred from the column A2 (Human), where mentions are shown in their original form, and serve as the ground truth. TP mentions are normalized after fuzzy matching, consistent with how the counts are calculated in Table 5.11.

false positive (FP). Also in the second example, A2 (Human) identifies *"condition for borrowing,"* while all systems except Joint (Haiku) completely omit this mention. Joint (Haiku) identifies only the fuzzy match *"borrowing."* which illustrates that matching the full span expected by human annotators is challenging. These cases contribute to false negatives (FNs).

- **Short Contextual Mentions Ignored by LLM**: This pattern is exemplified by one short Contextual mention *"second incident"* in the fourth example, identified by A2 (Human).[17] The failure of LLM-based models to identify this particular one may be

[17]This was annotated as a Contextual mention by A2, although their elaboration (*"another related situation"*) provides relatively weak contextual support. A stronger elaboration, as guided by the instructions, would incorporate context-specific information, such as: *"a second incident like that of the cafeteria prayer*

attributed to prompt examples that insufficiently showcase what Contextual mentions could look like. Consequently, even the strongest model, Joint (GPT), misses this mention, resulting in False Negatives (FNs).

- **Human Inclusion but Optional**: A2 (Human) occasionally identifies mentions that may be considered optional under the written guidelines, such as *"adopted,"* although this tendency is more common in A1's annotations. In this case, any system that omits the optional mention would be penalized with a false negative (FN), but the mention could also have been considered a true negative (TN). Another example is *"Bible discussion,"* an External mention whose elaboration is *"a conversation about religious topics from the Bible."* It is valuable to acknowledge that this pattern has random variations without a set guideline—humans sometimes also omit optional mentions, such as *"recess"* in the fourth example, extracted by some systems but not by A2, which penalizes Joint (Haiku) with a FP.

- **Human Deliberate Omission**: in the first example, *"that situation"* largely satisfies the prompt's instruction, which is identified by Corefence, Sequential, and Joint (Sonnet). However, A2 (Human), Joint (Haiku), and Joint (GPT) do not identify it, likely because the full preceding context (see problem formulation in Section 5.1), *"... We hope some modifications might still take place in Congress,"* lacks sufficient information to allow for elaboration, thus eliminating the value of identifying this mention. Deciding to omit a mention would be a complex reasoning process specific to a special case, where the guidelines do not provide clear direction. In this case, less conservative models may generate a false positive (FP) but they could also be considered a true positive (TP). Only more conservative models that choose to skip the mention are not penalized in this case.

leading to the court decision in a southern state."

- **Interchangeable Mentions**: The original text may include a subject (a noun or a noun phrase) and an object (a clause referring to the same entity). Selecting either as the mention often implies the other, as shown in the second example with *"it"* vs. *"the only institution"*. Despite using fuzzy matching, current evaluation does not account for syntactic relationships between mentions in the same sentence. As a result, systems like Coreference, Sequential, and Joint (Sonnet) may extract a form that is interchangeable with the reference but still be penalized with both a false positive (FP) and a false negative (FN), rather than credited with a true positive (TP). A similar issue arises in the third text, where identifying *"our"* is penalized, even though understanding its elaboration is largely sufficient to infer another mention *"the price we had to pay"* that is present in A2's annotations.

- **Invalid Mentions Not in Utterance**: A unique failure of the current Sequential design arises from the fact that Sequential works with an elaborated, intermediate text as input for the third pass. This leads to more false positive (FP) mentions which were not present in the original text and thus would not appear in human annotations. For example, *"IMF"* in the second example and *"70 percent income tax rate"* are not valid, as they are not part of the original utterance.

To succinctly answer the part of ***RQ6*** that involves mention identification, Joint (GPT) achieves the highest F-scores, exceeding 85% of the F-scores obtained by the human reference baseline. My inspection reveals a systematic gap: current systems, including Joint (GPT), sometimes miss or incompletely capture Contextual mentions that humans emphasize to convey intended meanings. Additionally, while F-scores provide an intuitive and informative way to compare LLM outputs with human performance, they may be imprecise due to natural variation in human annotations (which reflect divergence from individual differences) and limitations of the fuzzy matching process. A more formal validation

Table 5.14: The number of mentions identified by the systems, along with the counts attributed to External and Contextual categories based on their own attribution, are shown for the validation split.

Model Family	System	Overall	External	Contextual
OpenAI	EntLink	511	346	165
	Coreference	488	33	455
	Sequential	1,057	426	631
	Joint (GPT)	832	404	428

study of the precision measure, used in identifying mentions and contributing to F-scores, is presented in the next subsection.

Design Goal: Table 5.14 shows the effect of prompt design. Previous results provide some evidence regarding this question, such as the behaviors of the EntLink and the Coreference systems. Since all four proposed approaches have the same OpenAI GPT-4o model implementation and differ only in prompt design, I allow the proposed approaches to attribute mention categories, and count the number of External and Contextual mentions. No human annotation is involved: the counts are not processed through fuzzy matching, and therefore have a slight difference from the positive counts (sum of TP and FP) in Table 5.11. The EntLink approach identifies mentions but attributes only two-thirds (67.7%) of them as External, showing a necessity for elaboration of Contextual mentions that it cannot accomplish. The Coreference approach identifies mentions and attributes most of them (93.2%) as Contextual. The Sequential approach identifies more than the total number of mentions by EntLink plus Coreference, and further attributes more Contextual mentions than the Coreference approach (631 vs. 488). The Joint approach (in system GPT) identifies many mentions, yet still fewer than the Sequential approach, with half (48.6%) classified as External and the other half (51.4%) as Contextual. This suggests that mentions needing contextualization are present in both types in roughly balanced proportions. These results suggest that mentions identified for elaboration be chosen selectively, as is done by the first pass of Joint (GPT), rather than being accumulated over cascade of prompts, as in

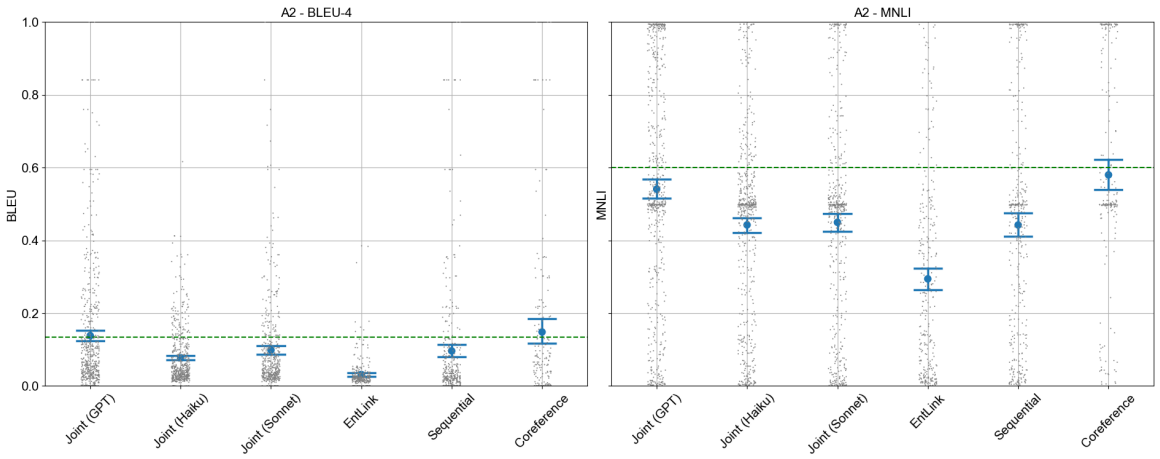


Figure 5.7: Blue dots and error bars represent the average, corpus-level BLEU-4 and MNLI scores with 95% confidence intervals for elaborations of true positive (TP) mentions on the validation split, based on A2's annotations. Overlaid grey dots in the scatter plots show individual scores for each TP mention's elaboration, highlighting the distribution and contrast between BLEU-4 and MNLI scores. A reference human baseline is shown as a green horizontal line, based on evaluating A1's elaborations against A2's elaborations as the ground truth. Systems are sorted consistently with those in Figure 5.7.

Table 5.15: Results on elaboration quality based on human annotations (A2) on the validation split.

System	TP			All Relevant (TP + FN) N=1,042	
	N	BLEU-4	MNLI	BLEU-4	MNLI
Human A1 (Reference)	642	0.135	0.601	0.083	0.370
Joint (GPT)	569	0.138	0.541	**0.076**	**0.296**
Joint (Haiku)	546	0.078	0.443	0.041	0.232
Joint (Sonnet)	479	0.099	0.450	0.045	0.207
EntLink	296	0.032	0.294	0.009	0.084
Sequential	364	0.096	0.442	0.034	0.155
Coreference	191	**0.148**	**0.580**	0.027	0.106

Sequential.

Elaboration: Figure 5.7 shows the average, corpus-level BLEU-4 and MNLI scores, with 95% confidence intervals [146], for true positive (TP) mentions' elaborations on the validation split. TP mentions are based on the mention identification results using A2 annotations (see Table 5.11). Table 5.15 presents the same average scores both for true positive (TP) mentions, and for all relevant mentions (where FN mentions receive a score of zero). The Coreference system ranks highest for measure values averaged over TP mentions. Ac-

cording to A2, the ranked order of systems based on BLEU-4 is Coreference, Joint (GPT), Joint (Sonnet), Sequential, Joint (Haiku), and EntLink. The ranked order based on MNLI is Coreference, Joint (GPT), Joint (Sonnet), Joint (Haiku), Sequential, and EntLink. BLEU-4 and MNLI give highly correlated system rankings: the Kendall's tau correlation between the two rankings is 0.867 (significant, p-value: 0.017), with the only rank swap occurring between Sequential and Joint (Haiku). The "All Relevant" grouping in Table 5.15 reflects how well the systems elaborate on mentions that, according to human annotations, are expected to be elaborated. The Joint (GPT) system ranks highest when averaged over all relevant mentions, primarily because it has elaborated a greater number of TP mentions than Coreference. Table 5.16 presents the same scores on the test split, showing that the top-performing systems are consistent with those on the validation split.

Figure 5.7 shows bootstrapped 95% confidence intervals (CIs) [146]. The figure includes an additional scatterplot layer, with gray dots representing individual scores for each true positive (TP) mention's elaboration. False negative (FN) mentions' elaboration texts receive a zero score and are excluded from the scatter plot. Coreference and Joint (GPT) are significantly better than all others by both BLEU-4 and MNLI, but they are not statistically distinguishable from each other by either measure by a bootstrap hypothesis test (not significant, p-value$>$0.05). Among the less competitive systems, Joint (Sonnet),

Table 5.16: Evaluating elaboration quality by human annotations (A2) on the test split.

System	TP			All Relevant (TP + FN) N=845	
	N=	BLEU-4	MNLI	BLEU-4	MNLI
Human (A1, Reference)	364	0.170	0.581	0.073	0.250
Joint (GPT)	423	0.167	0.558	**0.084**	**0.279**
Joint (Haiku)	406	0.070	0.442	0.034	0.213
Joint (Sonnet)	385	0.099	0.440	0.045	0.201
EntLink	244	0.034	0.310	0.010	0.090
Sequential	325	0.142	0.523	0.055	0.201
Coreference	180	**0.211**	**0.627**	0.045	0.134

Joint (Haiku) and Sequential are not statistically distinguishable by MNLI by bootstrap hypothesis tests (not significant, p-value>0.05). EntLink is clearly the least effective by both measures, with its point estimate consistently below the lower bound of the confidence interval for every other system. This indicates a statistically significant performance gap, likely due to its design: EntLink draws elaboration solely from world knowledge (e.g., Wikipedia content) in a stilted manner, without incorporating contextual information in the way human annotators tend to do. Interestingly, despite its under-performance, EntLink exhibits the smallest variance in BLEU-4 scores compared to the other systems. By contrast, the Coreference system achieves the highest average BLEU-4 score, but also exhibits the widest confidence interval and the greatest variance. Comparing the two measures: BLEU-4 and MNLI, I observe distinct score distributions from the grey dots. BLEU-4 values are concentrated in the lower range of $[0, 0.2]$, reflecting its sensitivity to exact n-gram matches and lexical overlap. In contrast, MNLI scores are more evenly distributed across the $[0, 1]$ range, indicating a stronger ability to capture semantic alignment between texts rather than relying on lexical similarity as BLEU-4 does.

To understand how MNLI represents elaboration quality (***RQ6***), Table 5.17 shows the mentions and elaboration texts for the same four sentences as in Table 5.17, but sorts them based on their MNLI score. Based on the score's percentile among the full set of measure scores, I mark them as low, middle, or high tertiles alongside the measure values. The following is a summary of five recurring patterns identified and marked in the table, each of which may influence the MNLI score when comparing systems' elaboration texts ("LLM elaboration") to human-annotated elaboration texts ("human elaboration"), potentially causing either an underestimation or overestimation. They are ordered by decreasing prevalence.

① **Comparable Elaboration**: Both the human and LLM elaboration are of similar

Utterance	Mention	A2 (Human)	System	Elaborations	BLEU-4	MNLI
Finally; as for Lebanon, that situation is only one facet of the whole Middle East problem.	Lebanon	country in the Middle East	Joint(Haiku)	a country in the Middle East located on the Mediterranean coast, bordered by Syria to the north and east, and Israel to the south, known for its complex political landscape and historical conflicts involving various religious and ethnic groups	0.883	0.497 Medium ②
	Middle East problem	referring to the broader set of issues in the Middle East, which could include political instability, territorial disputes, religious conflicts, etc.	Joint(Sonnet)	the ongoing conflicts, tensions, and disputes between various nations and groups in the Middle East region, including territorial disputes, religious differences, and political instability that had persisted since the mid-20th century	0.083	0.497 Medium, near High ②
At the same time it is the only institution which can impose economic reforms on extravagant countries as a condition for borrowing.	extravagant countries	nations with excessive or unsustainable spending	Joint(Haiku), and similarly Joint(Sonnet), Joint(GPT)	nations that spend money excessively or wastefully, often running large budget deficits or accumulating unsustainable levels of national debt without fiscal discipline	0.022	0.520 Medium, near High ②
	condition for borrowing	meaning that these reforms are mandatory for the country to receive loans from the IMF	Joint(Haiku)	borrowing from the International Monetary Fund (IMF)	0.042	0.462 Medium ⑤
	the only institution	referring to the IMF	Joint(Sonnet), and similarly Joint(GPT)	the International Monetary Fund (IMF)	0.107	0.984 High ① ③
It was added by a Democratic member of the Democratic-controlled House Ways and Means Committee, and again was part of the price we had to pay to get our economic program adopted.	economic program	the overall set of policies or reforms aimed at improving the economy	Joint(Haiku)	the economic policy aimed at reducing taxes and government spending, which was a key part of the Reagan administration's approach to stimulating economic growth, often referred to as "Reaganomics"	0.032	0.467 Medium ③
			Joint(Sonnet)	my administration's 1981 Economic Recovery Tax Act, which included major tax cuts and changes to federal spending aimed at stimulating economic growth and reducing inflation	0.030	0.427 Medium ③
	It	the reduction of the 70% income tax rate on unearned or investment income to 50%	Joint(GPT), and similarly Coreference, Sequential	the reduction of the 70 percent income tax rate on unearned or investment income to 50 percent	0.630	0.993 High ①
A second incident that now will be taken care of by the equal access bill had to do with a group of students holding a Bible discussion at recess on the school grounds.	equal access bill	the legislation ensuring student groups have equal rights to meet on school grounds	Sequential and EntLink	the equal access to cobra act (s.3182) was a bill which would amend the internal revenue code, the employee retirement income security act of 1974...	0.010	0.288 Low ④
			Joint(Sonnet), and similarly Joint(Haiku), Joint(GPT)	a law passed in 1984 that requires public schools to give religious groups the same access to school facilities as other student groups, protecting students' rights to meet for religious purposes during non-instructional time	0.050	0.498 Medium, near High ③
	second incident	or another related situation	Coreference	another event related to religious activities in schools	0.066	0.819 High ③

Table 5.17: Mentions (original form in A2's annotations) and elaboration texts from the validation split with BLEU-4 and MNLI scores. MNLI scores are followed by a label indicating their tertiles, using cutoff values of 0.357 (33rd) and 0.527 (67th) to divide the low, medium, or high tertiles. The circled digit corresponds to the summarized pattern.

high quality, and the MNLI score is in the high range. When the Coreference system produces correct elaborations (as judged based on human annotation and in my view), they are often similar in length, likely reflecting the limited contextual information available for elaboration.

(2) **Concise Human vs. Long LLM Elaboration, Both Correct**: Despite the difference in length, both are correct (as judged based on human annotation and in my view). The MNLI score falls in the medium range.

(3) **LLM Elaboration is Better, Both Correct**: The annotation interface (see Figure 5.3) provides a short inline input box for annotators to enter elaboration text. This design illustrates to annotators the effect of elaboration texts functioning as clauses within the sentence, but it may also encourage annotators to write more concise responses. In contrast, stronger LLM elaborations often provide richer details—for example, elaborating the *"economic program"* as the 1981 Economic Recovery Tax Act, or elaborating the *"equal access bill"* as a particular law passed in 1984. In such cases, automatic evaluation assign only a moderate MNLI score, underestimating quality due to the limitations of human annotations.

(4) **LLM Elaboration is Incorrect**: This pattern generally arises from EntLink Wikification errors (affecting both the Sequential and the EntLink systems), such as linking to a temporarily unavailable Wikipedia entry, or matching to an incorrect entity such as *"equal access bill"* linked to *"Equal Access to COBRA Act."* As a result, the MNLI score is low.

(5) **Fuzzy Matching Leading to Elaboration Differences**: Human and LLM elaborations differ for overlapping mentions due to fuzzy matching, such as *"condition for borrowing"* by A2 vs. *"borrowing"* by Joint (Haiku). This mismatch leads to an

underestimation of elaboration quality, with MNLI in the medium tertile, reflecting a cascading effect that originates from the mention identification step.

To succinctly answer ***RQ6*** on elaboration, Joint (GPT) and Coreference systems achieve the highest BLEU-4 and MNLI scores, all of which exceed 95% of the corresponding scores of the human reference baseline. My inspection reveals that most systems are capable of producing correct elaboration texts.[18] Some LLM elaborations even surpass human-written ones by offering more specific details and generating longer-form elaborations. Additionally, while MNLI scores generally serve as a reliable proxy for estimating elaboration quality, their accuracy is constrained by the use of only one human-written elaboration as the reference. However, BLEU-4 appears more rigid than MNLI in recognizing high-quality elaborations, e.g., the mention and elaboration of *"extravagant countries"* has a low BLEU-4 score but a MNLI score in the medium to high tertiles. BLEU-4 may still be useful as a corpus-level measure, due to its strong correlation with MNLI for ranking the systems. A more formal validation study of the MNLI measure is presented in the next subsection.

5.6.3 Validating Automatic Measures with Human Evaluation

Section 5.4 presents human evaluation results on system outputs, while Section 5.6.2 reports automatic evaluation results on the same outputs. When any comparative advantages among systems identified by automatic evaluation align with those from human evaluation, it suggests that the automatic evaluation measures reliably approximate the intended task effect, i.e., the effect as perceived by humans. Such measures are referred to as "surrogate endpoints." [132]—measures that can reliably predict outcomes of real-world importance.

Pearson correlation measures the linear relationship between two sets of scores [123].

[18] After my verification of human elaborations.

Pearson rank [64] is a head-weighted, score-sensitive variant of the Pearson correlation measure. I report Pearson rank as it emphasizes agreement among the most competitive systems, and like Pearson correlation, it is sensitive to all scoring differences. There is no standard interpretation for Pearson Rank values; however, I adopt the common interpretive ranges used for the Pearson correlation coefficient, on which Pearson Rank is based.

There are three comparable pairs, where automatic measures serve as prediction values and human ratings are treated as the ground truth.

- **Comparing precision to human ratings of mention identification**: Precision (Table 5.12) is a measure that indicates how often a mention identified by the LLM is also identified by a human annotator. Human ratings are available for a subset of mentions identified by LLMs. To align these ratings with the binary nature of precision, I convert ratings into binary labels ("Binary" in Table 5.5): a rating ≥ 3 is assigned a binarized value of 1 (meaning that "Identifying this mention has some value"), and a rating < 3 is assigned 0 ("Identifying this mention has no value"). Correlation is then computed between each system's average human rating (based on 50 sampled mentions from the test split) and its precision (computed over mentions from all test sentences).

- **Comparing MNLI to human ratings of elaboration texts**: MNLI is computed only for true positive mentions' elaborations. Averaging human ratings over these TP mentions' elaborations is appropriate for correlation analysis. It remains unclear whether human ratings for elaborations on false positive mentions are consistent with those for true positive mentions. Correlation is analyzed in two ways: (1) using each system's average human rating for elaborations of true positive mentions only (fewer than 50 per system) with its average MNLI score (same amount of TP elaborations that is fewer than 50); (2) using each system's average human rating for elaborations

of true positive mentions only (fewer than 50 per system) with its average MNLI score (TP elaborations over all test sentences). As noted in Section 5.6.2, MNLI may underestimate elaboration quality due to concise or suboptimal human references and fuzzy matching limitations. To address this, MNLI scores are also converted to binary values before averaging across all elaborations: a MNLI score ≥ 0.367 (the 33rd percentile of all MNLI scores on the test split) is assigned a value of 1, and a rating < 0.367 is assigned 0.

- **Comparing BLEU-4 to human ratings of elaboration texts**: The same correlation analysis applied to MNLI scores is applied to BLEU-4 scores, without doing binarization.

The Pearson Rank [64] between precision by A2's annotations and the binary, average human ratings is 0.897, which indicates a substantial positive linear relationship. For elaboration quality, each system is sampled with 50 random mentions, of which approximately 19.2 on average are true positive (TP) mentions based on human annotation with human-written elaboration texts and available MNLI scores. The Pearson Rank between the average human ratings of TP elaborations per system, and the average, binarized MNLI scores of all TP elaborations over all test sentences (i.e., the second way of calculating correlation described above; with varying amount of TP elaborations as shown in Table 5.16), is 0.501, indicating a moderate linear relationship again using the interpretation of Pearson correlation coefficient values. When the correlation is computed between the average human ratings and the average, binarized MNLI scores of the same set of TP elaborations (i.e., the first way of calculating correlation described above), the Pearson Rank is 0.249, indicating a week linear relationship again using the interpretation of Pearson correlation coefficient values. Given the limited number of data points underlying the average human ratings and MNLI scores, the latter Pearson Rank is more prone to random variation than

the former.

The Pearson Rank between the above average human ratings, and the average, original BLEU-4 scores of all TP elaborations across the full test set (i.e., the second way of calculating correlation) is -0.146. When the correlation is computed over the above average values and the average, original BLEU-4 sores of the same set of TP elaborations (i.e., the first way of calculating correlation), the Pearson Rank is -0.057. Both indicate a random to negative linear relationship and disfavor BLEU-4 for drawing conclusions about comparative system advantages.

Returning to Section 5.6.2, the results for automatically characterizing mention identification based on F-scores and for automatically characterizing elaboration quality based on MNLI scores remain valid. Joint (GPT) is best system by those measures, and comes closest to matching human performance on contextualization. Its primary limitation, compared to human annotations, is its tendency to miss Contextual mentions that humans consider useful for conveying the intended meanings of the text.

5.7 LLM-Generated Annotations as Human Substitutes

Among the proposed approaches (Section 5.2), the Joint approach (implemented as three systems) performs the contextualization task using a detailed and explanatory instruction (see Table 5.1), most closely resembling those given to real human annotators (see Appendix D). Previous sections have shown that a Joint system can achieve over 75% human performance for both identifying mentions and elaboration quality (Table 5.11 and 5.15). Combining the challenges of scaling human annotations due to time and cost constraints (See Section 5.5), the output by the Joint system, referred to as "LLM-generated annotations", may offer a promising alternative. In this section, I explore the potential of LLM-generated annotations to evaluate contextualization, helping to answer ***RQ7***: Can

Table 5.18: Results on mention identification using LLM-generated annotations by Joint (Sonnet) and Joint (Haiku), on the validation split (results separated by slash). The systems are ordered consistently with those in Figure 5.5.

System	Overall						
	TP	FP	FN	Precision	Recall	F_1	F_2
Joint (GPT)	534/586	295/242	228/316	**0.644**/0.708	**0.701/0.650**	**0.671/0.677**	**0.689/0.661**
EntLink	272/365	239/145	492/539	0.532/**0.716**	0.356/0.404	0.427/0.516	0.381/0.442
Sequential	429/442	625/606	335/463	0.407/0.422	0.562/0.488	0.472/0.453	0.522/0.473
Coreference	276/212	212/276	489/694	0.566/0.434	0.361/0.234	0.441/0.304	0.389/0.258

LLM-generated annotation substitute for human annotations in evaluation?

5.7.1 Design

The following experiment replaces human annotations with LLM-generated annotations, and repeats the previous automatic evaluation procedure. The fuzzy matching strategy is applied as with human annotations. One cautionary point is that LLMs trained in similar ways may exhibit similar biases and be prone to over-estimation when one model's annotations are used to evaluate another. To mitigate this, recall that Joint (Sonnet) and Joint (Haiku) are two implementations of the Joint approach that use independently developed LLMs, distinct from the OpenAI GPT-4o model (version: 2024-05-13) used by Joint (GPT), EntLink, Coreference, and Sequential. I use Joint (Sonnet) and Joint (Haiku) to generate the LLM-generated annotations for evaluating the other four approaches.

Table 5.19: Results on mention identification using LLM-generated annotations by Joint (Sonnet) and Joint (Haiku), on the test split (results separated by slash).

System	Overall						
	TP	FP	FN	Precision	Recall	F_1	F_2
Joint (GPT)	463/488	261/235	216/322	**0.640**/0.675	**0.682/0.602**	**0.660/0.637**	**0.673/0.616**
EntLink	252/333	200/119	428/478	0.558/**0.737**	0.371/0.411	0.445/0.527	0.397/0.450
Sequential	436/436	613/610	243/374	0.416/0.417	0.642/0.538	0.505/0.470	0.579/0.509
Coreference	297/231	201/267	383/580	0.596/0.464	0.437/0.285	0.504/0.353	0.461/0.309

5.7.2 Results

Table 5.18 reports results for identifying mentions on the validation split. Rankings based on Joint (Haiku)'s annotations in both F_1 and F_2 align with those of annotator A2 (Table 5.11). Kendall's tau for the two F_1 rankings and for the two F_2 rankings is 1.00 (p-value: 0.083). Since this perfect rank agreement does not meet the 0.05 significance threshold due to the small number of ranked items (N = 4), so Pearson Rank [64] is also computed. The Pearson Rank is 0.995 for F_1 and 0.999 for F_2, indicating a strong linear relationship if using the interpretation for Pearson correlation coefficient. In contrast, Sonnet's rankings diverge more from those of annotator A2. Kendall's tau is 0.333 for F_1 (p-value: 0.750) and 0.667 for F_2 (p-value: 0.333), with both rankings agreeing only in Joint (GPT) being the top-ranked system. However, the Pearson Rank remain high because it is head-weighted—0.974 for F_1, and 0.980 for F_2, still indicating a substantial linear relationship if using the interpretation for Pearson correlation. Table 5.19 presents results on the test split. The F_1 system rankings are identical to those observed on the validation split, and the F_2 rankings are likewise consistent with the validation split F_2 system rankings. The Pearson Rank to annotator A2's rankings (Table 5.12) all exceed 0.94. Based on the above limited results, the answer to ***RQ7*** might be affirmative: LLM-generated annotations can potentially substitute for human annotations for evaluation of mention identification.

Table 5.20 and 5.21 reports results for elaboration quality on the validation split. Focusing on the grouping of true positive (TP) elaborations, Joint (Sonnet) rankings align with annotator A2 for both BLEU-4 and MNLI, with a Kendall's tau of 1.000 (p-value: 0.083). Although the correlation value suggests automatic evaluation using LLM-generated annotations is at least moderately, and often perfectly, consistent with human annotations, this result does not meet the 0.05 significance threshold due to the small number of ranked items (N = 4). The Pearson Rank is also computed to provide extra evidence. It is 0.248

Table 5.20: Results on elaboration quality using LLM-generated annotations by Joint (Sonnet) on the validation split.

LLM	TP		
	N=	BLEU-4	MNLI
Joint (GPT)	534	0.178	0.524
EntLink	272	0.026	0.221
Coreference	276	**0.208**	**0.610**
Sequential	429	0.164	0.486

Table 5.21: Results on elaboration quality using LLM-generated annotations by Joint (Haiku) on the validation split.

LLM	TP		
	N=	BLEU-4	MNLI
Joint(GPT)	586	**0.102**	0.477
EntLink	365	0.029	0.267
Coreference	212	0.091	**0.534**
Sequential	442	0.059	0.396

for BLEU-4 and 0.251 for MNLI, indicating a weak linear relationship if using the same interpretation. Joint (Haiku) achieves a Kendall's tau of 0.667 (p-value: 0.333) for BLEU-4 and 1.000 (p-value: 0.083) for MNLI. The Pearson Rank is 0.423 for BLEU-4 and 0.261 for MNLI, indicating a weak to moderate linear relationship if using the same interpretation. Based on the above limited results, there is no firm answer to this part of ***RQ7***: it remains unclear whether LLM-generated annotations can reliably substitute for human annotations for evaluating the quality of elaboration texts. Results for elaboration quality on the test split are shown in Appendix G, where both LLM-generated annotations yield the same system rankings on the validation and test splits, respectively.

5.8 Summary

In this chapter, I introduce the novel contextualization task, which is motivated by the challenge of making historical texts comprehensible to modern audiences. I formulate the problem as identifying (contextualizable) mentions and then generating elaboration

texts to those mentions. The first research question (***RQ5***) concerns human performance in contextualization, with a specific focus on inter-annotator agreement. To address this, I designed human annotation study. The analysis reveals that mentions identified by different people can overlap but vary slightly in terms of the number of mentions per sentence, and the lengths of those mentions. To make better use of human annotations, I propose a fuzzy matching strategy, although it introduces its own limitations. Next, the analysis reveals that human-written elaboration texts vary in nature: some offer brief explanations linked to specific vocabulary, while others convey a more holistic interpretation of an entire sentence.

RQ6 investigates how well systems perform contextualization. Human evaluation directly captures perceptions of system outputs, revealing that most of the systems perform contextualization effectively: they identify mentions that provide at least some added value to the text, and they generate elaboration texts that tend to be of good quality. Complementing the human evaluation, I also collected human annotations for contextualization and propose automatic evaluation measures that approximate the contextualization effect. For identifying mentions, F-scores serve as intuitive and informative measures, indicating that the three Joint systems outperform the other three. Among them, Joint (GPT) emerges as the overall top-performing system, exceeding 85% of the F-scores obtained by the human reference baseline. For elaboration quality, the MNLI score generally serves as a reliable proxy for estimating quality. It indicates that the Joint (GPT) system is statistically significantly the most competitive among all systems, achieving performance that exceeds 95% of human levels. Meanwhile, manual inspection confirms that most systems are capable of generating correct elaboration texts. A necessary validation study correlated the F-score and MNLI measures with human ratings, supporting the validity of the aforementioned findings based on these automatic measures.

Finally, regarding whether LLM-generated annotations can serve as a substitute for human annotations in automatic evaluation (***RQ7***), the rankings based on LLM-generated

annotations—both for identifying mentions and for elaboration quality—largely align with those based on human annotations. However, the limited number of systems being compared prevents a definitive conclusion for ***RQ7***.

Chapter 6: Rewriting for Style Transfer

Conversational agents have a fundamental requirement to look like a single, consistent agent, which ensures a more consistent user experience, rather than feeling like "a chameleon." [145]. A specific set of qualities that consistently present in generated text is referred to as the *style* of a conversational agent according to Smith et al [145]. Style is especially important for my envisioned application that represents a particular historical figure engaging in a museum interaction. After retrieving (Chapter 3) and contextualizing texts (Chapter 5) from the historical collection, the texts still have diverse styles given that the texts are transcribed spoken texts or written texts from various scenarios. Text style transfer is a task of transferring a text to a particular style, having distinctive characteristics, while preserving its original content [82]. This chapter explores text style transfer.

The goal is to rewrite texts into a target style for the historical figure engaging in conversation with a general public user. I clarify what the target style is and what it is not. Author imitation [72, 80, 152, 173] is a prior application of text style transfer where texts from a general or alternative style are transformed into the specific style of a particular author. This chapter is not an author imitation task because the source and target texts are both authored by the same individual. On the other hand, the same individual's style can exhibit significant variations just as how styles can differ between different authors. Different communication scenarios serve different communication goals, affecting the styles, not all are the same as communication with a general public, such as museum visitor, which is the target style. To build a sense about the target style, I hypothesize two intuitive traits,

using short terms like how Smith et al. specify traits for conversational agents [145].

- **Conversational:** the target style sounds more like a part of a conversation, including but not limited to elements of interaction/engagement, where texts are generally more informal, use shorter sentences and simpler language.

- **Using public language:** the target style employs language suitable for a general audience. For example, while original texts may be direct and personal, using abbreviations and slang, the transferred texts aim for clarity and accessibility by adopting a more narrative tone and removing potential comprehension barriers.

Text style transfer is often framed as a machine learning task, using corpora that represent the source and target styles. In this setting, style is defined in a data-driven manner [82], although the short descriptors often used to label styles are inherently ambiguous. Conversely, styles defined by linguistic or rule-based criteria, which thoroughly delineates what does or does not constitute a given style, are often underexplored in machine learning, primarily due to the lack of matched data for such precise definition. This chapter works with the President Reagan collection and adopts a data-driven definition for the target style. The two traits—*being conversational* and *using public language*—are reasonable, though not exhaustive, characterization of the target style. While they may align with linguistic definitions of style, in this work I use them primarily as motivating concepts and as a

Table 6.1: Analysis of document types wrt. two traits of the target style.

Document type	**Conversational tone**	**Proper public language**
Diaries	No (written monologue)	No
Letters	No (written monologue)	No
Interview transcripts	Yes	Yes
Public papers	Partial (oral monologue)	Yes
Presidential debates	Yes (oral dialogue)	Yes

framework for organizing the historical document collection, thereby laying the foundation for an empirical (i.e., data-driven) investigation of the text style transfer problem.

The two corpora representing source and target styles can be either parallel—where each sentence in one style is paired with a counterpart conveying the same information in another style, or non-parallel. [82]. Parallel corpora are valuable for training neural methods for text style transfer [82]; however, non-parallel data is more prevalent in reality [86, 142], which is the case for the Reagan collection used in this dissertation. In Sec 4.1, I introduce the President Reagan collection, which comprises public papers, personal diaries, and interview transcripts. These documents are non-parallel across types: texts from different genres (e.g., letters and speeches) do not necessarily overlap in content, even if they occasionally address similar topics (e.g., a letter and a speech both concerning *Social Security Tax*). For the experiments in this chapter, the Reagan Collection is expanded to include personal letters from the book *Reagan: A Life in Letters* [144], and transcripts of presidential debates. Given that the target style is defined as conversational and employing public language, several document types align with this stylistic goal: particularly interview transcripts, presidential debates, and to a partial extent, public papers. Public papers consist primarily of speeches but also include written statements and notices, which, while publicly delivered, are structured as monologic rather than interactive discourse. Conversely, appropriate source texts are those that exhibit opposing stylistic traits, particularly non-conversational and private forms of communication. Personal letters, as a form of written, individual expression, are suitable for this role. Table 6.1 analyzes each document type in relation to the two traits of the target style.

I process the Reagan collection to create several experimental corpora, each derived from two document types designated as representing either a source style or a target style. Each *paired corpora* consists of two sets of texts, one for each style, with some degree of topic overlap between the sets. This overlap arises naturally from the fact that all texts

originate from the same historical figure, who addressed similar topics across different communicative contexts. However, these corpora remain non-parallel in nature, i.e., there is no one-to-one correspondence. The data are split into training, validation, and test sets. Importantly, the style transfer task is treated as a standalone component in this chapter: I work directly with original texts from the collection, rather than using outputs from previous retrieval or contextualization stages.

To explore this task, I propose five prompting-based approaches, each following a general schema adapted from the SUP-NATINST framework [167]. Approaches that use different prompts, even with minor variations, such as adding chain-of-thought reasoning in the instruction field of the schema, are treated as separate systems for evaluation. I describe a procedure for selecting *comparable examples* from these corpora, defined as pairs of source-style and target-style texts that share topical content, to support prompt-based in-context learning. Evaluation is conducted through human judgments of system outputs and automatic measures assessing:

- **Target style strength**, measured via a trained style classifier, and
- **Content preservation**, measured via the publicly available MNLI entailment model.

This chapter investigates the following research questions:

- ***RQ8***: How well can systems perform the proposed style transfer task?
- ***RQ9***: How does the number of prompting examples affect a system's ability to perform style transfer?
- ***RQ10***: To what extent can the selected automatic evaluation measures approximate human preferences in style transfer?

RQ8 is addressed with both human and automatic evaluations. ***RQ10*** examines the validity of the automatic evaluation measures through a correlation analysis with human judgments.

Experiments on the validation data suggest single-turn prompting outperforms chain-of-thought prompting, and that shorter examples are more effective than longer ones. To more fully understand the effect of prompt design, ***RQ9*** explores, specifically, how different numbers of prompting examples affect performance—by comparing N-shot (N=1, 2, 3) variants across different pairs of document types. To answer ***RQ10***, human evaluation produces both pointwise and pairwise assessments, and the results are used to examine how well automatic measures reflect human preferences.

6.1 Problem Formulation

The style transfer problem is formulated as training a model to generate texts ("text generator") using a *paired corpora* that has two sets of texts. The two sets are denoted by X_1 and X_2. X_1 contains m texts of the source style, i.e., style 1. X_2 contains n texts of the target style, i.e., style 2. This formulation assumes that the text generator works in the direction of style transfer from style 1 to 2.

$$Style\,1 : X_1 = \{x_1^1, ..., x_1^m\} \tag{6.1}$$

$$Style\,2 : X_2 = \{x_2^1, ..., x_2^n\} \tag{6.2}$$

The text generator is represented by a model $\mathbb{M}$.

At test time, given a text x_1^{test} in style 1, the text generator generates a style-transferred output $\bar{x}_2^{test}$, which is expected to reflect style 2. The overline indicates that $\bar{x}_2^{test}$ is a system-generated hypothesis rather than a ground-truth reference.

$$\bar{x}_2^{test} = \mathbb{M}(x_1^{test}) \tag{6.3}$$

NATINST [114]	PROMPT SOURCE [12]	SUP-NATINST [167]
Title:{{title}} **Definition**:{{definition}} **Emphasis & Caution**:{{emphasis}} **Things to Avoid**:{{avoid}} **Positive Example 1 -** **Input:** {{p_ex1.input}} **Output:** {{p_ex1.output}} **Reason:** {{p_ex1.reason}} **Positive Example 2 -** **...** **Negative Example 1 -** **Input:** {{n_ex1.input}} **Output:** {{n_ex1.output}} **Reason:** {{n_ex1.reason}} **Negative Example 2 -** **...** **Prompt**: {{prompt}} **Taks Instance**: **Input**: {{x.input}} **Output**:	**Can you write a sentence about the topic** {{p_ex1.concepts}}? ||| {{p_ex1.target}} **Can you write a sentence about the topic** {{p_ex2.concepts}}? ||| {{p_ex2.target}} ... **Can you write a sentence about the topic** {{x.concepts}} |||	**Definition**:{{definition}} **Positive Example 1 -** **Input:** {{p_ex1.input}} **Output:** {{p_ex2.output}} **Explanation:** {{p_ex2.exp}} **Positive Example 2 -** **...** **Negative Example 1 -** **Input:** {{n_ex1.input}} **Output:** {{n_ex2.output}} **Explanation:** {{n_ex2.exp}} **Negative Example 2 -** **...** **Now complete the following example-** **Input:** {{x.input}} **Output:**

Table 6.2: Previous approaches to designing in-context instructions. SUP-NATINST serves as the basis for designing the proposed approaches in this chapter. PROMPTSOURCE does not use a highly structured format as SUP-NATINST does, but instead tailors the text sequence—including several pairs of a conditioning text and a desired output—based on specific NLP tasks and datasets, such as CommonGen [99].

6.2 Proposed Approach

Prior work on style transfer include several Masked Language Models (MLM) [103, 104, 105] that are fine-tuned on style transfer data. The Generative Pre-trained Transformer (GPT) not only shows success in standard style transfer tasks (e.g., transferring sentiment), but in transferring to non-standard, arbitrary style that can be characterized in simple words ("being melodramatic", or "has metaphor") [131]. Because my style transfer goal is non-standard, I use GPT-based models to develop proposed approaches (see Section 2.3 for background on ChatGPT). I specifically leverage the most recent version available at the time this chapter was written in 2023: GPT-3.5-turbo.

A strength of GPT models lies in their capability to perform novel text generation tasks by only following some instructions, without model fine-tuning on the particular task. This is referred to as "in-context learning" [12, 167]. Such instructions are referred to

as "in-context instructions," involving plain language task definitions and/or few-shot examples. Table 6.2 gives a comparison of notable prior work on in-context learning. SUP-NATINST [167] and NATINST [114] are two benchmarks for the design of in-context instructions, where all instructions follow a uniform instruction schema. NATINST preceded SUP-NATINST.

PROMPTSOURCE is the programming IDE and templating language to develop in-context instructions that trade-off expressivity and explicit structure [12]. PROMPTSOURCE does not use a maximally structured format like SUP-NATINST does. Instead, it allows users to write expressive prompts using a two-part structure of conditioning text and the desired output, to fit every prompting need on the specific NLP tasks and datasets, such as the example task shown in Table 6.2, CommonGen [99]. The CommonGen example does not imply that other prompts written using PROMPTSOURCE should follow the exact same structure.

Because this work focuses on a single, specific task, style transfer, I follow the general schema of SUP-NATINST [167] to implement five proposed approaches (with their instructions shown in Table 6.3). Table 6.4 through 6.7 presents the examples used in the prompts corresponding to each paired corpora. Corpora are ordered based on their results order in Section 6.5.

Approach 1: NOEX The first proposed approach is the plain re-implementation of the prior SUP-NATINST benchmark [167]. I follow the general-purpose schema, and design a new in-context instruction for the style transfer task. NOEX, as a result, has a plain language task definition, but without any few-shot examples. It is a simple baseline.

Approach 2: SUP-NATINST The second proposed approach, referred to as SUP-NATINST, has a basic prompting format that consists of an instruction and one or a few positive ex-

Prompting Elements	NoEx	Sup-NatInst	PreparedDescr	StyleReasoning	EndGoal
Task Background	In this task, we ask you to read the input sentence that is from a letter written by Ronald Reagan, then [continue to Task Description]				
Task Description	express it in an output that speaks like something Ronald Reagan would say when he is **responding to a typical museum visitor**.		express it in an output that speaks like something Ronald Reagan would say when he is **diplomatic, contemplative, strategic, assertive, and pragmatic**.	**pinpoint the segments of the input that does not fit into a style that** Ronald Reagan would say when he is diplomatic, contemplative, strategic, assertive, and pragmatic.	pinpoint the segments of the input that does not fit into a style that Ronald Reagan would say when he is **responding to a typical museum visitor.**
Chain-Of -Thought Instruction	N/A			Then, starting with the statement "The rewritten version of the input is", provide your rewrite of the input to fit the diplomatic, contemplative, strategic, assertive, and pragmatic style.	Then, starting with the statement "The rewritten version of the input is", provide your rewrite of the input to fit into the style of responding to a typical museum visitor.
Limitations	The preferred output length is no longer than two times the length of the original input sentence.				
Positive Example(s) (One or many)	N/A	Input: `pos_example_input` Output: `pos_example_output`		Input: `pos_example_input` Output: `pos_example_explanation` The rewritten version of the input is `pos_example_output`	

Table 6.3: Proposed approaches involving different instructions. From let to right, instruction differences are in bold font. Texts in typewriter font represent example variables containing long texts, whose full contents are in the next four tables (6.4 to 6.7).

amples. Unlike prompting for classification tasks, where negative example's output can be "no," a text rewriting task does not have a clear choice for negative examples. Hence, Sup-NatInst does not have negative few-shot examples.

Approach 3: PreparedDescr The third proposed approach replaces the descriptive style requirement in Sup-NatInst and NoEx (*"responding to a typical museum visitor"*) with several declarative adjectives: *"diplomatic, contemplative, strategic, assertive, and pragmatic."* The adjectives are intended to prompt the LLMs to express the input sentence with these adjective characteristics. The adjectives were obtained by asking ChatGPT to analyze several texts from an interview transcript to describe the style characteristics that Reagan's speech would have.[1] This interview transcript was selected for its conversational tone and its use of proper public language suited to the anticipated target style.

[1]The transcript is titled as *"Interview With Robert L. Bartley and Albert R. Hunt of the Wall Street Journal on Foreign and Domestic Issues."* The question about style differences is phrased as *"Can you describe the style difference between the two texts below?"*

Using this set of five adjectives has limitations in how well it represents the target style, particularly when the target style is complex and not easily captured by adjectives alone. For example, I also asked ChatGPT to analyze the style characteristics of a Reagan letter to an American medical student evacuated from Grenada. The letter is a source style document and not conversational, but the audience is a member of the general public, so it shares some characteristics with the target style. ChatGPT suggested the adjectives *"conversational, warm, informal, personable, and politically strategic"*—reflecting some characteristics of the target style.

Approach 4: STYLEREASONING Chain-of-thought prompting divides a complex task for an LLM into simpler sub-tasks that it can solve. For style transfer, I divide the task into (1) pinpointing segments that do not fit a style goal, and (2) providing a rewrite of the input that fits the style goal. The LLMs will start with an explanation of (1), and then use the phrasing *"The rewritten version of the input is"* to indicate the end of explanation, and provide a rewrite output to (2). Here, the style goal is specified as *"a style that Ronald Reagan would say when he is diplomatic, contemplative, strategic, assertive, and pragmatic."*

Approach 5: ENDGOAL This approach adopts chain-of-thought prompting as Approach 4, but whose style goal is descriptive, i.e., the same as SUP-NATINST, i.e., *"when he is responding to a typical museum visitor."*

6.2.1 Selecting Comparable Prompting Examples

Within the corpora, examples that belong to the source and target styles have no exact content overlap. Given that the documents accumulate over the same time period and involved the same historical figure, they may still share topical overlap in certain cases.

The potential presence of topical overlap motivates me to identify such topical alignment, and use the aligned examples from two styles to prompt proposed approaches in a way that illustrates the style transfer effect. Prior work refers to corpora with this kind of topic alignment quasi-comparable corpora, and call texts that are not necessarily synonymous but share the same topic comparable [59]. Jin et al. further affirm that pseudo-parallel corpora resemble supervised data [82].

The starting point to find topical alignment are the target texts from a paired corpora, used as search queries, to find whether some source texts are comparable, i.e., having roughly the same content with them, such that they pair as comparable pairs. Using the target texts as the search queries, rather than the reverse, enables the resulting comparable example pairs to better reflect the characteristics of the target texts. Using the training split of the source and target texts from each paired corpora, I build an index consisting of all paragraphs from the source document type. For each target text, the set of non-stopwords forms a long query that is used to retrieve from the index, resulting in top three scoring paragraphs. The matching algorithm is the term-based BM25 method. Then, the matching paragraphs are segmented into sentences, from which a more effective but less efficient semantic matching model BERTScore [178] is used to measure content similarity between the target text and each sentence in a matching source paragraph. The cost of the BERTScore is limited by the number of matching paragraphs and the number of sentences in paragraphs. The highest-scoring source-target sentence pairs were manually screened by me to verify content similarity and then sorted by the length of the source sentence. There are two variants of comparable example(s) for each approach: long and short. For long comparable examples, the source sentence must exceeds 22 tokens if split by whitespace. For short examples, the source sentence must not exceed 22 tokens if split by whitespace. Note that finding topically aligned examples is a manual, one-time effort. I use up to three pairs of comparable examples for each length variant and for each paired corpora, in an

Examples	pos_example_input	pos_example_explanation	pos_example_output	Retouch
Paired Corpora: Letters–Student Interviews, Long Examples				
Example 1	Also much of that aid is to help those countries have a defense capability of their own which is less costly than if we had to provide that defense with our own forces.	The phrase **"which is less costly than if we had to provide that defense with our own forces"** emphasizes financial concerns in a way that may sound transactional rather than strategic, lie a cost-saving alternative to direct U.S. military involvement. **"To help those countries have a defense capability of their own"** is somewhat indirect and less assertive.	And some of that aid, also, is, very practically, military aid to those countries so that they can defend themselves.	In the output, "those countries" were originally referred to as "them."
Example 2	It is a program to use tax incentives to bring business and industry into depressed inner city neighborhoods to provide jobs and opportunity for the people there.	The phrase **"bring business and industry into depressed inner city neighborhoods"** implies a passive approach. The phrase **"depressed Inner City Neighborhoods"** may sound negative or overly bleak.	It's a program of using tax incentives at every level of government for businesses that will go into rundown areas, establish themselves there, and then hire the people that are there presently unemployed and getting welfare, and tax incentives to make this worthwhile.	In the output, "go into rundown areas" were originally referred to as "go in". "rundown areas" was an expression in the immediate prior context.
Example 3	The other purpose of our foreign aid is to help these lesser developed countries get their economies going so they can be better customers of ours.	The phrase **"so they can be better customers of ours**¨ makes foreign aid sound purely transactional and like a self-serving economic investment rather than a diplomatic and strategic effort. **"Get their economies going"** is informal and lacks the kind of dignified, strategic phrasing Reagan would typically use.	The aid that we're giving to foreign countries is aid to the lesser developed countries so that they can take their place in the world of nations with an economy in which they can have a living standard that is improved.	In the input, "The other purpose of our foreign aid" was originally "The other". The expansion was based on immediate prior context.
Paired Corpora: Letters–Student Interviews, Short Examples				
Example 1	All of us find apartheid repugnant but so do many people in South Africa.	The phrase **"but so do many people in South Africa"** is vague and passive, lacking a strong strategic or action-oriented perspective. **"a large element in South Africa"** sounds more substantial and deliberate. Saying **"many people"** does not specify who is opposing apartheid or what they are doing about it, making the statement less impactful.	There is a large element in South Africa that finds apartheid as repugnant as we do and is trying to do something about it.	N/A
Example 2	But, we inherited a government that was increasing in cost 17.5 percent a year.	The phrase **"we inherited a government"** suggests a lack of control, implying that the previous administration is to blame without emphasizing action. **"Increasing in cost"** is somewhat vague and less precise compared to **"increasing its spending"** which directly addresses fiscal responsibility.	But when we came here, the Government was increasing its spending by 17 percent a year.	N/A
Example 3	We have also reduced inflation from 12.4 percent to 3.7 for the last six months.	The phrase **"We have also reduced inflation"** implies direct control over inflation, whereas Reagan's style often acknowledged broader economic forces and policies influencing such changes. **"For the last six months"** is slightly ambiguous because it does not explicitly clarify that the lower rate is an ongoing trend.	Inflation, which was 12.4 percent, has for the last 6 months been running at only 3.7 percent.	In the output, "3.7 percent" was originally "four-tenths of 1 percent," which is a content mismatch.

Table 6.4: Formatting instructions with different content under a shared schema. This table presents prompting examples for the Letters–Student Interviews corpora. There are two variants of positive example(s) for each prompting approach: long and short. The column "Retouch" notes on changes to the original text to create comparable examples.

Examples	pos_example_input	pos_example_explanation	pos_example_output	Retouch
Paired Corpora: Letters–Interviews w/ Reporters, Long Examples				
Example 1	With the Democrats back in charge in both Houses of Congress, there are too many of them just waiting for a crack in the dam to flood us with tax changes to pay for a pack of new spending ideas.	Regarding the segment **"With the Democrats back in charge in"**, while Reagan criticized excessive spending, he typically framed policy disagreements in a more diplomatic way, avoiding outright partisan blame. The metaphor of a dam breaking is dramatic and negative. **"A pack of new spending ideas"** sounds dismissive and lacks the pragmatic, results-oriented tone Reagan often used.	There are factions in there that just want to keep on spending in the Congress.	In the input, "a pack of new spending ideas" were followed by "not deficit reduction," which were removed.
Example 2	Even people on welfare or at the poverty level of earnings have several hundred dollars more in purchasing power than they would have had if inflation had remained at the 1980 level.	The phrase **"even people on welfare"** may sound somewhat dismissive or divisive, which is not aligned with Reagan's typically optimistic and inclusive rhetoric. While factually meaningful, **"several hundred dollars"** sounds vague and lacks the weight of a more strategic economic framing.	Today, the person at the average family, the median income has $3,300 more in purchasing power than they would have had if we had stayed at the same tax and inflation rates of 1980.	N/A
Example 3	It is our hope that together these four nations, with the help of the private sector, can seek to alleviate the conditions that make so many people, particularly from the Caribbean states, want to come here.	**"It is our hope"** sounds weak and uncertain compared to a more decisive alternative. **"seek to"** introduces an unnecessary hedge, making the action sound less direct and effective. The expression **"make so many people, ... want to come here."** is passive and indirect in describing migration. Reagan often framed challenges in terms of opportunity and mutual benefit, rather than focusing on the issue of people.	What we envisioned with this is that together these four nations, with the help of the private sector, does all that it can to improve and see the conditions that affect so many people.	In the output, "together these four nations, with the help of the private sector" was originally "the government", and "the conditions that affect so many people" was originally "that there is no hunger."
Paired Corpora: Letters–Interviews w/ Reporters, Short Examples				
Example 1	Truth was Social Security didn't make it to July of '83.	The phrase **"Truth was"** is abrupt and informal. The wording around **"Social Security didn't make it to July of '83"** is vague and imprecise, almost as if Social Security is a person who failed to reach a goal, rather than a financial program running out of funds. Overall, it is blunt and could be seen as overly casual or lacking the gravitas typical of a statesman discussing such a serious issue.	But we were faced with Social Security nearing bankrupt in July of '83.	In the output, "nearing" was originally "bellying up."
Example 2	Right now we continue to try to persuade Congress to give us what we call "enterprise zones."	The phrase **"Right now"** is somewhat informal and overly immediate. Reagan often framed policies in a broader historical or long-term perspective rather than emphasizing urgency in a casual way. The phrase **"continue to try to persuade"** sounds hesitant and lacks assertiveness. The phrase **"to give us"** makes the administration sound like passive recipients rather than leaders advocating for policy. The phrase **"what we call"** suggests uncertainty or informality, which does not align with Reagan's clear rhetorical style.	This is why we have for a couple of years now been trying to get the enterprise zone legislation through the Congress.	N/A
Example 3	We have more people employed than ever before.	The phrasing **"We have more people employed"**, while factually correct, is somewhat simplistic and lacks the depth or strategic framing typical of Reagan's speeches or letters. The phrase "**than ever before"** is broad but imprecise.	Today the highest percentage of the potential pool consisting of all people, between the ages of 16 and 65, is employed than has ever been employed in our entire history.	In the output, "the potential pool... and 65," was originally "that potential pool."

Table 6.5: This table presents prompting examples for the Letters–Interviews w/ Reporters corpora.

Examples	pos_example_input	pos_example_explanation	pos_example_output	Retouch
Paired Corpora: Letters–Presidential Debates, Long Examples				
Example 1	I still believe if a defense can be found it is far more civilized than to continue relying on the threat of destroying millions of lives in retaliation for an enemy attack.	**"I still believe"** sounds personal and tentative. The rewritten version removes this and instead poses a rhetorical question. While **"civilized"** conveys a moral argument, it is somewhat abstract. The rewritten version replaces it with "far more humanitarian." The phrase **"to continue relying on the threat of destroying millions of lives in retaliation"** is long and somewhat grim.	If such a defense could be found, wouldn't it be far more humanitarian to say that now we can defend against a nuclear war by destroying missiles instead of slaughtering millions of people?	The input begins with "And (if such a...)", which was removed.
Example 2	We have expressed a willingness to help Lebanon with our multinational forces until such time as Lebanon forces can demonstrate they are capable of guarding the border.	The phrase **"expressed a willingness to help"** is tentative and passive. The phrasing **"until such time as"** is formal and bureaucratic, making it sound less pragmatic and conversational. This phrase **"demonstrate they are capable of guarding the border"** is vague and conditional, implying an indefinite commitment.	We went in with the multinational force to help remove, and did remove, more than 13,000 of those terrorists from Lebanon.	The input begins with "Then (we went in...)", which was removed.
Example 3	I was talking about the Lebanon situation and our effort to persuade Israel, Syria and the PLO to get out and give the Lebanese a chance to reestablish sovereignty over their own land.	The phrasing **"our effort to persuade Israel, Syria, and the PLO to get out"** makes the U.S. role sound like a straightforward negotiation effort. The phrase **"give the Lebanese a chance to reestablish sovereignty over their own land"** is somewhat informal and does not reflect a strong, assertive policy stance."	And then the Government of Lebanon asked us back in as a stabilizing force while they established a government and sought to get the foreign forces all the way out of Lebanon and that they could then take care of their own borders.	N/A
Paired Corpora: Letters–Presidential Debates, Short Examples				
Example 1	I believe we have much to gain by refusing to submit to blackmail.	The phrase **"submit to blackmail"** is emotionally charged and adversarial. While Reagan could be assertive and firm, he often avoided language that sounded accusatory, inflammatory, or morally absolute, and tended to favor strategic, values-driven framing such as moral leadership.	I don't think that our record of human rights can be assailed.	N/A
Example 2	As I say, the shah has a long record of friendship and service to the United States.	The phrase **"as I say"** is colloquial and repetitive, and can sound imprecise or casual. A more Reagan-esque version might omit it altogether or reframe it.	I did criticize the President because of our undercutting of what was a stalwart ally—the Shah of Iran.	The input begins with "In this case, (as I...)" and the output begins with "Morton, (I did...)" Both were removed.
Example 3	I asked them if it was possible to devise a weapon that could destroy missiles as they came out of their silos.	**"I asked them if in their thinking"** is somewhat informal and roundabout. It makes the speaker sound like a passive inquirer rather than a leader proposing a vision. The phrase **"destroy missiles as they came out of their silos"** has a blunt and aggressive tone. Framing it as a defensive measure rather than an offensive action would make it more palatable in a diplomatic context.	I propose that we research to see if there isn't a defensive weapon that could defend against incoming missiles.	N/A

Table 6.6: This table presents prompting examples for the Letters–Presidential Debates corpora.

Examples	pos_example_input	pos_example_explanation	pos_example_output	Retouch
Paired Corpora: Letters–Public Papers, Long Examples				
Example 1	We can't continue to urge the reestablishment of the family as an answer to society's breakdown while we bypass the family when it suits our convenience.	The phrase **"urge the reestablishment of the family as an answer to society's breakdown"** This phrase is ideologically charged and prescriptive, implying that restoring the family is the sole remedy for society's problems. While Reagan valued family, his rhetoric was more inclusive and forward-looking rather than prescribing a single solution. The segment **"while we bypass the family when it suits our convenience"** is accusatory and highlights hypocrisy in a manner that can seem judgmental.	We must do what we can to reverse the trend of abused children becoming abusive parents.	N/A
Example 2	We're going to try to create an economic infrastructure in their countries so there will be opportunities and jobs and a decent living for the people.	The phrase **"try to create"** feels a bit less assertive and more tentative. The phrase **"in their countries"** is vague and paternalistic. While positive, the segment **"so there will be opportunities and jobs and a decent living for the people"** is overly simplistic and lists benefits in a way that lacks the strategic and nuanced language characteristic of Reagan.	The Caribbean Basin Initiative will address the underlying economic and social problems that are retarding the development of the Caribbean Basin States.	N/A
Example 3	The difference is that in El Salvador the government has elections and is trying to be democratic while in Nicaragua the government has embraced Castro and the Soviets.	The segment **"has embraced Castro and the Soviets"** is overly blunt and ideologically charged.	Unlike El Salvador, the Nicaraguans don't want international oversight of their campaign and elections.	N/A
Paired Corpora: Letters–Public Papers, Short Examples				
Example 1	We are going to engage them in negotiations to reduce nuclear weapons.	The phrasing **"engage them in negotiations"** is rather neutral and passive. He would likely choose language that underscores a bold, proactive approach. The phrase **"reduce nuclear weapons"** is vague and doesn't capture the depth of strategic intent. He would be more specific, emphasizing not just reduction, but a radical decrease in both the scale and potential destructiveness.	Today we set out on a new path toward agreements which radically reduce the size and destructive power of existing nuclear missiles.	N/A
Example 2	Those demagogs who scream our tax reform benefited the rich are talking through their hats.	The label **"Those demagogs"** is overly inflammatory and confrontational. Another phrase **"who scream our tax reform benefited the rich"** is emotionally charged and accusatory. Reagan would lean toward framing issues in a more thoughtful and balanced manner rather than resorting to pejorative and hyperbolic language. The idiom **"talking through their hats"** is informal. The sentence could benefit from a more refined phrasing that criticizes unfounded rhetoric without sounding dismissive in a colloquial manner.	And those who say it can are merely engaging in empty rhetoric as usual.	N/A
Example 3	We must all work together to bring an end to terrorism.	The phrase **"We must all work together"** is quite generic and lacks the specificity he sometimes used when rallying allies. The phrase **"to bring an end to terrorism"** is somewhat vague and broad.	The United States calls upon other nations to join with us in isolating the terrorists and their supporters.	N/A

Table 6.7: This table presents prompting examples for the Letters–Public Papers corpora.

effect to better support generalization.

For content that differ between the target text and the selected source text, a retouching step tries to edit the content, while retaining the stylistic features. The retouching step works under the assumption that stylistic features and contents are not coupled, i.e., separable within a text, as seen in some prior style transfer tasks such as verb tense or Yelp review sentiment transfer [82]. However, for one paired corpora, Letters–Public Papers, I did not perform retouching due to a substantial content gap between the optimal comparable pairs, which may reflect an inherent characteristic of this paired corpora.

6.3 Data Collection

The problem formulation requires a paired corpora consisting of two sets of texts in order to develop[2] a style transfer model. Inference is then using the model on the source set of texts. I create these corpora from selected document types from the Reagan collection. This section provides details of this process.

The Reagan collection introduced in Chapter 4 serves as the starting point, from which a prototype retrieves content to answer user questions about Ronald Reagan. Table 6.8 is an updated version of the Reagan collection that includes three new document types: letters, student interviews, and presidential debates. Letters come from a book of Reagan's letters, with annotations [144]. The student interviews and presidential debates are in a text transcript format from the Ronald Reagan Presidential Library and Museum website.[3] Among these document types (Table 6.1), I characterize each by two aspects that distinguish the source style from the target style. I anticipate four corpora to be possible for the style transfer modeling: (1) Letters–Student Interviews, (2) Letters–Interviews w/ Reporters, (3) Letters–Presidential Debates, and (4) Letters–Public Papers.

[2] I use "develop" to refer to prompt development rather than training or fine-tuning a model.

[3] `https://www.reaganlibrary.gov/`

Table 6.8: Statistics of the updated President Reagan collection, which now includes Reagan letters and presidential debate transcripts. Data are split at the document level to ensure that sentences from the same document do not appear in multiple splits. Document types with small numbers of documents tend to have long documents.

Document Type	# Document				# Paragraphs			# Sentences		
	Total	Train	Valid	Test	Train	Valid	Test	Train	Valid	Test
Letters	1,058	755	151	152	2,811	562	563	8,846	1,909	1,918
Student Interviews	7	5	1	1	42	8	9	609	107	96
Interviews w/ Reporters	247	176	35	36	677	125	124	6,672	1,315	1,196
Presidential Debates	4	2	1	1	22	4	5	360	64	102
Public Papers	8,147	5,819	1,163	1,165	-	-	-	256,992	48,597	50,538

Table 6.9: Amount of texts for each paired corpora after balancing.

Paired Corpora	Train	Validation	Test
Letters–Student Interviews (primary)	462	83	82
Letters–Interviews w/ Reporters	1,768	395	360
Letters–Presidential Debates	326	55	88
Letters–Public Papers	1,768	395	360

Processing Letters The efforts to process letters are briefly described in Chapter 5, and covered in more detail here. The book *Reagan: A life in letters* has 1,416 pages containing letters written by Ronald Reagan throughout his lifetime. For research use, and without public distribution, I convert the book into a text format. Then I extract individual letters by segmenting based on salutations as the beginning boundaries and several variants of Reagan signatures as the ending boundaries. There are 1,058 letters, which are divided into train, validation, and test folds in a $5:1:1$ split. Each letter contains one or more paragraphs, where a paragraph is usually cohesive on one topic and has one or more sentences. The end of a paragraph is identified based on the heuristic of a line whose character count is less than 74, which is the minimum number of characters for a full line in the book. I segment the letters into separate paragraphs based on the paragraph boundary. Then, I segment the paragraphs into sentences using the spaCy library.[4] A sentence is an instance for style

[4]Each sentence can be traced back and connected with preceding sentences within the same paragraph.

transfer in my experiments.

Processing Interviews The Reagan collection contains 254 interview transcripts, drawn from interview-formatted webpages on the Ronald Reagan Presidential Library and Museum website. They are divided into interviews with news reporters (247; see Chapter 4, Section 4.1), and student interviews (7). Student interviews comprise a small subset of seven interviews whose titles contain the keywords *"Students"* and *"Question-and-Answer Session."* These are recognized as presidential Q&A sessions with high school students, rather than student interviews—although I adopt the latter term in this dissertation for consistency. The 247 interviews with news reporters are divided into train, validation, and test folds in a 5 : 1 : 1 split. The 7 student interviews are divided into 2, 1, and 1 document for train, validation, and test splits. Despite the single document in one split, student interviews are specially arranged sessions that last more than one hour, whose transcript document thus tends to be longer. Student interviews provide the closest examples to the target style, i.e., the intended style transfer goal: the historical figure engaging in conversations with the general public, such as a user with a high school education, and that the QA session is a public conversation.

I segment each interview transcript based on the website HTML's element tree path to create paragraphs, and use any leading speaker identity, such as "Mr. President" to identify utterances by Ronald Reagan. I then create text units for style transfer from the paragraphs. For interviews, each unit consists of two consecutive, non-overlapping sentences. Interview utterances are often highly colloquial, short, and filled with filler phrases, so this heuristic better matches the information content of single sentences in letters.

Processing Presidential Debates Presidential debates are crawled from the website `www.debates.org` for four presidential debates between Ronald Reagan and another pres-

Chapter 5 particularly concerns with the letter sentences and their preceding sentences as instances for contextualization.

idential candidate.[5] They are divided as 2, 1, and 1 documents for train, validation, and test. Given that a presidential debate also consists of turn-taking spoken utterances by multiple speakers, the paragraph segmentation follows the same procedure as for interviews. Paragraph boundaries are determined based on the HTML element tree paths from the source website.

Processing Public Papers There are 8,147 public papers from the original Reagan collection. I divide them into train, validation, and test folds in a 5 : 1 : 1 split. Despite the public papers having HTML's element tree path segmented paragraphs, I create text units (sentences) directly from the document, without first segmenting them into paragraphs, as public papers were not used for contextualization in my experiments Therefore, the number of paragraphs are not listed in Table 6.8.

Filtering and Balancing Corpora From this point onwards, I refer to any text unit derived from the above-mentioned document types and intended for style transfer as a *text*. After obtaining texts from the four document types, I filter them to retain only these that may have style transfer importance. There are three cases to be filtered out. First, examples that have too few (less than ten) tokens (after splitting by whitespace) are filtered out. This includes salutations such as "Dear Dr. Smith." Second, appreciation statements are filtered out, but only from letters. Based on my manual screening, appreciation statements often contain pronouns, e.g., *"A letter like yours helps to make me feel I made the correct decision."* or *"I'm also most grateful for the inscription—I hope I can live up to it."* I filter appreciation statements through the regular expression match of a lexical set of pronouns *"you," "me," "I," "my," "your,"* and *"yours."* Using a lexical set to filter is efficient but aggressive. It achieves a 94.8% precision at filtering out appreciation statements, compared to the silver standard of asking the OpenAI GPT-3.5-turbo model whether a text is an

[5] On Sept 21 and Oct 28 of 1980, and Oct 7 and 21 of 1984.

appreciation statement.[6] That silver standard would be costly to scale to all letter texts. Third, the first sentence of a paragraph is not used because they do not have the context to fit the problem formulation in Chapter 5.

Finally, I balance the paired corpora to have the same number of texts for both types, by down-sampling based on the the amount of texts in the fewer document type. Which texts are omitted during down-sampling is determined by processing order, based on filenames or the letters' sequence in the book. Doing so mitigates the risk of the trained model that I use for automatic evaluation (see Section 6.5) having a bias in favor of one class over another. See Table 6.8 for the number of texts before filtering or balancing, and Table 6.9 for the number of texts after filtering and balancing.

6.4 Human Evaluation

Research in text rewriting, including text style transfer, lacks a standardized evaluation protocol [120, 172]. Researchers generally evaluate text style transfer using a mixture of human and automatic evaluation, as seen in politeness transfer [103], and stylized utterance generation for desirable dialog quality [145]. There are several dimensions along which to evaluate the overall success of style transfer, and for which several measures have been designed [82]: (1) Strength of the target style ("style strength"), (2) Preservation of essential source content ("content preservation") [173], (3) Fluency/naturalness, which refers to general language quality. Surrounding these dimensions, I describe human evaluation in this section and automatic evaluation in Section 6.5. As in Chapter 5, fluency is not evaluated in

[6]The prompt to GPT-3.5-turbo model is *"You are given a text from a letter. You need to determine if the text primarily expresses appreciation and/or validation, or if primarily expresses something else that is neither appreciation or validation. Indicate your answer as "Yes" or "No" respectively. Example 1 - Input: A letter like yours helps to make me feel I made the correct decision. Output: Yes Example 2 - Input: And let me tell you, Henry made it very clear that negotiating was the only option because of that reason. Output: No Example 3 - Input: I'm also most grateful for the inscription-I hope I can live up to it. Output: Yes Now complete the following example - Input:* {{`x.input`}} *Output:"*

this chapter. This is a minor limitation, as large language models (LLMs) are well known for generating natural and fluent text [25], and all proposed approaches rely on prompting LLMs to generate texts without model fine-tuning. Between the two types of evaluation, it is valuable to first present human evaluation that directly probes human preferences on system outputs. That can inform automatic evaluation. By contrast, automatic evaluation uses measures to model the problem, sometimes relying on simplified assumption and revealing limited facets that may fail at diverse phenomenon of language. Human evaluation also comes with drawbacks, however, such as its high cost and difficulty in replication.

6.4.1 Design

For text rewriting tasks, there are three principal ways in which people can provide preferences in human evaluation [82]:

- Pointwise scoring, where a human evaluator (shortened as "evaluator") examines an output and decides whether the output satisfies some criterion, without comparing with other outputs, which is straight-forward.
- Pairwise comparison, where the evaluator examines two rewrites of the same text at once, and decides whether one rewrite is better than the other.
- Ranking, where evaluator examines more than two rewrites of the same text at once, and ranks the outputs based on some criterion.

I collected human preferences using a mixture of point-wise scoring and pairwise comparison. The same evaluator from Chapter 5 (see Section 5.5 for demographic information), after completing their work on Chapter 5, was asked to provide their preferences for a random sample of style transfer output, at the same compensation rate. An Excel spreadsheet

was used for this evaluation. During the training session conducted using Zoom, the evaluator developed guidelines for rating rewrites on two scales from 1 to 5 (two 5-point Likert scales). One assesses style strength, and the other assesses content preservation.

For style strength, the evaluator first reviewed some source and target texts from the primary corpora: 40 from letter texts in the validation split, and 40 from student interviews in the validation split. After building a sense for how styles between the two sets differ from each other, they summarized stylistic features that distinguish between the two sets in their own words. Then, they were instructed to translate those stylistic features into a guideline for the assignment of style strength rating labels on an ordinal scale from 1 to 5. I provided the following instruction: *"Intuitively, a rating of 5 should represent the strongest presentation of the style, while a rating of 1 should indicate little to no presence of the style. You are free to define the specific wording and decision boundaries based on your discretion."* They then examined a mixture of target texts and style-transferred output from NoEx, Sup-NatInst, and EndGoal on the validation split, iterating the specific wording and decision boundaries between each rating scale, and could revise the guideline until satisfied.[7] For content preservation, they were instructed to see 45 original and style transferred text pairs sampled from the same three systems from the validation split. They examined whether the text pairs have the rewritten text preserving the content from the original text. They then created a guideline for the assignment of content preservation rating labels.

The evaluator worked on pointwise scoring first. The test split of the primary corpora, Letters–Student Interviews, contains 82 source texts (see Table 6.9). Given five approaches that totals nine systems (each approach has long and short variants of prompting examples, except NoEx), I reduce the scale by randomly sampling 30 source texts by the same three

[7]Including the target texts during training and actual human evaluation helps validate the evaluator's performance, as the target texts are assumed to exhibit perfect style strength, for which the evaluator should assign a label of 5.

(a) Labels for the 5-point rating scale on style strength.

Point	Label
5	This text sounds like the spoken word. It may have characteristics such as emotion markers, first-person perspective, and simplified language.
4	This text is similar to a spoken style but contains minor differences (e.g., rhetorical devices, passive voice, or complex vocabulary).
3	This text bears some resemblance to a spoken style.
2	This text does not resemble a spoken style.
1	This text resembles a written style.

(b) Labels for the 5-point rating scale on content preservation.

Point	Label
5	The content is well-preserved.
4	The content was preserved but has some minor flaws (inferred meaning, omitted but useful details, slight tonal change, or other potential inaccuracies).
3	There is additional content, and I don't know how to judge the content accuracy.
2	The content may be impaired.
1	The content is definitely impaired, wrecked, broken, etc., e.g., its meaning changed completely.

Table 6.10: Labels for the two aspects: style strength and content preservation. Wordings are not grammatically perfect.

systems: NoEx, Sup-NatInst (Long), and EndGoal (Long). The 30 source texts each have style-transferred rewrites from three systems, resulting in 90 rewrites. In addition, I included 30 texts from the student interviews that were unpaired with the source texts. In total, 120 texts were rated for style strength, presented in random order.

For content preservation, the 90 source texts and their rewrites were evaluated as 90 pairs. Pairs were grouped by the 30 source texts, with the three rewrites for each source text randomly shuffled within the group. This arrangement reduces cognitive effort in understanding the source text.

After completing pointwise scoring for 120 texts on style strength, and 90 pairs on content preservation, the evaluator was trained for pairwise comparison. Training involved 10 pairs for style strength and 10 pairs (with their corresponding source texts) for content

preservation, all drawn from the validation split. For each pair, the evaluator was asked: *"which text (left or right) is better based on your guideline about style strength/content preservation. Type 0 for left and 1 for right."* Ties could be specified with a justification, but were to be used only when necessary.

Following training, the evaluator worked on 50 pairs for style strength and 50 pairs (with their source texts) for content preservation, sampled from the test split and were not used for pointwise scoring. Each pair consisted of one rewrite by SUP-NATINST (Long) and one rewrite by ENDGOAL (Long), both doing style transfer on the same source text. The order of the pairs was then randomly shuffled.

Let a, b, t denote the counts corresponding to preference for SUP-NATINST, preference for ENDGOAL, and no preference. A common test to examine whether humans actually prefer one system over the other is the two-tailed sign test [49, 52, 91], where the probability that humans prefer either one is 0.5 (equal).[8] The null hypothesis is that SUP-NATINST is no different from ENDGOAL in their style strength as perceived by human. A binominal distribution, with the number of successes defined as $x = b + 0.5t$ and the number of trials as $n = a+b+t$, was used to estimate probability under the null hypothesis. This probability was then compared to the significance level of 0.05.

[8]Dras recommends Log-Linear Bradley-Terry models over the classical sign test, primarily due to the ambiguous effect of various ways of modeling of ties [49]. In my human evaluation, ties are rare.

Table 6.11: Median and mean pointwise human ratings of different systems, each on a sample of 30 rewrites from the test split. A set of 30 target texts from the test split was also rated as a reference. Style strength was rated without reference to the corresponding source text, while content preservation was rated with respect to the source text.

System	Style Strength		Content Preservation	
	Median	Average	Median	Average
Target texts (student interviews)	5	4.63	-	-
NOEX	5	4.60	3	2.90
SUP-NATINST (Long)	5	4.53	4	4.07
ENDGOAL (Long)	4	3.93	3	2.80

6.4.2 Results

Table 6.10 lists the rating scales for style strength and content preservation, developed by the human evaluator. I calculate the median pointwise score for each system, separately for style strength and content preservation. For style strength, there is a reference, median pointwise score on 30 unpaired target texts, i.e., from student interviews. The median score is reported in line with the setup of ordinal rating scales, rather than evenly spaced numerical scales. The median is 5 on style strength for the target texts, NOEX, and SUP-NATINST, and 4 for ENDGOAL. The median is 4 on content preservation for SUP-NATINST, and 3 for both NOEX and ENDGOAL. The median alone does not offer strong differentiation among systems. I also report the average scores as supplementary evidence. Table 6.11 summarizes the median and the average of pointwise scoring.

Next, for pairwise comparison, and on style strength, the counts corresponding to preference for SUP-NATINST, preference for ENDGOAL, and no preference are 35, 14, and 1 respectively. Using the two-sided sign test, the p-value is 0.003,which is below the significance level at 0.05. Hence I reject the null hypothesis, and conclude that SUP-NATINST is better than ENDGOAL on style strength. For pairwise comparison on content preservation, the corresponding counts are 26, 21, and 3. Using the two-sided sign test, the p-value is 0.479. Hence I cannot reject the null hypothesis, and therefore conclude that SUP-NATINST does not differ significantly from ENDGOAL on content preservation.

6.5 Automatic Evaluation

Automatic evaluation has advantages in terms of scalability and reproducibility. A notable drawback is the potential bias associated with the chosen measures that model the task. To mitigate this issue, it is advisable to diversify the measures for different dimen-

sions, and/or use several measures for one dimension [82]. The measure choices that correspond to each dimension are summarized in Table 6.12. Using these measure choices, I respond to ***RQ8*** regarding system performance.

6.5.1 Design

The evaluation dimension of style strength measures the extent that the output demonstrates a target style. The target document types of student interviews, interviews with news reporters, presidential debates, and public papers show different degree of the target style. I approximate target strength using a binary classifier that predicts texts between letters and one of those target types. The classifier is trained by fine-tuning a HuggingFace RoBERTa-base model,[9] using the training split of the balanced corpora corresponding to each document pair (see Table 6.9). Prior work define an *Accuracy* measure for style strength using such a classifier [76, 97, 142]. *Accuracy* is the percentage of texts classified as having the target type. In this work, the automatic evaluation measure is defined as the average probability assigned to the target type across all validation or test texts, referred to as "Average Probability" (Avg. Prob.). However, the predicted probability of the target type is not always a reliable indicator of model confidence, as the choice of training objective—such as the cross entropy, or focal loss—can influence whether the model outputs are smooth or sharply peaked. Therefore, I also report Accuracy in the same way as in prior work.

[9]RoBERTa-base was pre-trained on a large corpus of English data in a self-supervised fashion. It is commonly intended to be fine-tuned on a downstream task. The URL is `https://huggingface.co/FacebookAI/roberta-base`.

Table 6.12: Evaluation measures across two dimensions of style transfer quality.

Evaluation dimension	**Automatic measures**
Style Strength	Accuracy; Average Probability (Avg. Prob.)
Content Preservation	MNLI

During the early development, the initial style classifier exhibited a defect, making predictions biased to statements that contain the style cues, such as *"Well, you see"* and *"Well, you know."* This originates from the training data, particularly texts from interviews, using connective words such as *"and" "but", "now," "here," "Well, let me tell you, young fella,"* or simply *"well."* The connective words exist on sentence boundaries of the interview scripts, because the interviews are naturally occuring dialog from a public speaker who tends to use those style cues. None of the approaches provides explicit instructions regarding connective words. However, initial results showed that SUP-NATINST's outputs on the validation split often contain such style cues. When a model fails to generate those words at the beginning, the predicted probability of the target style is low regardless of other style changes in the text. I therefore further process the training data before training the style classifier such that three connective words/expressions that appear at the beginning of a text are removed, including *"And," "Well"* and *"I think."* It remains an open question for whether removing these connective words is appropriate, specifically, whether such words should be considered part of the writing style. Indeed, stopwords and punctuation are important stylometric features [89], and systematically removing them may alter the stylistic characteristics of the content. In my case, I remove these connective words because their consistent placement at the beginning of responses appears to be a direct result of the interview format, serving more as turn-taking or boundary markers than as stylistic features. As such, they are less representative of the stylistic patterns within the content that I aim to analyze.

The second dimension, content preservation, is based on the intuition that style transfer should not change the content of the text. Content preservation is not a new concept: prior work of text style transfer used different terms such as consistency, entailment, etc. [82], where BLEU is a common measure [80, 173]. However, a high BLEU score based on exact token matching remains a rough estimate about content preservation, with drawbacks

such as not modeling semantic similarity. MNLI is the overall best performing measure according to the text consistency evaluation benchmark QAFactEval [55]. MNLI uses a pre-trained RoBERTa model trained from the MNLI textual entailment dataset to predict whether a premise entails, contradicts, or is neutral to a hypothesis. For content preservation, I use MNLI to predict whether a source text (as the premise) entails, contradicts, or is neutral to the style-transferred text (as the hypothesis) for subsequent results in this section. A key limitation of MNLI is the mismatch between the definition of entailment and the notion of content preservation. Content preservation refers to maintaining the original semantics [82] and can vary in degree from high to low. By contrast, a clear entailment relationship does not necessarily imply high content preservation. For example, the sentence *"Pat is a fluffy cat that loves to play"* entails the sentence *"Pat is a cat."* but much of the original content is lost. In Chapter 5, I calculate MNLI in both directions and take the average, which may be an alternative measure.

A harmonic mean between Average Probability and MNLI is calculated to show a combined, balanced effect of both style strength and content preservation. The harmonic mean weighs Average Probability and MNLI as equally important.

There are four experimental conditions, differing by the paired corpora, and ordered with the primary corpora of Letters–Student Interviews presented first because student interviews fit the concept of the target style particularly well. Conducting experiments with a specific corpora means three things: (1) the number of texts in the validation and test split is balanced in a manner specific to this corpora, (2) comparable prompting examples are selected by searching using target texts as queries to find comparable source texts, and (3) a classifier that reports on style strength is trained from a balanced number of texts from both types in this corpora. In each table, the first two rows are reference baselines on the source and target texts, measured on the two dimensions without doing any style transfer. The Accuracy and Average Probability on source and target texts validates the accuracy classi-

fier: one should be low and the other should be high. The two reference rows show MNLI values of 1, which would indicate the perfect content preservation. In this way, the second row represents an ideal, high baseline, of strong target style strength and perfect content preservation. In each table, results for the system rows are averages over three using different one-shot examples in the prompt, repeatedly. In addition to automatic evaluation results, I perform qualitative analysis on model outputs, and summarize insights about the model performance.

6.5.2 Results on the Letters–Student Interviews Corpora

Letters–Student Interviews is my primary paired corpora, as student interviews model the target style most closely. Table 6.13 presents the system performance results on this corpora. The first two rows serve as reference baselines. The high accuracy on target texts (97.6%) indicates that the classifier is generally reliable. The source texts, which are expected to yield low accuracy, result in 21.7% of the texts being classified as target. This outcome suggests a limitation of using the classifier for evaluating system performance (***RQ10***): the classifier may have been trained on mixed data, and the source texts themselves may exhibit stylistic diversity, with some texts resembling the target style. I will further discuss this limitation in Section 6.5.7. In the context of this corpora, it implies that certain source texts (letter sentences) are already conversational or colloquial in tone, using language aimed at a public audience. This explanation is plausible, as the editorial notes in the book indicate that some Reagan letters include Reagan's responses to citizen correspondence.[10] For example, the classifier predicts a 0.997 probability of the target label for the following source text: *"Yes, this country is of the people, by the people, and for the*

[10]As noted in two chapter headers: *"Most of the letters that Reagan wrote did not involve responding to critics, but some did. Reagan's critics were a varied lot. Most were concerned citizens who wrote him about a particular policy."* and *"Reagan wrote many letters to children and young people. He replied both to individual letters and to collections of letters he received from school children."*

Table 6.13: Results on the Letters–Student Interviews (source–target) corpora, on the validation split. Accuracy shows the percentage of style transferred texts classified as the target.

System	Prompting Example	Accuracy	Avg. Prob.	MNLI	H-mean
All source (reference)	-	21.7	0.211	1	0.348
All target (reference)	-	97.6	0.974	1	0.987
NoEx	-	**98.8**	**0.984**	0.410	0.578
Sup-NatInst	Long	68.3	0.677	0.782	0.726
	Short	68.3	0.684	0.811	**0.742**
PreparedDescr	Long	26.5	0.276	0.714	0.399
	Short	16.1	0.170	0.771	0.279
StyleReasoning	Long	15.7	0.167	0.682	0.268
	Short	10.0	0.114	0.762	0.198
EndGoal	Long	24.9	0.244	0.760	0.369
	Short	14.1	0.143	**0.813**	0.243

people and what we need today, is more by the people." Similarly, *"Well, probably we are all responsible to the extent that we tend to be apathetic or to go our own way with our own affairs and not look too closely at government."* (target label probability: 0.990) illustrates how some source texts naturally exhibit features of the target style.

With the low and high baselines established upfront, the remaining rows report system performance addressing ***RQ8***. Figure 6.1 presents the 95% bootstrapped confidence intervals for two measures (Avg. Prob. and MNLI), following the same procedure described in Section 5.6. The Avg. Prob. plot reveals three distinct sets of systems, corresponding to high, medium, and low regions: (1) NoEx; (2) Sup-NatInst; and (3) PreparedDescr, StyleReasoning, and EndGoal. Comparisons between systems across these sets yield statistically significant differences. The long and short variants of the same system are consistently in the same set. Additional bootstrap hypothesis tests were conducted on system pairs where necessary.[11] No statistically significant differences were found within a set. NoEx achieves the highest average probability (and transfer accuracy), with relatively small variance. It is followed by Sup-NatInst (Long and Short), though both

[11] In Figure 6.1, for any two systems where one's point estimate overlaps with the other's confidence interval, bootstrap hypothesis tests were conducted following the procedure described in Appendix E. Whether the p-value falls below the 0.05 threshold determines statistical significance.

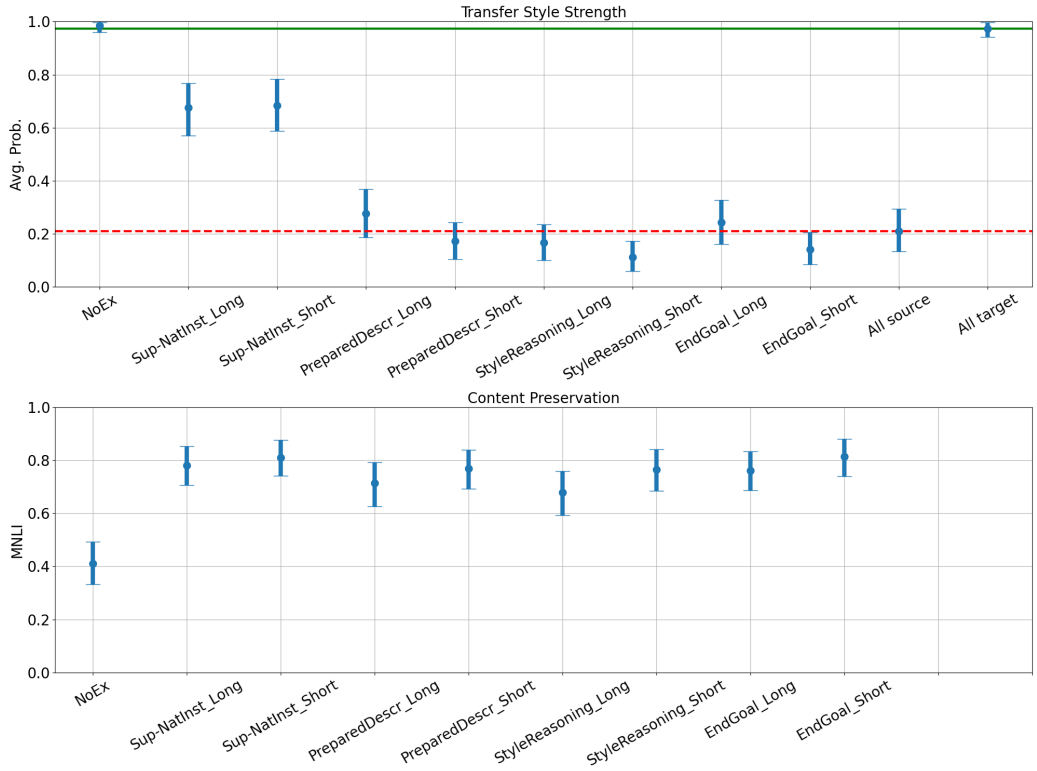


Figure 6.1: Point estimates of Average Probability (Avg. Prob.) and MNLI scores, with bootstrapped 95% confidence intervals on the primary corpora of Letters–Student Interviews. The green solid line indicates the target reference, i.e., Avg. Prob. of all target texts, while the red dashed line indicates the source reference, i.e., Avg. Prob. of all source texts. Systems are ordered as in Section 6.2, with instructions progressing from basic to complex.

exhibit substantially larger variance in average probability compared to NoEx. The remaining systems achieve no more than 0.28 in average probability, performing at or below the source reference. A strong and intuitive linear correlation is observed between transfer accuracy and average probability. Based on the 11 paired values in Table 6.13, the Pearson correlation coefficient, appropriate for assessing the strength of a linear relationship, is 0.999 (significant, p-value: 6.5e-17).

MNLI reports quite opposite results. NoEx has the lowest MNLI score. EndGoal (Short) has the highest MNLI score. The MNLI plot shows two sets of systems: (1) NoEx, and (2) all other systems. Systems from different sets are significantly different; no further statistically significant differences were found within a set.

Original Text	Style Transferred Text	System	Probability	Δ Prob.	Issue
Probability Quintile: Low (0–20%)					
Evidently we have to blame the postal service for this one.	It appears that the responsibility for this issue lies with the postal service.	PREPAREDDESCR Long	0.003 (4%)	-1e-04 (35%)	①
We did it with an unprecedented use of federal forces including the military in cooperation with local and state forces.	Through an unprecedented collaboration of federal, military, local, and state forces, we successfully achieved our goal.	STYLEREASONING Long	0.005 (19%)	-4e-03 (21%)	①
Probability Quintile: Medium (40-60%)					
In 1962 JFK's budget for military spending was 48 percent of the total budget.	In 1962, President Kennedy allocated 48 percent of the total budget to military spending, reflecting the importance of national defense in our strategic priorities.	PREPAREDDESCR Long	0.011 (42%)	0.004 (56%)	①
Inflation which was 12.4 percent is now running at 4.5 percent.	The rate of rising prices, which was at 12.4 percent, has now decreased to 4.5 percent.	ENDGOAL Short	0.300 (60%)	0.192 (70%)	①
Inflation which was 12.4 percent is now running at 4.5 percent.	The rate of rising prices, which was at 12.4 percent, has now decreased to 4.5 percent.	ENDGOAL Short	0.300 (60%)	0.192 (70%)	①
They are the result of almost 40 years of deficit spending and excessive taxation.	They are the outcome of nearly four decades of financial decisions that have led to budget deficits and high levels of taxation.	STYLEREASONING Long	0.011 (42%)	0.006 (59%)	①
We found many well-to-do people taking such low interest loans and reinvesting those loans in U.S. treasury bills for which the taxpayers were paying an interest rate double and even triple the interest on the college loan.	It has come to our attention that some individuals have been taking advantage of low interest loans and reinvesting them in U.S. treasury bills, resulting in taxpayers paying double or even triple the interest on college loans.	ENDGOAL Short	0.037 (53%)	-0.1986 (11%)	①
Probability Quintile: High (80-100%)					
Again, our credibility is compromised because they crossed the border and nothing was done.	Well, let me tell you, folks, our credibility takes a hit when those folks cross that border and we don't do a darn thing about it.	SUP-NATINST Long	0.997 (94%)	0.0286 (65%)	① ②
Should a leader utter such threats if he knows he cannot respond if his threat is ignored?	Well, let me tell you, a true leader doesn't go around making empty threats. If you can't back it up, you shouldn't be saying it in the first place.	NOEX Long	0.997 (91%)	0.989 (54%)	① ②
Inflation which was 12.4 percent is now running at 4.5 percent.	I am pleased to inform you that the inflation rate, which was previously at 12.4 percent, has now decreased to 4.5 percent.	ENDGOAL Long	0.996 (83%)	0.888 (78%)	①

Table 6.14: Original and their style-transferred texts are shown, with probability reported for the style-transferred texts and the corresponding Δ probability between the two. Percentile ranks are indicated in parentheses, and key observations are annotated with circled digits.

SUP-NATINST (Short), i.e., prompting with basic instruction and short examples achieves the highest harmonic mean (0.742, and 75% of the target reference), while more complex prompt designs tend to yield lower harmonic means compared to NOEX, which has no prompting examples at all.

Original Text	Style Transferred Text	System	MNLI	Issue
MNLI Quintile: Low (0–20%)				
In 1962 JFK's budget for military spending was 48 percent of the total budget.	In 1962, President Kennedy allocated 48 percent of the total budget to military spending, reflecting the importance of national defense in our strategic priorities.	PREPAREDDESCR Long	0.002 (5%)	③
Should a leader utter such threats if he knows he cannot respond if his threat is ignored?	Well, let me tell you, a true leader doesn't go around making empty threats. If you can't back it up, you shouldn't be saying it in the first place.	NOEX Long	0.111 (17%)	③
Inflation which was 12.4 percent is now running at 4.5 percent.	I am pleased to inform you that the inflation rate, which was previously at 12.4 percent, has now decreased to 4.5 percent.	ENDGOAL Long	0.108 (17%)	③
MNLI Quintile: Medium (40-60%)				
They are the result of almost 40 years of deficit spending and excessive taxation.	They are the outcome of nearly four decades of financial decisions that have led to budget deficits and high levels of taxation.	STYLEREASONING Long	0.973 (56%)	④
We did it with an unprecedented use of federal forces including the military in cooperation with local and state forces.	Through an unprecedented collaboration of federal, military, local, and state forces, we successfully achieved our goal.	STYLEREASONING Long	0.964 (52%)	④
MNLI Quintile: High (80-100%)				
Evidently we have to blame the postal service for this one.	It appears that the responsibility for this issue lies with the postal service.	PREPAREDDESCR Long	0.993 (86%)	④
Inflation which was 12.4 percent is now running at 4.5 percent.	The rate of rising prices, which was at 12.4 percent, has now decreased to 4.5 percent.	ENDGOAL Short	0.992 (79%)	④
But that was when we were buying Haile Selassie a million-dollar yacht, sawmills for a country that had no trees and paving a highway in a land where there were no automobiles.	And that was during a time when we were purchasing a million-dollar yacht for Haile Selassie, building sawmills for a country without trees, and constructing a highway in a land without automobiles.	SUP-NATINST Long	0.993 (82%)	④

Table 6.15: Original and their style-transferred texts, with MNLI scores reported between each pair. Percentile ranks and key observations are noted, as in Table 6.14.

6.5.3 Examples on the Letters–Student Interviews Corpora

To better understand an overall style transfer effects across different systems (***RQ8***) and the validity of the chosen evaluation measures (***RQ10***), I inspect texts sampled from three quintiles (0-20%, 40-60%, 80-100%) based on the sorted sequences of probability and MNLI scores, by all systems. The three ranges are referred to as "Low," "Medium," and

"High" in Tables 6.14 and 6.15, respectively. Texts are sampled judgmentally within each quintile for balanced coverage across systems. I also aim to have the same text examined for both probability and MNLI scores, to better illustrate their tradeoff on the same content. Additionally, I add a column of Δ probability to show the difference of probability scores of the original text and style transferred text. The quintile boundaries of the probability scores are 0.003, 0.005, 0.010, 0.272, 0.995, and 0.997, respectively. The quintile boundaries of the MNLI scores are 0.000, 0.241, 0.888, 0.979, 0.992, and 0.995, respectively. This inspection highlights four observations about the selected measures and the systems, as detailed below:

(1) **Style classifier for Letters–Student Interviews emphasizes discourse markers**: The three texts in the High Avg. Prob. range all contain some form of discourse marker (DM) [139], such as *"Well,"* or *"I am pleased to inform you that."* This observation also occurs in other style transferred texts not listed in the table. The DMs are sometimes brief while others are clausal length[12] *"Well"* is one example of a DM that serves particular conversational function about the utterance quality [139, 141]. *"Let me tell you"* which contains the pronouns of "you" and "me" is another DM that indicates listener and speaker participation [139]. When DMs are present, the style classifier (trained specifically on this corpora, Letters–Student Interviews) tends to assign a high probability score for the target style label. This is intuitive, as those markers appear frequently in the target texts (student interviews), on which the style classifier was trained. By contrast, the style transferred texts in the Low and Medium quintiles contain no such DMs.

(2) **SUP-NATINST and NOEX generate more DMs**: These two systems tend to begin with phrases like *"Well, let me tell you,"* more frequently than others. This suggests

[12] Although later critiques consider clausal-length discourse indicators as not DMs in the strict sense [57].

that they benefit from the basic instruction provided, which is the distinguishing part from other systems. The part of the instruction, *"speaks like something Ronald Reagan would say when he is responding to a typical museum visitor,"* may prompt the models to adopt DMs that support conversational flow. As a further plausible explanation, the LLMs may generalize the basic instruction by drawing on conversational patterns learned during pretraining on public-facing content, which may have included parts of the same Reagan collection from the web.

(3) **MNLI effectively penalizes hallucinations**: In at least two of the three (and probably all three) texts in the Low range of the MNLI table (6.15), there is additional but ungrounded content. The first text includes the phrase *"reflecting the importance of..."* which, in my view, does not obey the original text's intended emphasis on excessive spending by a prior administration. Similarly, in the second text, *"If you can't back it up..."* slightly diverges in meaning from the original message, which emphasizes an inability to respond if a threat is ignored. In the third text, the added phrase *"I am pleased to inform..."* is also questionable, as the original letter features Reagan explaining his economic and other policies to a concerned citizen who believed that the economy was bleak. In this context, a frank and serious tone would be more appropriate than the overly cheerful one used. Prior work described a phenomenon known as "hallucination" [77], in which LLMs generate content that is coherent, readable, and grammatically correct, yet not factually accurate. In these cases, the low MNLI scores appropriately penalize hallucinations.

(4) **MNLI effectively reflects content preservation**: In contrast to the previous point, this observation is supported by the five texts in the medium and high ranges, where style changes are evident while content remains largely preserved. Side-by-side inspection shows minimal content addition, such as replacing the more colloquial and

personal term *"spending"* with *"financial decision,"* and *"blame"* with the more formal phrase *"responsibility for ... lies with..."*

In summary, my inspection supports the use of MNLI scores for measuring content preservation, while the utility of the style classifier (trained on this corpora) for style strength (the first dimension) seems to be strongly influenced by discourse markers. With that one caveat, on the Letters–Student Interviews corpora, I find that prompting with basic instructions and short examples achieves the best balance between transfer style strength and content preservation.

6.5.4 Results on the Letters–Interviews (w/ Reporters) Corpora

I repeat this analysis for the other three paired corpora and report any new insights. The classifiers may behave differently, partly because they were trained on different corpora. Table 6.16 shows the next corpora of Letters–Interviews with Reporters. This is the most similar corpora to the previous one, but differs in the content of the target texts and the quantity of balanced texts, due to the inclusion of more interviews, and a different set of interviews, which Reagan conducted with news reporters (see Table 6.8). The source reference achieves an accuracy of 4.6%, which is lower than that observed with Letters–Student Interviews corpora, even though both source texts consist of letter sentences, albeit in different quantities. As expected, the target reference achieves an accuracy of 95.2%, consistent with the classification results from the primary corpora.

Comparing the two paired corpora, aside from the difference in quantity (462 vs. 1,768, for the training split), the gap in source reference accuracy (21% vs. 4.6%) suggests two possibilities: (1) a sampling difference, where the primary corpora may have a non-random sampling bias; and (2) fitness for the direction of style transfer, as texts from interviews with news reporters likely differ more stylistically from letters than do interviews with

Table 6.16: Results on the Letters–Interviews (w/ reporters) corpora, on the validation split.

System	Prompting Example	Accuracy	Avg. Prob.	MNLI	H-mean
All source (reference)	-	4.6	0.047	1	0.091
All target (reference)	-	95.2	0.951	1	0.975
NoEx	-	**96.5**	**0.962**	0.416	0.581
Sup-NatInst	Long	52.9	0.531	**0.808**	0.641
	Short	70.0	0.700	0.745	**0.722**
PreparedDescr	Long	17.5	0.183	0.707	0.290
	Short	21.2	0.213	0.655	0.321
StyleReasoning	Long	19.2	0.197	0.653	0.302
	Short	19.0	0.191	0.663	0.296
EndGoal	Long	13.9	0.145	0.711	0.241
	Short	15.6	0.159	0.741	0.262

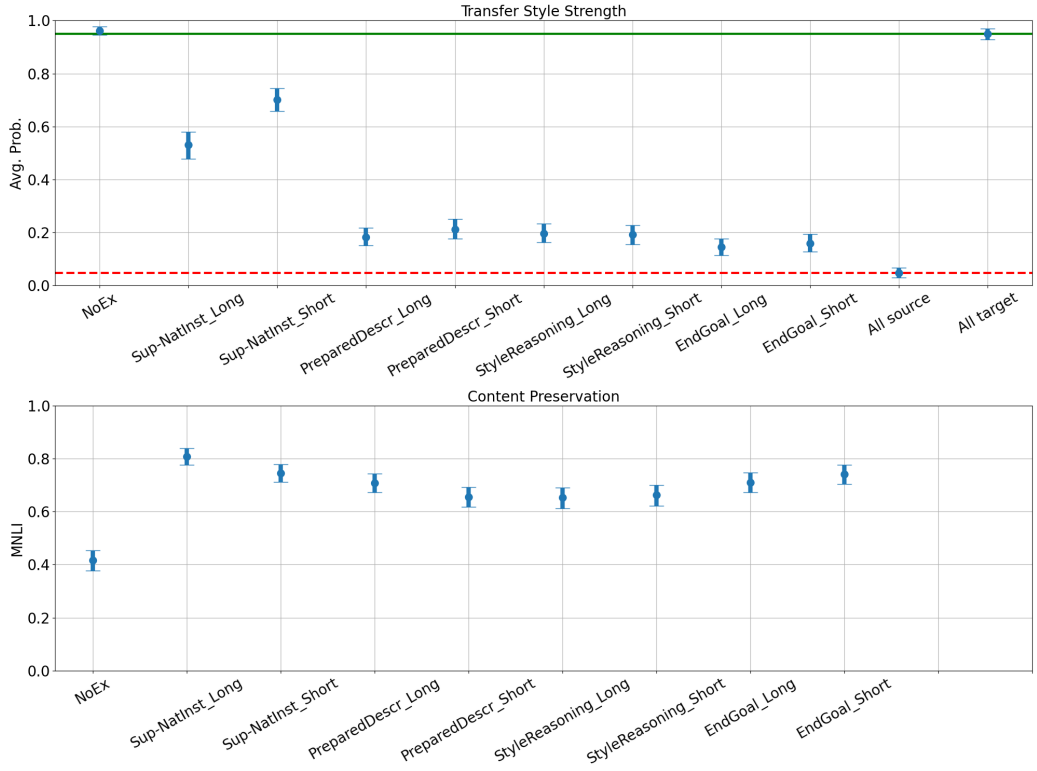


Figure 6.2: Point estimates of Average Probability (Avg. Prob.) and MNLI scores, with bootstrapped 95% confidence intervals on the Letters–Interviews w/ Reporters corpora.

high school students. Regarding (1), both sets of source texts (from this corpora and the previous corpora) were sequentially sampled from letter sentences in the order of their appearance in the book, where letters are organized into chapters based on topics. I notice that all original (source) texts in Table 6.14 pertain to domestic policy—a topic that, in my

view, aligns well with the target style despite being letter sentences. This topical bias may influence stylistic features and contribute to the higher accuracy. For (2), Reagan may use more complex vocabulary in interviews with news reporters, making these utterances more distinguishable from letter sentences.

The results for the systems in Table 6.16 are largely consistent with those in the previous table (6.13). For instance, SUP-NATINST remains the top performer by the harmonic mean, reinforcing the effectiveness of prompting with basic instructions and short examples. Figure 6.2 plots the point estimates and bootstrapped confidence intervals for the Letters-Interviews w/ Reporters corpora. There are a few new insights. SUP-NATINST (Long) has a significantly more competitive MNLI score than other systems. Meanwhile, it performs significantly worse than its short variant on Avg. Prob. While the differences in Avg. Prob. and MNLI between the two variants of SUP-NATINST are not significant in the Letters-Student Interviews corpora, they are significant for this corpora. Given this well-controlled comparison, the observed effect can be attributed to the prompting examples used. As described in the design of the proposed approaches (Section 6.2.1), comparable examples were selected by searching using target texts as queries to find source texts. Comparing the `pos_example_output` columns in Tables 6.4 and 6.5, the target texts from interviews with reporters are generally more complex, especially in the long examples. Having more complex expressions in the input may enhance content preservation, as there is more information to convey and thus perhaps leaves less room for hallucination.

6.5.5 Results on the Letters–Presidential Debates Corpora

Table 6.17 presents results for the Letters–Presidential Debates corpora. The source reference achieves an accuracy of 50.9%, and it has a high variance. This relatively high reference compared to the primary corpora helps explain the observed increase in accuracy

Table 6.17: Results on the Letters–Presidential Debates corpora, on the validation split.

System	Prompting Example	Accuracy	Avg. Prob.	MNLI	H-mean
All source (reference)	-	50.9	0.509	1	0.674
All target (reference)	-	96.4	0.947	1	0.973
NoEx	-	**97.6**	**0.954**	0.406	0.569
Sup-NatInst	Long	86.1	0.829	0.624	0.712
	Short	87.9	0.845	0.666	**0.745**
PreparedDescr	Long	70.3	0.685	0.543	0.606
	Short	70.9	0.682	0.522	0.591
StyleReasoning	Long	51.5	0.492	0.608	0.544
	Short	48.5	0.471	0.588	0.523
EndGoal	Long	49.7	0.484	**0.736**	0.584
	Short	49.7	0.463	0.694	0.555

across most systems compared to the previous corpora. For example, systems with complex designs, i.e., those other than NoEx and Sup-NatInst, previously achieved no more than 27% accuracy for the primary corpora and no more than 22% for the Letters–Interviews corpora, but now every system exceed 48% in accuracy. Manual inspection reveals that the classifier attends to stylistic features beyond just discourse markers. For example, style-transferred texts beginning with phrases like *"We have consolidated..." "One of the more challenging issues is..."* and *"This situation is of concern to us..."* all receive an absolute score above 0.9.

MNLI declines for all systems except NoEx, which continues to yield consistently low scores. Since MNLI is not customized for this specific corpora, the observed downward trend in MNLI scores is likely due to the use of different prompting examples rather than characteristics of the corpora itself.

6.5.6 Results on the Letters–Public Papers Corpora

Table 6.18 presents results for the Letters–Public Papers corpora. Public papers in the Reagan collection are transcripts of speeches delivered publicly by Reagan. The source and target references achieve expected accuracy, 13.4% and 90.4% respectively, potentially due

Table 6.18: Results on the Letters–Public papers corpora, on the validation split.

System	Prompting Example	Accuracy	Avg. Prob.	MNLI	H-mean
All source (reference)	-	13.4	0.148	1	0.258
All target (reference)	-	90.4	0.905	1	0.950
NOEX	-	**99.8**	**0.997**	0.412	0.583
SUP-NATINST	Long	90.5	0.900	0.538	0.674
	Short	96.8	0.965	0.519	**0.675**
PREPAREDDESCR	Long	86.1	0.855	0.360	0.507
	Short	88.9	0.887	0.414	0.565
STYLEREASONING	Long	78.1	0.783	0.495	0.606
	Short	76.4	0.772	0.509	0.613
ENDGOAL	Long	69.6	0.694	0.540	0.607
	Short	66.6	0.668	**0.615**	0.640

to the two types of texts having very distinct styles.

Among the systems, NOEX and SUP-NATINST (Short) achieve the best accuracy. Inspection of the style transferred texts reveals that these two systems contain Discource Markers (DMs) as stylistic enhancements. Content preservation is consistently lower than for other paired corpora (see Section 6.2.1). This may be because no retouch was done to the prompting examples for this corpora. For a balanced effect of style strength and content preservation, SUP-NATINST (Short) remains the top system.

6.5.7 Style Classifier Limitations

The above experiments use different classifiers to measure transfer style strength. The classifiers were trained with different paired corpora, and therefore whose prediction has different interpretations. For example, the classifier for Letters–Student Interviews seems to almost trivially predict based on the presence of discourse markers, while the classifier for Letters–Presidential Debates seems to predict based on stylistic cues that are more complex and diverse. These differences can be seen as different inductive biases arising from the limited stylistic coverage in each of the four paired corpora.

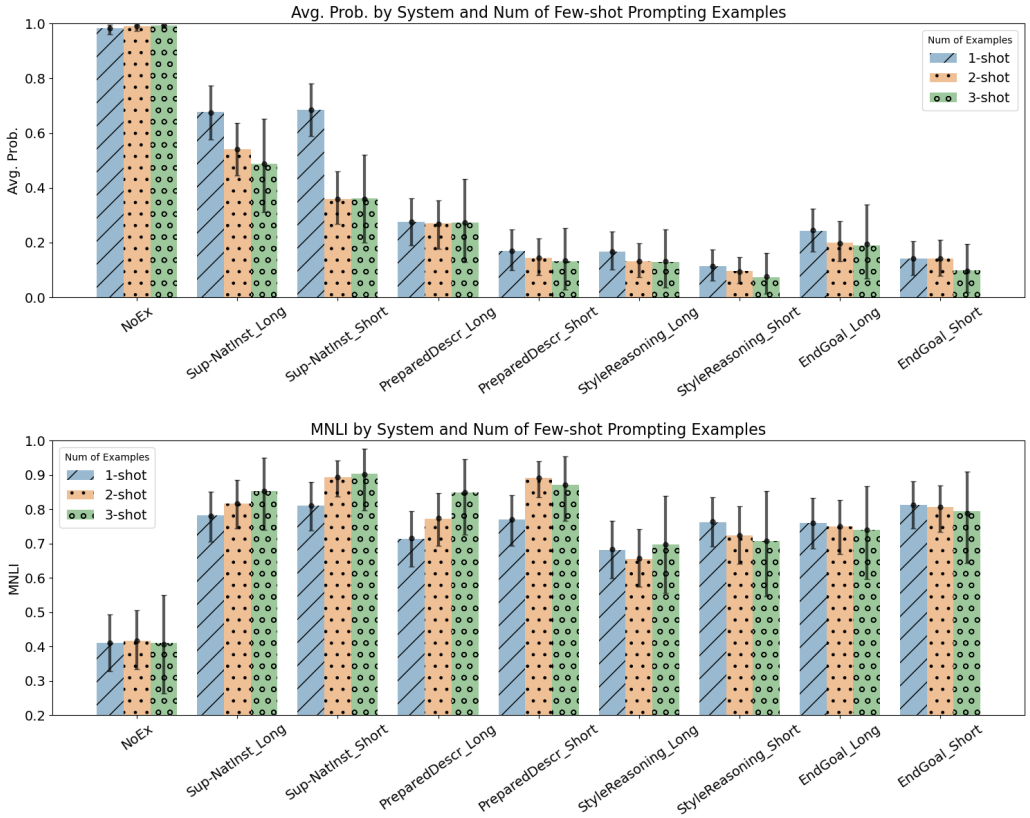


Figure 6.3: Average Probability and MNLI scores across systems with varying numbers of few-shot examples on the validation split, for the Letters–Student Interviews paired corpora, along with their bootstrapped 95% confidence intervals. For MNLI, as all values are above 0.2, the y-axis starts from 0.2 for clearer visualization.

6.5.8 Effect of the Number of Prompting Examples

To address ***RQ9***, I investigate the impact of varying the number of prompting examples for each system. Based on the the available set of three prompting examples per system, I experiment with selecting N = 1, 2, 3 examples for few-shot prompting, and compute the measure values on the validation split. For each setting, the process is repeated over all possible combinations of selecting N out of three examples. When computing point estimates and bootstrapped 95% confidence intervals, all measurement values from different repeats are pooled together. Figure 6.3 presents the results on the primary corpora. The plot for Accuracy is omitted as its overall pattern closely resembles that of Avg. Prob.

For most systems, no significant differences in both average probability and MNLI are observed across 1-shot, 2-shot, or 3-shot prompting. However, for Avg. Prob., SUP-NATINST (both long and short variants) achieve statistically significantly higher scores with 1-shot prompting. This observation is relevant to the discussion on classifier biases. It is possible that 1-shot prompting directs at a smaller set of features the classifier should attend to, whereas 2-shot prompting is less well-directed. For MNLI, both SUP-NATINST (Short) and PREPAREDDESCR (Short) show statistically significant improvement with 2-shot prompting compared to 1-shot. Additionally, PREPAREDDESCR (Long) further improves its MNLI score when increasing from 2-shot to 3-shot prompting. It appears that content preservation, as measured by MNLI, can sometimes benefit from a larger number of prompting examples. Meanwhile, target style strength, at least as measured by Accuracy or average probability, tends to score higher with fewer prompting examples.

6.5.9 Validating Automatic Measures with Human Evaluation

Section 6.4 presents human evaluation results on a subset of system outputs, while Section 6.5 reports automatic evaluation results on the full set of system outputs. This section repeats the validation study to understand whether automatic measures can reliably reflect human preferences.

Given the pointwise human ratings on style strength and content preservation, and two measures that approximate the same, it is possible to compute the correlation coefficient for the two sets of paired scores. The correlation coefficient is the Spearman's rank correlation [148]. Spearman's rank correlation is suitable for large datasets (N>20) and ordinal variables (as human ratings are), and often used to correlate between automatic measures and human evaluation in natural language generation (NLG) and machine translation tasks [61, 111, 132]. It assesses the presence or absence of a monotonic relationship: a

value closer to 1 or -1 indicates a stronger positive or negative monotonic relationship between the two variables. There are 120 human ratings for style strength—30 each from three systems' output and target texts, see Section 6.4)—and 90 human ratings for content preservation (30 each from three systems' output). Corresponding classifier probabilities (used to compute Avg. Prob.), and MNLI scores are also available in matching quantities to those human ratings. Spearman's rank correlation between style strength and classifier probability is 0.209 (significant, p-value: 0.022, N=120). This is a weak correlation, which may be explained by the earlier discussion about classifier limitations. Spearman's rank correlation between human content preservation ratings and MNLI is 0.494 (significant, p-value: 7.5e-07, N=90). This is a moderate correlation, which is consistent with my earlier inspection (Section 6.5.3).

6.6 Summary

In this chapter, I investigated style transfer from historical letters to a more conversational and public-facing style. I characterized the target style using four types of documents. I proposed approaches that prompt LLMs with style transfer instructions expressed through general prompting schema, with and without chain-of-thought reasoning. I further proposed to create comparable examples from the paired corpora as few-shot examples within the prompt. Experiment results show that systems that prompt LLMs with basic instructions (SUP-NATINST) and short examples are most balanced between target style and content preservation, achieving about 75% of an ideal high baseline. For the best-performing system, the number of prompting examples impacts system performance: 2-shot prompting yields optimal results for content preservation, as reflected by MNLI scores. By contrast, 1-shot prompting appears most effective for style strength, based on Accuracy and average probability. Given the limitations of these style strength measures

and the moderate alignment of MNLI with human preferences, I recommend developing style transfer systems using SUP-NATINST with two short examples for few-shot prompting.

Chapter 7: Conclusions

This dissertation envisions automated systems that engage in conversational interaction as emulated historical figures, supported by historical document collections. This vision is traditionally known as *digital immortality*. One natural setting for such a system is historical museums. This dissertation aims to inform the museum studies community about the potential to realize this vision using emerging technologies, such as retrieval technologies for conversations, and large language models capable of overcoming low-resource language challenges. This prospective application also invites reflection on new challenges it may introduce. This is an ongoing, iterative inquiry in which technical development and conceptual understanding continually inform one another.

My inquiry begins with foundational conversational search techniques that enable access to any document collection, including the *historical document collection* defined in this dissertation. I develop a multi-stage pipeline—comprising query rewriting, passage retrieval, and re-ranking—evaluated using the TREC Conversational Assistance Track 2021 test collection (Chapter 3). Automatic query rewriting addresses the common challenge of interpreting user utterances in context for conversational search [41]. To evaluate its impact, I compare retrieval effectiveness with and without rewriting using nDCG@3, which emphasizes top-ranked results suitable for voice assistant delivery. I similarly evaluate three additional features of the retrieval pipeline. Results show statistically significant improvements, primarily at lower-ranked positions, though this may be influenced by pooling bias in the TREC CAsT test collection. This chapter offers practical insights into design-

ing effective multi-stage retrieval systems, detailing the configurations that shape retrieval performance and, ultimately, the quality of conversational information access.

I then shift from the general TREC CAsT test collection to a custom historical document collection based on the archival resources of Ronald Reagan. I build a prototype system supporting single-turn, retrieval-based interaction, and conduct a need-finding study with four experts in libraries, archives, and museums (LAM). After a think-aloud session issuing queries and reviewing retrieved results, participants affirmed the prototype's value in returning useful content. In the semi-structured contextual interviews, they offered insights into the envisioned system's potential in historical museums. Iterative coding of interview transcripts surfaced three design implications: (1) the need to create immersive experiences, such as being multi-sensory elements, or temporal-spatial continuity to contextualize history, (2) raising awareness of archival limitations, including strategic challenges related to digitization resources, and engaging curious, one-time museum visitors in long-term, in-depth inquiry, highlighting the opportunity for curators and system designers to foster synergy between archival content and public audiences.

The insights expose a gap between the current retrieval-based prototype and future immersive systems, such as those enabling "in-the-moment" experience with historical figures. A major step towards bridging this gap is rewriting retrieved historical content. I propose two sequential rewriting tasks: (1) contextualization, which augments unfamiliar or unclear mentions in the retrieved content (referred to as *uncontextualized mentions*) with elaboration texts, (2) style transfer, which adapts the retrieved content to reflect the speech style of a historical figure conversing with general museum visitors.

For contextualization, I collect human annotations on Reagan's letter sentences, chosen for their brevity and insider tone. The data helps to examine inter-annotator agreement. Results show that while human-identified mentions often overlap, differences exist, requiring fuzzy matching to make the most of the annotations. Human-generated elaborations also

vary in length and interpretive depth. I then develop six LLM-based approaches that perform one to several passes of few-shot prompting, with each pass dedicated to either identifying mentions, or generating elaboration texts. A comprehensive evaluation methodology assesses *how well systems can perform contextualization?* It involves human evaluation directly to system outputs, and automatic evaluation comparing against human annotations, accompanied with qualitative manual inspection on those outputs. Results show that most proposed approaches, except one baseline, successfully identify helpful mentions and generate high-quality elaborations.

For style transfer, I adopt a data-driven definition of source and target styles [82] using four non-parallel corpora from the Reagan collection. Letter sentences serve as the source style, and public verbal materials (e.g., interviews, presidential debates, speeches) represent the target style. I propose LLM-based prompting systems built on a general schema called SUP-NATINST, and using few-shot, *comparable examples*—sentence pairs on similar or the same topics across source and target texts. To evaluate *how well can systems perform style transfer?*, there are two standard dimensions: target style strength (measured via a trained classifier as Avg. Prob.), and content preservation (using the MNLI entailment score). Across all corpora, prompting without chain-of-thought reasoning and with short comparable examples yields the highest harmonic mean between the two measures. Meanwhile, to interpret this result as system performance, I conduct correlation analysis and manual inspection to understand *how they approximate real-world user preferences.* MNLI aligns moderately well with human judgments of content preservation. However, style strength measures (e.g., Avg. Prob. and Accuracy) correlate weakly with human ratings, likely due to overreliance on discourse markers rather than broader stylistic features. This limitation stems in part from adopting a data-driven definition of style, where stylistic diversity may be underrepresented or unevenly modeled.

While this dissertation makes promising contributions (see Section 7.1), it also faces

certain limitations (Section 7.2) and highlights directions for future research (Section 7.3).

7.1 Contributions

The contributions of this dissertation fall into three categories: system (S), methodology (P), and research (R). The proposed envisioned system was implemented in modular components, including a general-purpose conversational search pipeline, a retrieval-based prototype over the domain-specific Reagan collection, and two LLM-based systems that use few-shot prompting to perform contextualization and style transfer, respectively. I addressed ten research questions through the design, implementation, and analysis of the envisioned system, each of which constitutes a research contribution. Technical development and conceptual understanding are bridged through a diverse set of research methodologies: the procedure for creating a historical document collection enables the transition from general to domain-specific conversational search; the human annotation protocol allows human annotators to effectively demonstrate human performance to the contextualization task; and a multi-faceted evaluation framework supports analysis of whether each system component fulfills its intended function.

7.1.1 System Contributions

- ***S1:*** Development of a multi-stage, general-purpose conversational search pipeline incorporating query rewriting, passage retrieval, and re-ranking, along with empirical evaluation of its specifications and their impact on retrieval effectiveness. (Chapter 3, Section 3.2.1)

- ***S2:*** Development of a single-turn, retrieval-based prototype of the envisioned system, integrating passage retrieval, re-ranking, and sentence extraction over the Reagan historical document collection. This prototype served as a boundary object for a

need-finding study with museum experts. (Chapter 4, Section 4.2)

- ***S3:*** Development of LLM-based approaches for contextualization using few-shot prompting, with separate passes for identifying mentions and generating elaboration texts, to perform contextualization. (Chapter 5, Section 5.2)

- ***S4:*** Development of LLM-based prompting approaches for style transfer using *comparable examples* with topical overlap for few-shot prompting. (Chapter 6, Section 6.2)

7.1.2 Methodology Contributions

- ***M1:*** A procedure to scrape, segment, and index a historical document collection focused on Ronald Reagan, consisting of diverse document types including interview transcripts, letters, diaries, public papers, and presidential debate transcripts. (Chapter 3, Section 4.1, and Chapter 6, Section 6.3)

- ***M2:*** Development of a human annotation protocol for contextualization, including an AMT-based tool and a fuzzy matching strategy to address inter-annotator variability. (Chapter 5, Section 5.5)

- ***M3:*** A multi-faceted evaluation framework for contextualization, combining human evaluation, automatic evaluation against human annotations, and manual inspection. (Chapter 5.6)

7.1.3 Research Contributions

- ***R1:*** Validating the effectiveness of query rewriting at improving retrieval. (Chapter 3, Section 3.3.2)

- ***R2:*** Validating the effectiveness of retrieval techniques, including augmenting with document-level evidence and combining dense and sparse retrieval through sharding at improving retrieval. (Chapter 3, Section 3.3.2)

- ***R3:*** Validating the utility of the envisioned system, at a retrieval-based prototype form, including presentation of authentic, primary source information. (Chapter 4, Section 4.3.5)

- ***R4:*** Identifying three key design implications for deploying the envisioned system in historical museum settings, based on expert interviews: designing immersive experiences, considering archival assurance, and supporting visitor engagement for long-term scholarly inquiry. (Chapter 4, Section 4.3.6)

- ***R5:*** Analysis of human performance on contextualization. Found that human-identified mentions can overlap but vary slightly, and their elaboration texts vary in briefness and interpretiveness. (Chapter 5, Section 5.5)

- ***R6:*** Analysis of system performance on contextualization. Found that most system-identified mentions add value to the original content, and most system-written elaboration texts (except that by one baseline) are correct and high quality. (Chapter 5, Section 5.4, and 5.6)

- ***R7:*** Meta-valuation of LLM-generated annotations as substitute for human annotations in evaluation. Found that they may be potentially useful as a basis for recognizing promising systems. (Chapter 5, Section 5.7)

- ***R8:*** Analysis of system performance on style transfer. Found that prompts excluding chain-of-thought reasoning and including short, comparable examples produced the strongest overall performance. (Chapter 6, Section 6.5)

- ***R9:*** Analysis of the effect of the number of prompting examples on style transfer effectiveness. Results show that 2-shot prompting achieves the best content preservation, while 1-shot prompting yields the best target style strength. (Chapter 6, Section 6.5.8)

- ***R10:*** Discussion of the strengths and limitations of automatic measures by comparing them with human judgments and conducting manual inspection. While the content preservation metric correlates reasonably well with human preferences, style strength measures exhibit weaker alignment. (Chapter 6, Section 6.5)

7.2 Limitations

I would like to highlight the following limitations on language resources, to readers interpreting the results reported in this dissertation.

7.2.1 Other Historical Document Collections

While the Reagan collection (Section 4.1) provides a readily available language resource to populate the prototype and enables experiments on contextualization and style transfer, the envisioned system is not intended to be limited to emulating Ronald Reagan or any specific historical figure. Extending this dissertation to other historical collections would be valuable for evaluating its generalizability. Appendix C enumerates a list of alternative historical document collections of historical figures, including scientists Albert Einstein, Richard Feynman, Ada Lovelace, Grace Hopper, and other professions. The set of available collections will continue to grow. In general, the choice will depend on a digitized collection, ideally in English, and with copyright permission.

7.2.2 Systematic User Study

At the current stage of system development, no real user interaction has been incorporated due to the lack of a fully integrated end-to-end system. The completed need-finding study involved participants prompting the prototype with spontaneous single-turn utterances, which were typically simple and general questions. For example, *"What was the inaugural like?"* It may be beneficial to conduct a larger-scale user study involving real users interacting with certain components or the full end-to-end system.

7.2.3 Text Rewriting for Longer Texts

For the contextualization and style transfer tasks, I process the Reagan collection to perform rewriting on sentences. A sentence-length unit can be intuitive: (1) users of conversational interaction may have limited attention to listen to texts beyond certain length; (2) Sentences come with its original context, such as preceding text in the paragraph for contextualization. However, current findings on how effectively systems can contextualize or perform style transfer is limited to sentences. If inputs are longer texts, the proposed approaches can be applied, although their performance may be different.

7.3 Future Work

Future research could advance this dissertation by addressing systematic ethical issues, constructing a retrieval test collection, and building an end-to-end system with dialogue management capabilities.

Figure 7.1: One example interview between Reagan and reporters. The question by the reporter is complex and can be simplified as a user utterance in the retrieval test collection.

7.3.1 Systematic Ethical Investigations

The complex issues of authenticity, consent, legislation, and the emotional impact of AI-powered virtual immortality have long been recognized by scholars [136] and are increasingly acknowledged by users as well [60]. In Chapter 4, the need-finding study highlights potential risks associated with representing the deceased to museum visitors, such as the uncanny valley effect, which may cause user discomfort. Savin-Baden et al. identified various concerns, including ethical practices related to the hosting environment (often a third-party service provider), preservation of personal repositories (referred to here as personal historical collections), legislative ambiguity regarding posthumous digital presence, and distinctions between casual, formal, or entrepreneurial forms of digital immortality [138]. It is therefore important to systematically investigate ethical concerns arising from systems that posthumously represent historical figures. A valuable direction for future work would be to conduct a structured interview study with ethicists, examining the system both at the component level and as an integrated end-to-end system.

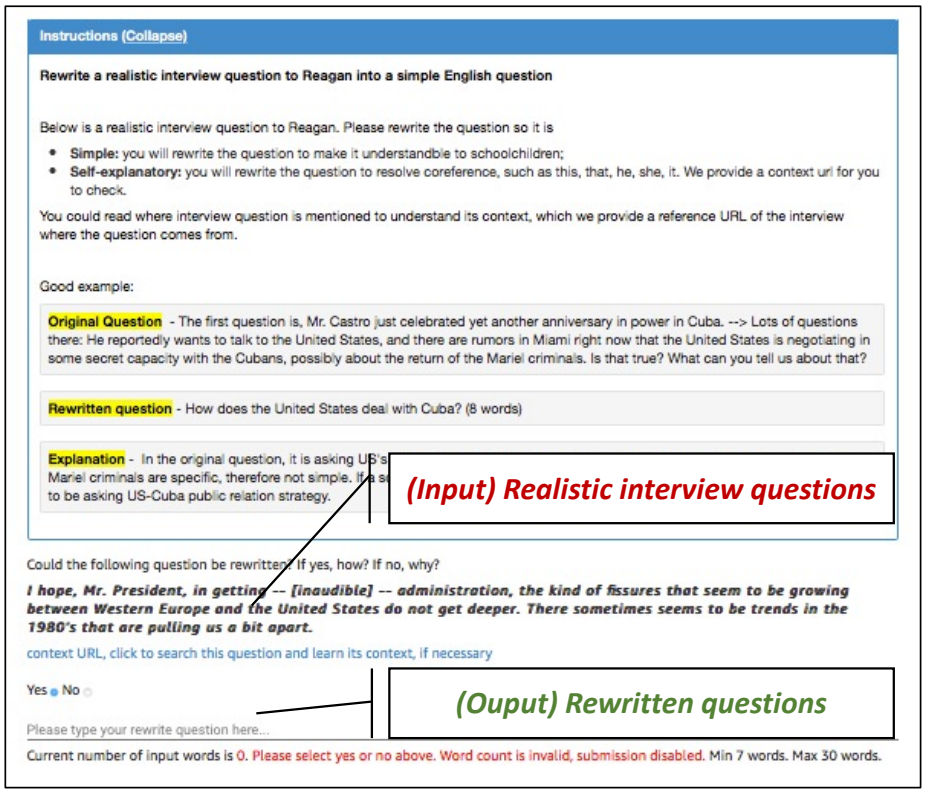


Figure 7.2: Annotation interface for collecting user utterances.

7.3.2 Constructing a Test Collection for Retrieval

User utterances and relevance judgments centered on the the current Reagan collection or alternative historical document collections are necessary to evaluate the performance of a retrieval component in a conversational search setting. Moreover, a sufficient quantity of such data is valuable to develop a retrieval component within the framework. There are two issues to consider when constructing such a collection. One is to construct single-turn user utterances that users might actually use with a historical figure that express some information need. The second is to collect relevance judgments to those utterances on the collection.

For the first issue, some utterances can be adapted from interview questions from the transcripts, primarily those raised by students, but possibly those raised by reporters. The

original interview questions are contextual and specific. To be representative of questions a museum visitor might ask, they would need to be made more general. Taking the interview example in Figure 7.1, the original question in the blue box is a descriptive one on South African Apartheid Policy, asked by a radio journalist. This question could be made more general as *"How do you react to actions taken in South Africa toward freedom for the black majority?"* Questions that were asked regarding other presidential administrations may be suitable, such as The Nixon Interviews between President Nixon and journalist David Frost [113], and the Exit Interview Project from the Jimmy Carter Presidential Library [81]. Questions asked to other historical figures that are general may be suitable too. For example, prior work constructed a list of interview questions to Astronaut Alfred Worden [170] or Holocaust survivors [159]. Past queries from web search logs about a historical figure may reflect underlying information needs that search engine users have and can potentially be rewritten as user utterances, not necessarily in a question format. Local experts in history could be recruited to write or validate the user utterances. Figure 7.2 illustrates a potential interface to obtain user utterances from above language resources.

Relevance judgments are binary or graded labels that indicate whether a document (or passage) being retrieved is relevant to a query (in my context, a user utterance). These labels are used in retrieval evaluation to examine the effectiveness of a retrieval system. One common method for collecting them is through pooling techniques, as described in the TREC CAsT 2020 overview [42].

7.3.3 End-to-end System with Dialog Management

With components for retrieval, contextualization, and style transfer developed to interact with user utterances in a pipelined design, the next step would be to integrate these components into an end-to-end system. For example, rather than applying the style transfer

component directly to original texts, it would operate on texts that have already been retrieved and contextualized, producing output that reflects the cascading effect of earlier two components. The cascading effect brings two questions. First, will contextualization output negatively affect the current style transfer process? For the first question, future work can conduct style transfer evaluation against non-contextualized input and contextualized input. Second, could style transfer negate the effect of contextualization—for example, by removing references that are precisely the ones being contextualized? Most text style transfer (TST) work assumes that style and content in a text are separable [82]. Nevertheless, future work could evaluate the impact of style transfer on contextualization by comparing outputs with and without post-hoc style transfer.

Additionally, dialogue management can be incorporated to generate well-situated responses based on the outputs of retrieval and rewriting. At this stage, multiple valid responses may be available from earlier components. A straightforward strategy is to select the highest-ranked response based on retrieval scores (the "pick-one" strategy). Alternatively, the system could summarize multiple responses (the "summarize" strategy), allowing users to ask clarifying questions or explore specific aspects through further interaction—ultimately guiding them toward a more deeply engaged conversation. Another possibility for dialog management is to incorporate mediation strategies to improve user experience—for example, experimenting with an antagonistic conversational agent emulating a museum visitor. Preliminary work on the benefits of introducing antagonism in AI suggests a promising direction, in which the envisioned system might shift from being merely sycophantic to to adopting a more corrective stance [30].

7.3.4 Utilizing Multimodal Content

The envisioned system focuses on textual material as input, and generates textual outputs that can be spoken by a virtual representation of a historical figure. However, historical document collections may be enriched with multimodal content such as images, audio, and video. At the basic level, close-up footage of Ronald Reagan speaking—though constituting only a small portion of the overall Reagan collection—can support video synthesis (as discussed in Chapter 1) by enabling the construction of lifelike virtual agents capable of realistic gestures and facial expressions. Additional multimodal content, including news photographs from the same time period and audiovisual recordings of relevant events, can serve as grounding context for interactive conversations. While the proposed system focuses on unimodal text-based retrieval and rewriting, recent work in multimodal retrieval-augmented generation (RAG) explores how to integrate diverse modalities, such as texts, images, audios, and videos, to enhance the quality and contextual richness of generated outputs [2]. From a learning perspective, both textual and non-textual modalities contribute to meaning-making processes [116]. An exciting opportunity lies in developing multimodal conversational interfaces in which a virtual historical figure dynamically presents visual or auditory content, such as showing a photo or playing a clip, to deepen engagement, much like a human presenter. Furthermore, multimodal augmented reality (AR) applications may construct immersive environments, such as virtual dioramas of a newsroom or a conference, enabling visitors to feel embedded within historical scenes and potentially fostering deeper engagement with cultural heritage content [70].

7.4 Final Thoughts

For conversational interaction with emulated figures supported by historical document collection, a system that retrieves, contextualizes, and does style transfer can be valuable from a museum studies perspectives, and the realization of this idea with prevailing retrieval techniques and revolutionary Large Language Models (LLMs) is promising as verified from independent investigation to each. This dissertation is a work with interdisciplinary significance. Situated at the intersection of cultural heritage data and the computational modeling of an emulated conversational agent, it may be seen as a two-way street—the former motivates the development of techniques that are effective for novel tasks, and the latter highlights the potential of cross-disciplinary approaches to foster meaningful engagement and innovative uses of cultural heritage data.

Appendix A: IRB Approval Letter for Chapter 4

1204 Marie Mount Hall
College Park, MD 20742-5125
TEL 301.405.4212
FAX 301.314.1475
irb@umd.edu
www.umresearch.umd.edu/IRB

DATE:	March 25, 2021
TO:	Xin Qian
FROM:	University of Maryland College Park (UMCP) IRB
PROJECT TITLE:	[1727782-1] Ask Me Anything: Bringing information retrieval technologies into museums of historical figures
REFERENCE #:	
SUBMISSION TYPE:	New Project
ACTION:	APPROVED
APPROVAL DATE:	March 25, 2021
EXPIRATION DATE:	March 24, 2022
REVIEW TYPE:	Expedited Review
REVIEW CATEGORY:	Expedited review category # 7

Thank you for your submission of New Project materials for this project. The University of Maryland College Park (UMCP) IRB has APPROVED your submission. This approval is based on an appropriate risk/benefit ratio and a project design wherein the risks have been minimized. All research must be conducted in accordance with this approved submission.

Prior to submission to the IRB Office, this project received scientific review from the departmental IRB Liaison.

This submission has received Expedited Review based on the applicable federal regulations.

This project has been determined to be a MINIMAL RISK project. Based on the risks, this project requires continuing review by this committee on an annual basis. Please use the appropriate forms for this procedure. Your documentation for continuing review must be received with sufficient time for review and continued approval before the expiration date of March 24, 2022.

Please remember that informed consent is a process beginning with a description of the project and insurance of participant understanding followed by a signed consent form. Informed consent must continue throughout the project via a dialogue between the researcher and research participant. Unless a consent waiver or alteration has been approved, Federal regulations require that each participant receives a copy of the consent document.

Please note that any revision to previously approved materials must be approved by this committee prior to initiation. Please use the appropriate revision forms for this procedure.

All UNANTICIPATED PROBLEMS involving risks to subjects or others (UPIRSOs) and SERIOUS and UNEXPECTED adverse events must be reported promptly to this office. Please use the appropriate reporting forms for this procedure. All FDA and sponsor reporting requirements should also be followed.

All NON-COMPLIANCE issues or COMPLAINTS regarding this project must be reported promptly to this office.

Please note that all research records must be retained for a minimum of seven years after the completion of the project.

If you have any questions, please contact the IRB Office at 301-405-4212 or irb@umd.edu. Please include your project title and reference number in all correspondence with this committee.

This letter has been electronically signed in accordance with all applicable regulations, and a copy is retained within University of Maryland College Park (UMCP) IRB's records.

Appendix B: IRB Determination (Not HSR) for Chapter 5 and 6

1204 Marie Mount Hall
College Park, MD 20742-5125
TEL 301.405.4212
FAX 301.314.1475
irb@umd.edu
www.umresearch.umd.edu/IRB

DATE:	September 26, 2024
TO:	Xin Qian
FROM:	University of Maryland College Park (UMCP) IRB
PROJECT TITLE:	[2240678-1] Developing computational methods to rewrite text from historical documents
SUBMISSION TYPE:	New Project
ACTION:	DETERMINATION OF NOT HUMAN SUBJECT RESEARCH
DECISION DATE:	September 26, 2024

Thank you for your submission of New Project materials for this project. The University of Maryland College Park (UMCP) IRB has determined this project does not meet the definition of human subject research under the purview of the IRB according to federal regulations.

We will retain a copy of this correspondence within our records.

If you have any questions, please contact the IRB Office at 301-405-4212 or irb@umd.edu. Please include your project title and reference number in all correspondence with this committee.

This letter has been electronically signed in accordance with all applicable regulations, and a copy is retained within University of Maryland College Park (UMCP) IRB's records.

Appendix C: Alternative Historical Document Collections

Figure	Language	Area	Interview	Diary (life or discipline)	E-book (biography)	Speech / papers	Letters
Einstein	Bilingual	Physics	Not quite available, has single QA pairs of quotes	The Travel Diaries of Albert Einstein: The Far East, Palestine, and Spain, 1922 - 1923	Einstein on Einstein: Autobiographical and Scientific Reflections	484 writings by Einstein and 3,450 letters written by and to him. An additional 3,441 documents appear in abstract.	
Feynman	English	Physics	Totaling 5 long interviews (all by Charles Weiner in several gap years)	Notebook at at the Niels Bohr Library & Archives in College Park	Surely You're Joking, Mr. Feynman! (Adventures of a Curious Character)	Nobel lecture notes and Cargo Cult Science notes	Perfectly Reasonable Deviations from the Beaten Track: The Letters of Richard P. Feyman
Edison		Invention		Two weeks of diary transcribed			
Admiral Grace Hopper		Computer science			A few biography books by others, not including two children's book		
Neil Armstrong		Space science					
Nikola Tesla		Invention			My Inventions: The Autobiography of Nikola Tesla		
Darwin	Bilingual	Biology		Beagle Diary (travel log)			English/Serbo-Croatian diary comparisons, Serbo-Croatian diary commentary, and Tesla/Scherff Colorado Springs correspondence
Sally Ride	English	Space science		Mostly work log e.g., shuttle training notes or Kidsat team		A sample of transcribed papers	

Figure C.1: Analysis of alternative historical document collections.

Appendix D: Instructions to Annotators on Contextualization

Figure D.1 shows a screenshot of the instructions provided to annotators for the contextualization task.

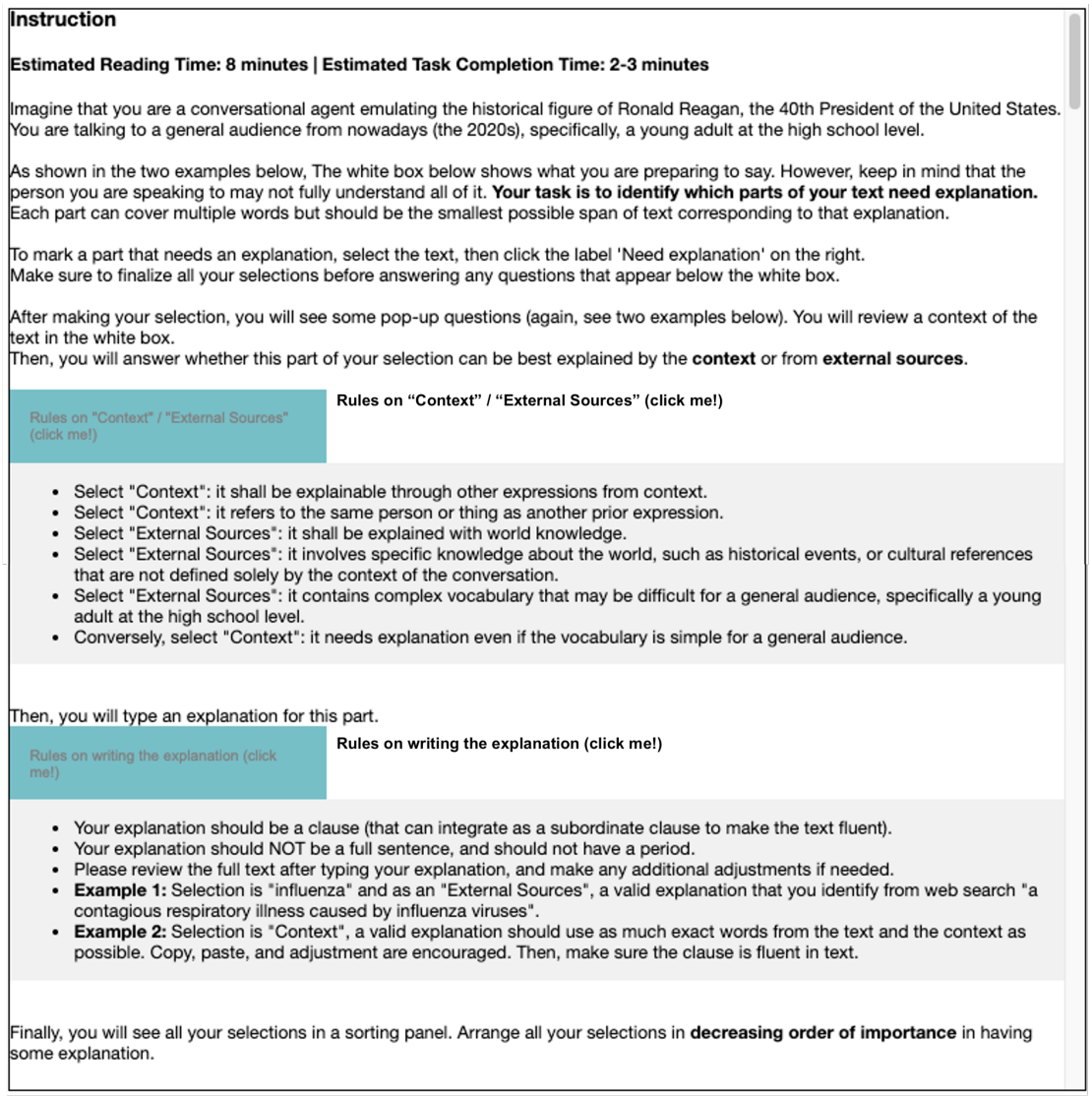


Figure D.1: Annotator instructions for contextualization on Amazon Mechanical Turk.

Appendix E: Procedure of a Bootstrap Hypothesis Test

A bootstrap hypothesis test determines statistical significance on whether two statistical measure values, such as the F_1 scores or the BLEU-4 scores from two different systems (as seen in Chapter 5.6), differ from each other. It does so by resampling the observed data with replacement to generate an empirical distribution of their difference and calculating the two-sided p-value under the null hypothesis. The null hypothesis states that the measure value of interest, calculated from two systems, come from the same distribution. [1] The alternative hypothesis is two-sided, stating that the F-score calculated from one system differs from that of the other system. Note that if two systems yield the same measure value in the original sample (i.e., no observed difference), the null hypothesis cannot be rejected under any circumstances. In such cases, a bootstrap hypothesis test will always produce a p-value of 1.

Take the measure value as F-score as an example, the procedure is as follows:

1. Combine the original two samples to create universal sample (under the null hypothesis).

2. Repeat N(=50,000) times

   (a) Draw a sample of the same amount, with replacement, to create Sample 1 and compute its F-score, denoted as F_1.

[1] Procedures based on lecture slides from the University of Washington, Summer 2020, CSE 312: Foundations of Computing II for one-sided alternative hypothesis, and Stanford University Winter 2021, CS109: Probability for Computer Scientists for two-sided alternative hypothesis.

(b) Draw another sample of the same amount, with replacement, to create Sample 2 and compute its F-score, denoted as F_2.

(c) Calculate the difference between two F-scores: $F_1 - F_2$.

3. Calculate the p-value as the proportion of absolute differences in $F_1 - F_2$ that are at least as extreme as the observed absolute difference in F-scores from the original two samples.

Appendix F: Supplementary Experiments on Contextualization for Mention Identification

This Appendix provides supplementary experiment results on identifying mentions from Chapter 5. To expand on the identifying mentions, Table F.1 provides a separate view within the overall mention identification results, on the validation split. Based on the earlier design (Section 5.6), the table divides into two groups: External and Context, based solely on the human decisions on whether a mention is best explained from external resources or the context ("relaxed count"). A mention considered by human as External is a true positive (TP) as long as a system extracts it; otherwise it is a false negative (FN) if a system does not extract it. There is no FP per category reported in this table, because there is no human decisions on a mention that human does not extract; hence no Precision reported. A1I annotations consider that the Joint (GPT) system has the top rank in Recall for mentions that human think best explained by external sources, while the Joint (Sonnet) system has the top rank in Recall for mentions best explained by context. A1J annotations place both Joint (GPT) and Joint (Haiku) as the top rank in Recall for mentions by external sources, and place the Joint (GPT) as the top rank in Recall for mentions by context. Focusing on the Recall values highlights the task's difficulty for LLMs: in this task, when a human annotation extracts a mention, the best-performing LLM has a 77.5-91.0% chance of extracting it. If a Joint system is used to mimic human annotation, they can extract both External and Context types as human do, with a recall of 50-81%. The

Table F.1: Results on mention identification with human annotations (relaxed count) on the validation split, through a separate view on the component values (i.e., sub-measures) of the overall F_1. There are two groups, External and Context, based on human annotations. The counts are relaxed counts in that they do not divide by the LLM's' attribution of mention types.

Model Family	Approach	External			Context		
		TP	FN	Recall	TP	FN	Recall
Human (reference)		497/459	110/260	0.819/0.638	145/183	49/140	0.747/0.567
Claude 3	Joint (Sonnet)	307/322	300/397	0.506/0.448	132/157	62/166	**0.680**/0.486
	Joint (Haiku)	380/411	227/307	0.626/**0.572**	86/135	108/188	0.443/0.418
OpenAI	Joint (GPT)	385/409	221/306	**0.635/0.572**	112/160	82/163	0.577/**0.495**
	NER	209/223	398/496	0.344/0.310	43/73	151/250	0.222/0.226
	Coreference	60/79	547/640	0.099/0.110	119/112	75/211	0.613/0.347
	Sequential	202/227	405/492	0.333/0.316	128/137	66/186	0.660/0.424

Coreference system has a low Recall (0.099/0.110) for mentions best explained by external sources, while the NER system has a low Recall (0.222/0.226) for mentions best explained by the context. Both cases are presumably tied to the design goal of tackling one specific mention type over the other. Additionally, the Sequential system has competitive Recall (Context) according to A1J, occupying the third rank and the value is 85.7% of that of the top rank. The Kendall's tau of the two rankings by two annotators on the five LLMs: 0.828 (p-value: 0.022) for recall of external mentions, and 0.467 (p-value: 0.272) for recall of coreference mentions. Only the External mentions show a strong correlation in rankings by either version of annotations.

Table F.2: Results on mention identification with human annotations (relaxed count) as Table F.1, but on the test split.

Model Family	Approach	External			Context		
		TP	FN	Recall	TP	FN	Recall
Human (reference)		203/190	75/388	0.730/0.329	161/174	71/93	0.694/0.652
Claude 3	Joint (Sonnet)	201/244	77/334	0.723/0.422	125/141	107/126	0.539/0.528
	Joint (Haiku)	228/304	50/274	**0.820/0.526**	71/102	161/165	0.306/0.382
OpenAI	Joint (GPT)	228/286	50/292	**0.820**/0.495	107/137	125/130	0.461/0.513
	NER	193/178	85/400	0.694/0.308	38/66	194/201	0.164/0.247
	Coreference	46/56	232/522	0.165/0.097	119/124	113/143	0.513/0.464
	Sequential	193/175	85/403	0.694/0.303	131/150	101/117	**0.565/0.562**

Table F.3: Cohen's κ between LLMs and human annotations for all validation split examples, on classifying External or Context mentions. The two columns correspond to two annotator versions. N is the total number of human-annotated mentions, followed by that of External and Context mentions, separated by a slash.

LLM version	Annotator #1 A1J (N=801, 607/194)	Annotator #2 A1I (N=1,042, 719/323)
Claude 3.5 - Sonnet	0.588	0.417
Claude 3.5 - Haiku	0.519	0.343
GPT-4o	0.539	0.389
GPT-4o-mini	0.619	0.424
GPT-3.5-turbo	0.577	0.374

During the experiments, I also allowed the systems to attribute their extracted mentions into two categories, which provides information to report separate counts based on both human labels and the attributions of types by proposed approaches. Like what human annotators work in the interface Figure 5.3, each of the LLM-based approaches provides class labels on each mention annotated by human as either Context, i.e., a mention that is explainable with context, or External, i.e., a mention that is explainable with external knowledge. Different LLMs annotate the same number of mentions as the human annotations. Since the classification prompt is independently executed on the LLM, separate from the prompts used for contextualization, the classification accuracy reflects the performance of the underlying LLM models. In addition to the Claude 3.5 Sonnet and Haiku versions, I also experiment with two additional LLMs: GPT-4o-mini (a smaller model) and GPT-3.5-turbo (an earlier release). Table F.3 reports the LLM's classification agreement with human annotation using Cohen's Kappa. The results validate that LLM-based approaches have a good to substantial agreement with A1J annotations on classification, and an average good agreement with A1I annotations on classification. The gap between the current agreement and a perfect agreement suggests the influence of two factors: random or systematic human errors and LLM's classification accuracy.

Table F.4 and F.6) both report on the validation split with the separate view of External

Table F.4: Results on mention identification with human annotations (strict count), on the validation split. This table reports on **External** mentions as annotated by human annotators. An LLM needs to attribution the types consistently with human annotations to be counted as a TP. A mention is a Positive of the External type when LLM extracts this mention and attributes its type as External, otherwise Negative.

Model Family	System	External						
		TP	FP	FN	Precision	Recall	F_1	F_2
Human (reference)		434/434	285/173	173/285	0.604/0.715	0.715/0.604	0.655/0.655	0.690/0.623
Claude 3	Joint (Sonnet)	246/245	128/127	361/474	**0.658**/0.659	0.405/0.341	0.502/0.449	0.439/0.377
	Joint (Haiku)	322/337	226/210	285/381	0.588/0.616	**0.530/0.469**	**0.558/0.533**	**0.541/0.493**
OpenAI	Joint (GPT)	266/267	138/134	340/448	**0.658/0.666**	0.439/0.373	0.527/0.478	0.470/0.409
	NER	176/186	170/160	431/533	0.509/0.538	0.290/0.259	0.369/0.349	0.317/0.289
	Coreference	18/20	15/13	589/699	0.545/0.606	0.030/0.028	0.056/0.053	0.037/0.034
	Sequential	143/149	282/272	464/570	0.336/0.354	0.236/0.207	0.277/0.261	0.251/0.226

and Context respectively, but have a strict definition for true positives (TP). It requires consistency between LLM and human attribution of mention types ("strict count"), which give a closeup explanation on mention identification. This definition of Positive is consistent with NER evaluation [175] that considers NER types. Microsoft Azure AI service [1] [175] also has a similar definition for their NER tasks, where a TP (true positive) mention is a mention extracted and correctly predicted for its type. In Table F.4, the Joint (Haiku) system has the top rank for F_1 and F_2 for External mentions. In Table F.6), the Joint (Sonnet) system has the top rank for F_1 and F_2 for Context mentions by A1J annotations, while Joint (GPT) has the top rank by A1I annotations. The top system all stand out with the most

[1] According to the Azure AI documentation.

Table F.5: Results on mention identification for External mentions (strict count), on the test split.

Model Family	Approach	External						
		TP	FP	FN	Precision	Recall	F_1	F_2
Human (reference)		170/170	408/108	108/408	0.294/0.612	0.612/0.294	0.397/0.397	0.503/0.328
Claude 3	Joint (Sonnet)	159/192	116/84	119/386	0.578/**0.696**	0.572/0.332	0.575/0.450	0.573/0.371
	Joint (Haiku)	188/241	243/190	90/337	0.436/0.559	**0.676/0.417**	0.530/**0.478**	**0.609/0.439**
OpenAI	Joint (GPT)	170/191	127/106	108/387	0.572/0.643	0.612/0.330	**0.591**/0.437	0.603/0.366
	NER	166/143	120/143	112/435	0.580/0.500	0.597/0.247	0.589/0.331	0.594/0.275
	Coreference	17/17	9/9	261/561	**0.654**/0.654	0.061/0.029	0.112/0.056	0.075/0.036
	Sequential	156/119	232/264	122/459	0.402/0.311	0.561/0.206	0.468/0.248	0.520/0.221

Table F.6: Results on mention identification with human annotations (strict count), on the validation split. This table reports on **Coreference** mentions as annotated by human annotators. It has the same setup as Table F.4 except it is about Coreference.

Model Family	System	Context						
		TP	FP	FN	Precision	Recall	F_1	F_2
Human (reference)		120/120	203/74	74/203	0.372/0.619	0.619/0.372	0.464/0.464	0.546/0.404
Claude 3	Joint (Sonnet)	128/116	263/275	66/207	**0.327/0.297**	**0.660**/0.359	**0.438**/0.325	**0.548**/0.345
	Joint (Haiku)	79/85	279/273	115/238	0.221/0.237	0.407/0.263	0.286/0.250	0.348/0.258
OpenAI	Joint (GPT)	110/124	318/304	84/199	0.257/0.290	0.567/**0.384**	0.354/**0.330**	0.457/**0.360**
	NER	41/43	124/122	153/280	0.248/0.261	0.211/0.133	0.228/0.176	0.218/0.148
	Coreference	119/110	336/345	75/213	0.262/0.242	0.613/0.341	0.367/0.283	0.483/0.315
	Sequential	126/115	503/514	68/208	0.200/0.183	0.649/0.356	0.306/0.242	0.448/0.299

competitive Recall to their respective annotations, even when they do not achieve the highest Precision. The Kendall's tau between rankings by two annotators are: 1.000 (p-value: 0.003) for F_1 score on external mentions, 1.000 (p-value: 0.003) for F_2 score on external mentions, 0.600 (p-value: 0.136) for F_1 score on context mentions, and 0.733 (p-value: 0.056) for F_2 score on context mentions.

There is "mass movement" of identified mentions between the relaxed count and the strict count. Nevertheless, the two conditions differ only by definition. A sanity check verifies that the number of human-annotated External mentions—calculated as the sum of TP and FN in the External column of the relaxed count Table F.1 (total: 607/719)—matches the corresponding TP and FN sum in the External column of the strict Table F.4. Additionally, the number of all mentions extracted by a particular LLM—calculated as the

Table F.7: Results on mention identification for Context mentions (strict count), on the test split.

Model Family	System	Context						
		TP	FP	FN	Precision	Recall	F_1	F_2
Human (reference)		140/141	127/91	92/126	0.524/0.608	0.603/0.528	0.561/0.565	0.586/0.542
Claude 3	Joint (Sonnet)	120/122	285/283	112/145	**0.296/0.301**	0.517/0.457	**0.377/0.363**	**0.450/0.414**
	Joint (Haiku)	58/76	322/303	174/191	0.153/0.201	0.250/0.285	0.190/0.235	0.222/0.263
OpenAI	Joint (GPT)	104/122	326/308	128/145	0.242/0.284	0.448/0.457	0.314/0.350	0.383/0.407
	NER	33/46	133/120	199/221	0.199/0.277	0.142/0.172	0.166/0.212	0.151/0.186
	Coreference	119/121	353/351	113/146	0.252/0.256	0.513/0.453	0.338/0.327	0.425/0.393
	Sequential	125/131	539/529	107/136	0.188/0.198	**0.539/0.491**	0.279/0.283	0.393/0.379

sum the sum of TP and FP in the overall Table 5.11—matches the sum of TP (External), TP (Context) and FP (External, Context) in the strict Tables F.4 and F.6. Finally, the TP count from Overall class can be larger or equal to the sum of TP count of External and TP count of Context in the strict tables. The former is larger when a system extracts a mention but attributes it to a different class than the human label, which will result in a TP in the relaxed count table, but also one FN and one FP in the respective strict table. Under the External grouping, take A1J annotations for example, the six systems each have 61 (out of 307, 19.9%), 58 (out of 380, 15.3%), 119 (out of 385, 30.9%), 33 (out of 209, 15.8%), 42 (out of 60, 70%), and 59 (out of 202, 29.2%) mentions extracted but incorrectly attributed to the Context type by the system. Under the Context grouping and take A1J annotations, the six systems each have 4 (out of 132, 3.03%), 7 (out of 86, 8.14%), 2 (out of 112, 1.79%), 2 (out of 43, 4.65%), 0, and 2 (out of 128, 1.56%) mentions extracted but incorrectly attributed to the External type by the system. This highlights a difference in mis-classification patterns by LLM: a human-annotated context mention is less likely to be misclassified by LLM as External, whereas a human-annotated external mention is more likely to be misclassified by LLM as Context. Among the systems, the Coreference method not only extracts very few External mentions but when it does, it tends to classify them as Context mentions.

By looking at the two conditions on separate counts and single-value measures, the advantage of the three Joint systems are still clear. Among the three other systems, their rankings are topped by the NER system for external mentions, and topped by the Sequential system for Context mentions in the relaxed sense, or topped by the Coreference system for Context mentions in the strict sense. One explanation is that the Sequential system captures more ambiguous mentions (e.g., those that could qualify as both External and Context, or neither) as Context alongside mentions that have certainty (with a high Recall), whereas the Coreference system for Context mentions focuses on extracting mentions that have certainty (with a high Precision).

Appendix G: Supplementary Experiments on Contextualization using LLM-generated Annotations

This appendix presents supplementary results from an automatic evaluation of elaboration quality on the test split, using LLM-generated annotations.

Table G.1: Results on elaboration quality using LLM-generated annotations by Joint (Sonnet) on the test split.

System	TP			All Relevant (TP + FN) N=679	
	N	BLEU-4	MNLI	BLEU-4	MNLI
Joint (GPT)	463	0.173	0.503	**0.118**	**0.343**
NER	252	0.024	0.197	0.009	0.073
Coreference	297	**0.200**	**0.589**	0.087	0.257
Sequential	436	0.139	0.476	0.090	0.306

Table G.2: Results on elaboration quality using LLM-generated annotations by Joint (Haiku) on the test split.

System	TP			All Relevant (TP + FN) N=810	
	N	BLEU-4	MNLI	BLEU-4	MNLI
Joint (GPT)	488	**0.095**	0.451	**0.057**	**0.272**
NER	333	0.025	0.235	0.010	0.096
Coreference	231	0.085	**0.548**	0.024	0.156
Sequential	436	0.060	0.411	0.032	0.221

Bibliography

[1] Abraham Lincoln Presidential Museum and Library. https://presidentlincoln.illinois.gov/. [Online; accessed 26-January-2021].

[2] Mohammad Mahdi Abootorabi, Amirhosein Zobeiri, Mahdi Dehghani, Mohammadali Mohammadkhani, Bardia Mohammadi, Omid Ghahroodi, Mahdieh Soleymani Baghshah, and Ehsaneddin Asgari. Ask in any modality: A comprehensive survey on multimodal retrieval-augmented generation. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Findings of the Association for Computational Linguistics: ACL 2025*, pages 16776–16809, Vienna, Austria, July 2025. Association for Computational Linguistics.

[3] Herman Aguinis and Angelo M Solarino. Transparency and replicability in qualitative research: The case of interviews with elite informants. *Strategic Management Journal*, 40(8):1291–1315, 2019.

[4] Lauren Algee. Volunteers leverage ocr to transcribe library of congress digital collections. https://blogs.loc.gov/thesignal/2025/05/volunteers-ocr/, 2025. [Online; accessed 39-July-2022].

[5] Jacopo Amidei, Paul Piwek, and Alistair Willis. The use of rating and Likert scales in natural language generation human evaluation tasks: A review and some recommendations. In Kees van Deemter, Chenghua Lin, and Hiroya Takamura, editors, *Proceedings of the 12th International Conference on Natural Language Generation*, pages 397–402, Tokyo, Japan, October–November 2019. Association for Computational Linguistics.

[6] Ion Androutsopoulos and Prodromos Malakasiotis. A survey of paraphrasing and textual entailment methods. *Journal of Artificial Intelligence Research*, 38:135–187, 2010.

[7] Ron Artstein, Alesia Gainer, Kallirroi Georgila, Anton Leuski, Ari Shapiro, and David Traum. New Dimensions in Testimony Demonstration. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 32–36, 2016.

[8] Ron Artstein and Massimo Poesio. Inter-coder agreement for computational linguistics. *Computational linguistics*, 34(4):555–596, 2008.

[9] Ron Artstein, David Traum, Oleg Alexander, Anton Leuski, Andrew Jones, Kallirroi Georgila, Paul Debevec, William Swartout, Heather Maio, and Stephen Smith. Time-offset Interaction with a Holocaust Survivor. In *Proceedings of the 19th international conference on Intelligent User Interfaces*, pages 163–168, 2014.

[10] Javed A Aslam and Mark Montague. Bayes optimal metasearch: a probabilistic model for combining the results of multiple retrieval systems. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and development in Information Retrieval*, pages 379–381, 2000.

[11] Mark Auslander. Touching the past: materializing time in traumatic "living history" reenactments. *Signs and Society*, 1(1):161–183, 2013.

[12] Stephen Bach, Victor Sanh, Zheng Xin Yong, Albert Webson, Colin Raffel, Nihal V. Nayak, Abheesht Sharma, Taewoon Kim, M Saiful Bari, Thibault Fevry, Zaid Alyafeai, Manan Dey, Andrea Santilli, Zhiqing Sun, Srulik Ben-david, Canwen Xu, Gunjan Chhablani, Han Wang, Jason Fries, Maged Al-shaibani, Shanya Sharma, Urmish Thakker, Khalid Almubarak, Xiangru Tang, Dragomir Radev, Mike Tian-jian Jiang, and Alexander Rush. PromptSource: An integrated development environment and repository for natural language prompts. In Valerio Basile, Zornitsa Kozareva, and Sanja Stajner, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 93–104, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[13] Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, et al. MS MARCO: A Human Generated Machine Reading Comprehension Dataset. *arXiv preprint arXiv:1611.09268*, 2016.

[14] Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the ACL workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72, 2005.

[15] Siddhartha Banerjee, Prakhar Biyani, and Kostas Tsioutsiouliklis. Transforming Chatbot Responses to Mimic Domain-specific Linguistic Styles. In *Second Workshop on Chatbots and Conversational Agent Technologies*, 2016.

[16] Liam Bannon, Steve Benford, John Bowers, and Christian Heath. Hybrid design creates innovative museum experiences. *Commun. ACM*, 48(3):62–65, March 2005.

[17] Toby Glenn Bates. *The Reagan rhetoric: History and memory in 1980s America*. Northern Illinois University Press, 2011.

[18] Gordon Bell and Jim Gray. Digital Immortality. *Commun. ACM*, 44(3):28–31, March 2001.

[19] Linda Bell and Joakim Gustafson. Utterance Types in the August Dialogues. In *ESCA Tutorial and Research Workshop (ETRW) on Interactive Dialogue in Multi-Modal Systems*, 1999.

[20] Anja Belz and Eric Kow. Comparing rating scales and preference judgements in language evaluation. In John Kelleher, Brian Mac Namee, and Ielka van der Sluis, editors, *Proceedings of the 6th International Natural Language Generation Conference*. Association for Computational Linguistics, July 2010.

[21] J Martin Bland and Douglas G Altman. Multiple significance tests: the bonferroni method. *BMJ*, 310(6973):170, 1995.

[22] Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. A large annotated corpus for learning natural language inference, 2015.

[23] Henry William Brands. *Reagan: The Life*. Anchor, 2016.

[24] Svend Brinkmann. *Qualitative Interviewing*. Oxford University Press, 2013.

[25] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901, 2020.

[26] Rose Buchanan, Keau George, Taylor Gibson, Eric Hung, Daria Labinsky, Diana E Marsh, Rachel Menyuk, Lotus Norton-Wisla, Selena Ortego-Chiolero, Nathan Sowry, et al. Toward Inclusive Reading Rooms: Recommendations for decolonizing practices and welcoming indigenous researchers. *Archival Outlook*, 2021.

[27] Chris Buckley, Darrin Dimmick, Ian Soboroff, and Ellen Voorhees. Bias and the limits of pooling for large collections. *Information retrieval*, 10:491–508, 2007.

[28] Chris Buckley and Ellen M Voorhees. Retrieval evaluation with incomplete information. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 25–32, 2004.

[29] Francesco Cafaro and Stella A. Ress. Time Travelers: Mapping museum visitors across time and space. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct*, UbiComp '16, page 1492–1497, New York, NY, USA, 2016. Association for Computing Machinery.

[30] Alice Cai, Ian Arawjo, and Elena L Glassman. Antagonistic ai. *arXiv preprint arXiv:2402.07350*, 2024.

[31] John Carroll, Guido Minnen, Yvonne Canning, Siobhan Devlin, and John Tait. Practical Simplification of English Newspaper Text to Assist Aphasic Readers. In *Proceedings of the AAAI-98 Workshop on Integrating Artificial Intelligence and Assistive Technology*, pages 7–10. Citeseer, 1998.

[32] Lucy Chabot. Nixon Library Technology Lets Visitors 'Interview' Him. `https://www.latimes.com/archives/la-xpm-1990-07-21-mn-346-story.html`, 1990. [Online; accessed 10-January-2022].

[33] Boxing Chen and Colin Cherry. A systematic comparison of smoothing techniques for sentence-level bleu. In *Proceedings of the ninth workshop on statistical machine translation*, pages 362–367, 2014.

[34] Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. Reading Wikipedia to Answer Open-domain Questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1870–1879, 2017.

[35] Cheng-Han Chiang and Hung-yi Lee. Can large language models be an alternative to human evaluations? In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15607–15631, Toronto, Canada, July 2023. Association for Computational Linguistics.

[36] Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46, 1960.

[37] Charles Cole. Inducing Expertise in History Doctoral Students via Information Retrieval Design. *The Library Quarterly: Information, Community, Policy*, 70(1):86–109, 2000.

[38] Richard J Cox. America's pyramids: Presidents and their libraries. *Government Information Quarterly*, 19(1):45–75, 2002.

[39] J Shane Culpepper, Fernando Diaz, and Mark D Smucker. Research frontiers in information retrieval: Report from the third strategic workshop on information retrieval in lorne (swirl 2018). In *ACM SIGIR Forum*, volume 52, pages 34–90. ACM New York, NY, USA, 2018.

[40] Dennis A Daellenbach. The ronald reagan presidential library. *Government Information Quarterly*, 11(1):23–35, 1994.

[41] Jeffrey Dalton, Chenyan Xiong, and Jamie Callan. Trec cast 2019: The conversational assistance track overview, 2020.

[42] Jeffrey Dalton, Chenyan Xiong, and Jamie Callan. Cast 2020: The conversational assistance track overview. In *In Proceedings of TREC*, 2021.

[43] Jeffrey Dalton, Chenyan Xiong, and Jamie Callan. TREC cast 2021: The conversational assistance track overview. In Ian Soboroff and Angela Ellis, editors, *Proceedings of the Thirtieth Text REtrieval Conference, TREC 2021, online, November 15-19, 2021*, volume 500-335 of *NIST Special Publication*. National Institute of Standards and Technology (NIST), 2021.

[44] Cody Delistraty. How a chatbot helped me talk to my dead mother. `https://www.wsj.com/tech/ai/talking-to-my-deceased-mother-e5586300`, 2024.

[45] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[46] Lee R Dice. Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302, 1945.

[47] Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. Wizard of wikipedia: Knowledge-powered conversational agents, 2019.

[48] Emma Dixon and Amanda Lazar. Approach Matters: Linking practitioner approaches to technology design for people with dementia. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–15, 2020.

[49] Mark Dras. Squibs: Evaluating human pairwise preference judgments. *Computational Linguistics*, 41(2):309–317, June 2015.

[50] Aparna Elangovan, Ling Liu, Lei Xu, Sravan Babu Bodapati, and Dan Roth. ConSiDERS-the-human evaluation framework: Rethinking human evaluation for generative large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1137–1160, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[51] Ahmed Elgohary, Denis Peskov, and Jordan Boyd-Graber. Can you unpack that? learning to rewrite questions-in-context. In *Empirical Methods in Natural Language Processing*, 2019.

[52] John D. Emerson and Gary A. Simon. Another look at the sign test when ties are present: The problem of confidence intervals. *The American Statistician*, 33(3):140–142, 1979.

[53] Emmett Till Interpretive Center. Emmett Till Interpretive Center. `http://archives.cpajournal.com/old/11287210.htm`. [Online; accessed 12-April-2021].

[54] Meredith R Evans. Presidential libraries going digital. *The Public Historian*, 40(2):116–121, 2018.

[55] Alexander Fabbri, Chien-Sheng Wu, Wenhao Liu, and Caiming Xiong. QAFactEval: Improved QA-based factual consistency evaluation for summarization. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz, editors, *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2587–2601, Seattle, United States, July 2022. Association for Computational Linguistics.

[56] Raya Fidel, Rachel K Davies, Mary H Douglass, Jenny K Holder, Carla J Hopkins, Elisabeth J Kushner, Bryan K Miyagishima, and Christina D Toney. A Visit to the Information Mall: Web searching behavior of high school students. *Journal of the American Society for Information Science*, 50(1):24–37, 1999.

[57] Bruce Fraser. What are discourse markers? *Journal of Pragmatics*, 31(7):931–952, 1999. Pragmatics: The Loaded Discipline?

[58] Zhenxin Fu, Xiaoye Tan, Nanyun Peng, Dongyan Zhao, and Rui Yan. Style Transfer in Text: Exploration and evaluation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), Apr. 2018.

[59] Pascale Fung and Percy Cheung. Multi-level bootstrapping for extracting parallel sentences from a quasi-comparable corpus. In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, pages 1051–1057, 2004.

[60] Demetria Gallegos. Ai can re-create your loved ones after they die. is that good or bad? `https://www.wsj.com/lifestyle/relationships/ai-ethics-morals-dead-people-grief-f8e3c7f7`, May 2024.

[61] Michel Galley, Chris Brockett, Alessandro Sordoni, Yangfeng Ji, Michael Auli, Chris Quirk, Margaret Mitchell, Jianfeng Gao, and Bill Dolan. deltaBLEU: A discriminative metric for generation tasks with intrinsically diverse targets. In Chengqing Zong and Michael Strube, editors, *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 445–450, Beijing, China, July 2015. Association for Computational Linguistics.

[62] Sudeep Gandhe, Andrew Gordon, Anton Leuski, David R Traum, and Douglas W Oard. First steps toward linking dialogues: Mediating between free-text questions and pre-recorded video answers. In *Transformational Science And Technology For The Current And Future Force*, pages 331–338. World Scientific, 2006.

[63] Mingqi Gao, Xinyu Hu, Li Lin, and Xiaojun Wan. Analyzing and evaluating correlation measures in NLG meta-evaluation. In Luis Chiruzzo, Alan Ritter, and Lu Wang, editors, *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2199–2222, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics.

[64] Ning Gao, Mossaab Bagdouri, and Douglas W Oard. Pearson rank: a head-weighted gap-sensitive score-based correlation coefficient. In *Proceedings of the 39th International ACM SIGIR conference on research and development in information retrieval*, pages 941–944, 2016.

[65] Carlos Gemmell and Jeffrey Dalton. Glasgow representation and information learning lab (GRILL) at the conversational assistance track 2020. In Ellen M. Voorhees and Angela Ellis, editors, *Proceedings of the Twenty-Ninth Text REtrieval Conference, TREC 2020, Virtual Event [Gaithersburg, Maryland, USA], November 16-20, 2020*, volume 1266 of *NIST Special Publication*. National Institute of Standards and Technology (NIST), 2020.

[66] Alan S. Glazer. Auditing Museum Collections. `http://archives.cpajournal.com/old/11287210.htm`, 1991. [Online; accessed 12-April-2021].

[67] Mark Greene and Dennis Meissner. More Product, Less Process: Revamping traditional archival processing. *The American Archivist*, 68(2):208–263, 2005.

[68] Cynthia Griffin. The Museum Library. *University Museum Bulletin*, 19(2):23–27, 1955.

[69] Joakim Gustafson, Nikolaj Lindberg, and Magnus Lundeberg. The August Spoken Dialogue System. In *Sixth European Conference on Speech Communication and Technology*, 1999.

[70] Maria CR Harrington. Connecting user experience to learning in an evaluation of an immersive, interactive, multimodal augmented reality virtual diorama in a natural history museum & the importance of story. In *2020 6th International Conference of the Immersive Learning Research Network (ILRN)*, pages 70–78. IEEE, 2020.

[71] Marjorie L Harth. Learning from Museums with Indigenous Collections: Beyond repatriation. *Curator: the museum journal*, 42(4):274–284, 1999.

[72] Junxian He, Xinyi Wang, Graham Neubig, and Taylor Berg-Kirkpatrick. A probabilistic formulation of unsupervised text style transfer. *arXiv preprint arXiv:2002.03912*, 2020.

[73] Alex Hern. Amazon's alexa could turn dead loved ones' voices into digital assistant. https://www.theguardian.com/technology/2022/jun/23/amazon-alexa-could-turn-dead-loved-ones-digital-assistant, 2022. [Online; accessed 1-July-2025].

[74] Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. spaCy: Industrial-strength Natural Language Processing in Python, 2020.

[75] Brian Hosler, Davide Salvi, Anthony Murray, Fabio Antonacci, Paolo Bestagini, Stefano Tubaro, and Matthew C Stamm. Do deepfakes feel emotions? a semantic approach to detecting deepfakes via emotional inconsistencies. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1013–1022, 2021.

[76] Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P. Xing. Toward controlled generation of text. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1587–1596. PMLR, 06–11 Aug 2017.

[77] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55, January 2025.

[78] Kalervo Järvelin and Jaana Kekäläinen. Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems*, 20(4):422–446, October 2002.

[79] Grishma Jena, Mansi Vashisht, Abheek Basu, Lyle Ungar, and Joao Sedoc. Enterprise to Computer: Star trek chatbot. *arXiv preprint arXiv:1708.00818*, 2017.

[80] Harsh Jhamtani, Varun Gangal, Eduard Hovy, and Eric Nyberg. Shakespearizing modern language using copy-enriched sequence to sequence models. In *Proceedings of the Workshop on Stylistic Variation*, pages 10–19, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

[81] Jimmy Carter Presidential Library and Museum. Oral histories - Research - the Jimmy Carter Presidential Library and Museum. https://www.jimmycarterlibrary.gov/research/oral$_{\text{h}}$istories#/assets/documents/oral$_{\text{h}}$istories/exit$_{\text{i}}$nterviews/. [Online; accessed 10-January-2022].

[82] Di Jin, Zhijing Jin, Zhiting Hu, Olga Vechtomova, and Rada Mihalcea. Deep learning for text style transfer: A survey, 2021.

[83] Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, November 2020.

[84] Maurice G Kendall. A new measure of rank correlation. *Biometrika*, 30(1-2):81–93, 1938.

[85] Omar Khattab and Matei Zaharia. Colbert: Efficient and effective passage search via contextualized late interaction over BERT, 2020.

[86] Kalpesh Krishna, John Wieting, and Mohit Iyyer. Reformulating unsupervised style transfer as paraphrase generation. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 737–762, Online, November 2020. Association for Computational Linguistics.

[87] Logan Kugler. Raising the dead with ai. *Commun. ACM*, 67(6):17–19, may 2024.

[88] Steinar Kvale. *Doing Interviews*. Sage, 2008.

[89] Ksenia Lagutina, Nadezhda Lagutina, Elena Boychuk, Inna Vorontsova, Elena Shliakhtina, Olga Belyaeva, Ilya Paramonov, and PG Demidov. A survey on stylometric text features. In *2019 25th Conference of Open Innovations Association (FRUCT)*, pages 184–195. IEEE, 2019.

[90] J. Richard Landis and Gary G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174, 1977.

[91] Samuel Läubli, Rico Sennrich, and Martin Volk. Has machine translation achieved human parity? a case for document-level evaluation. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4791–4796, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

[92] Dawn Lawrie, James Mayfield, Douglas W Oard, Eugene Yang, Suraj Nair, and Petra Galuščáková. Hc3: A suite of test collections for clir evaluation over informal text. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2880–2889, 2023.

[93] Anton Leuski, Jarrell Pair, David Traum, Peter J McNerney, Panayiotis Georgiou, and Ronakkumar Patel. How to Talk to a Hologram. In *Proceedings of the 11th International Conference on Intelligent User Interfaces*, pages 360–362, 2006.

[94] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. *arXiv preprint arXiv:2005.11401*, 2020.

[95] Haitao Li, Qian Dong, Junjie Chen, Huixue Su, Yujia Zhou, Qingyao Ai, Ziyi Ye, and Yiqun Liu. LLMs-as-judges: a comprehensive survey on llm-based evaluation methods. *arXiv preprint arXiv:2412.05579*, 2024.

[96] Jiwei Li, Michel Galley, Chris Brockett, Georgios Spithourakis, Jianfeng Gao, and William B Dolan. A persona-based neural conversation model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 994–1003, 2016.

[97] Juncen Li, Robin Jia, He He, and Percy Liang. Delete, retrieve, generate: a simple approach to sentiment and style transfer. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1865–1874, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[98] David Lillis, Fergus Toolan, Rem Collier, and John Dunnion. Probfuse: a probabilistic approach to data fusion. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 139–146, 2006.

[99] Bill Yuchen Lin, Wangchunshu Zhou, Ming Shen, Pei Zhou, Chandra Bhagavatula, Yejin Choi, and Xiang Ren. CommonGen: A constrained text generation challenge for generative commonsense reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1823–1840, Online, November 2020. Association for Computational Linguistics.

[100] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.

[101] Sheng-Chieh Lin, Jheng-Hong Yang, Rodrigo Nogueira, Ming-Feng Tsai, Chuan-Ju Wang, and Jimmy Lin. Conversational question reformulation via sequence-to-sequence architectures and pretrained language models, 2020.

[102] Shayne Longpre, Robert Mahari, Ariel Lee, Campbell Lund, Hamidah Oderinwale, William Brannon, Nayan Saxena, Naana Obeng-Marnu, Tobin South, Cole Hunter, Kevin Klyman, Christopher Klamm, Hailey Schoelkopf, Nikhil Singh, Manuel Cherep, Ahmad Mustafa Anis, An Dinh, Caroline Chitongo, Da Yin, Damien Sileo, Deividas Mataciunas, Diganta Misra, Emad Alghamdi, Enrico Shippole, Jianguo Zhang, Joanna Materzynska, Kun Qian, Kush Tiwary, Lester Miranda, Manan Dey, Minnie Liang, Mohammed Hamdy, Niklas Muennighoff, Seonghyeon Ye, Seungone

Kim, Shrestha Mohanty, Vipul Gupta, Vivek Sharma, Vu Minh Chien, Xuhui Zhou, Yizhi Li, Caiming Xiong, Luis Villa, Stella Biderman, Hanlin Li, Daphne Ippolito, Sara Hooker, Jad Kabbara, and Sandy Pentland. Consent in crisis: The rapid decline of the ai data commons. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 108042–108087. Curran Associates, Inc., 2024.

[103] Aman Madaan, Amrith Setlur, Tanmay Parekh, Barnabas Poczos, Graham Neubig, Yiming Yang, Ruslan Salakhutdinov, Alan W Black, and Shrimai Prabhumoye. Politeness transfer: A tag and generate approach. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1869–1881, Online, July 2020. Association for Computational Linguistics.

[104] Jonathan Mallinson, Aliaksei Severyn, Eric Malmi, and Guillermo Garrido. FELIX: Flexible text editing through tagging and insertion. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1244–1255, Online, November 2020. Association for Computational Linguistics.

[105] Eric Malmi, Aliaksei Severyn, and Sascha Rothe. Unsupervised text style transfer with masked language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8671–8680, 2020.

[106] Satota Mandal, Bratati Ghosh, Sunen Chakraborty, and Ruchira Naskar. Can deepfakes mimic human emotions? a perspective on synthesia videos. In *TENCON 2024 - 2024 IEEE Region 10 Conference (TENCON)*, pages 306–309, 2024.

[107] Donald Marinelli and Scott Stevens. Synthetic Interviews: the art of creating a'dyad'between humans and machine-based characters. In *Proceedings 1998 IEEE 4th Workshop Interactive Voice Technology for Telecommunications Applications. IVTTA'98 (Cat. No. 98TH8376)*, pages 43–48. IEEE, 1998.

[108] Pierre-Emmanuel Mazare, Samuel Humeau, Martin Raison, and Antoine Bordes. Training millions of personalized dialogue agents. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2775–2779, 2018.

[109] Patrick McCarthy. "living history" as the "real thing": a comparative analysis of the modern mountain man rendezvous, renaissance fairs, and civil war reenactments. *ETC: A Review of General Semantics*, 71(2):106–123, 2014.

[110] Carole McGranahan. A Presidential Archive of Lies: Racism, twitter, and a history of the present. *International Journal of communication*, 13:19, 2019.

[111] Massimo Melucci. On rank correlation in information retrieval evaluation. *SIGIR Forum*, 41(1):18–33, June 2007.

[112] Rada Mihalcea and Andras Csomai. Wikify! Linking documents to encyclopedic knowledge. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 233–242, 2007.

[113] Marvin Minoff, John Birt, and Jorn H. Winther. David Frost Interviews Richard Nixon, 1977. [Online; accessed 10-January-2022].

[114] Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi. Cross-task generalization via natural language crowdsourcing instructions. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3470–3487, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[115] National Archives. Strategy for digitizing archival materials. `https://www.archives.gov/digitization/strategy.html`, 2014. [Online; accessed 12-April-2021].

[116] Wendy Nielsen and Jennifer Yeo. Introduction to the special issue: Multimodal meaning-making in science. *Research in Science Education*, 52(3):751–754, 2022.

[117] Rodrigo Nogueira and Kyunghyun Cho. Passage re-ranking with BERT. *arXiv preprint arXiv:1901.04085*, 2019.

[118] Rodrigo Nogueira, Zhiying Jiang, Ronak Pradeep, and Jimmy Lin. Document ranking with a pretrained sequence-to-sequence model. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 708–718, Online, November 2020. Association for Computational Linguistics.

[119] Rodrigo Nogueira, Wei Yang, Kyunghyun Cho, and Jimmy Lin. Multi-stage document ranking with BERT. *arXiv preprint arXiv:1910.14424*, 2019.

[120] Phil Ostheimer, Mayank Nagda, Marius Kloft, and Sophie Fellenz. A call for standardization and validation of text style transfer evaluation. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10791–10815, Toronto, Canada, July 2023. Association for Computational Linguistics.

[121] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In Pierre Isabelle, Eugene Charniak, and Dekang Lin, editors, *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics.

[122] Simon Parkin. Back-up Brains: The era of digital immortality. https://www.bbc.com/future/article/20150122-the-secret-to-immortality, 2015. [Online; accessed 10-January-2022].

[123] Karl Pearson. Note on regression and inheritance in the case of two parents. *Proceedings of the royal society of London*, 58(347-352):240–242, 1895.

[124] Fabio Petroni, Aleksandra Piktus, Angela Fan, Patrick Lewis, Majid Yazdani, Nicola De Cao, James Thorne, Yacine Jernite, Vladimir Karpukhin, Jean Maillard, Vassilis Plachouras, Tim Rocktäschel, and Sebastian Riedel. KILT: a benchmark for knowledge intensive language tasks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2523–2544, Online, June 2021. Association for Computational Linguistics.

[125] Trevor J Pinch and Wiebe E Bijker. The Social Construction of Facts and Artefacts: Or how the sociology of science and the sociology of technology might benefit each other. *Social studies of science*, 14(3):399–441, 1984.

[126] Xin Qian and Douglas W Oard. Full-collection search with passage and document evidence: Maryland at the TREC 2021 conversational assistance track. In *TREC*, 2021.

[127] Xin Qian, Douglas W Oard, and Joel Chan. Conversational interaction with historical figures: What's it good for? In *International Conference on Information*, pages 32–49. Springer, 2022.

[128] Filip Radlinski and Nick Craswell. A theoretical framework for conversational search. In *Proceedings of the 2017 Conference on Human Information Interaction and Retrieval*, pages 117–126, 2017.

[129] Sudha Rao and Joel Tetreault. Dear Sir or Madam, May I Introduce the GYAFC Dataset: Corpus, benchmarks and metrics for formality style transfer. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 129–140, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[130] Sravana Reddy and Kevin Knight. Obfuscating gender in social media writing. In *Proceedings of the First Workshop on NLP and Computational Social Science*, pages 17–26, Austin, Texas, November 2016. Association for Computational Linguistics.

[131] Emily Reif, Daphne Ippolito, Ann Yuan, Andy Coenen, Chris Callison-Burch, and Jason Wei. A recipe for arbitrary text style transfer with large language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational*

Linguistics (Volume 2: Short Papers), pages 837–848, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[132] Ehud Reiter. A structured review of the validity of bleu. *Computational Linguistics*, 44(3):393–401, 09 2018.

[133] S. Ress, F. Cafaro, D. Bora, D. Prasad, and D. Soundarajan. Mapping History: Orienting museum visitors across time and space. *Journal on Computing and Cultural Heritage*, 11(3), August 2018.

[134] Van Rijsbergen. Information retrieval; ; butterworth, 1978. *J. librariansh.*, 11:237, 1979.

[135] Russell L. Riley. Presidential Oral History: The Clinton presidential history project. *The Oral History Review*, 34(2):81–106, 2007.

[136] Martine Aliana Rothblatt. *Virtually Human: The Promise—and the Peril—of Digital Immortality*. Macmillan, 2014.

[137] Maggi Savin-Baden and David Burden. Digital Immortality and Virtual Humans. *Postdigital Science and Education*, 1(1):87–103, 2019.

[138] Maggi Savin-Baden, David Burden, and Helen Taylor. The ethics and impact of digital immortality. *Knowledge Cultures*, 5(2):178–196, 2017.

[139] Deborah Schiffrin. *Discourse markers*. Number 5. Cambridge University Press, 1987.

[140] Patrick Schober, Christa Boer, and Lothar A Schwarte. Correlation coefficients: appropriate use and interpretation. *Anesthesia & analgesia*, 126(5):1763–1768, 2018.

[141] Lawrence C Schourup. *Common discourse particles in English conversation*. Routledge, 2016.

[142] Tianxiao Shen, Tao Lei, Regina Barzilay, and Tommi Jaakkola. Style transfer from non-parallel text by cross-alignment. *Advances in Neural Information Processing Systems*, 30, 2017.

[143] Wei Shen, Jianyong Wang, and Jiawei Han. Entity linking with a knowledge base: Issues, techniques, and solutions. *IEEE Transactions on Knowledge and Data Engineering*, 27(2):443–460, 2014.

[144] Kiron K Skinner, Annelise Anderson, and Martin Anderson. *Reagan: A life in letters*. Simon and Schuster, 2004.

[145] Eric Michael Smith, Diana Gonzalez-Rico, Emily Dinan, and Y-Lan Boureau. Controlling style in generated dialogue. *CoRR*, abs/2009.10855, 2020.

[146] Mark D. Smucker, James Allan, and Ben Carterette. A comparison of statistical significance tests for information retrieval evaluation. In *Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management*, CIKM '07, page 623–632, New York, NY, USA, 2007. Association for Computing Machinery.

[147] Haoyu Song, Yan Wang, Wei-Nan Zhang, Xiaojiang Liu, and Ting Liu. Generate, Delete and Rewrite: A three-stage framework for improving persona consistency of dialogue generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5821–5831, Online, July 2020. Association for Computational Linguistics.

[148] Charles Spearman. The proof and measurement of association between two things. 1961.

[149] Byron Spice and Eric Sloss. Carnegie Mellon's Synthetic Interview Technology Enables Virtual Chats with Darwin's Ghost. `https://www.cs.cmu.edu/news/2009/carnegie-mellons-synthetic-interview-technology-enables-virtual-chats-darwins-ghost`, 2009. [Online; accessed 23-February-2021].

[150] Neha Srikanth and Junyi Jessy Li. Elaborative simplification: content addition and explanation generation in text simplification. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 5123–5137, Online, August 2021. Association for Computational Linguistics.

[151] Susan Leigh Star and James R Griesemer. Institutional ecology, translations' and boundary objects: Amateurs and professionals in berkeley's museum of vertebrate zoology, 1907-39. *Social studies of science*, 19(3):387–420, 1989.

[152] Bakhtiyar Syed, Gaurav Verma, Balaji Vasan Srinivasan, Anandhavelu Natarajan, and Vasudeva Varma. Adapting language models for non-parallel author-stylized rewriting. *AAAI Press*, 34:9008–9015, 2020.

[153] Tianyi Tang, Hongyuan Lu, Yuchen Jiang, Haoyang Huang, Dongdong Zhang, Xin Zhao, Tom Kocmi, and Furu Wei. Not all metrics are guilty: Improving NLG evaluation by diversifying references. In Kevin Duh, Helena Gomez, and Steven Bethard, editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6596–6610, Mexico City, Mexico, June 2024. Association for Computational Linguistics.

[154] The Dalí (Salvador Dalí Museum). Dalí Museum. `https://thedali.org/about-the-museum/center-avant-garde/research-resources/`. [Online; accessed 12-April-2021].

[155] Paul Thomas, Seth Spielman, Nick Craswell, and Bhaskar Mitra. Large language models can accurately predict searcher preferences. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, page 1930–1940, New York, NY, USA, 2024. Association for Computing Machinery.

[156] Yufei Tian, Tenghao Huang, Miri Liu, Derek Jiang, Alexander Spangher, Muhao Chen, Jonathan May, and Nanyun Peng. Are large language models capable of generating human-level narratives? In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17659–17681, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[157] Angela Tinwell, Mark Grimshaw, Debbie Abdel Nabi, and Andrew Williams. Facial expression of emotion and perception of the uncanny valley in virtual characters. *Computers in Human behavior*, 27(2):741–749, 2011.

[158] S. M Towhidul Islam Tonmoy, S M Mehedi Zaman, Vinija Jain, Anku Rani, Vipula Rawte, Aman Chadha, and Amitava Das. A comprehensive survey of hallucination mitigation techniques in large language models, 2024.

[159] David Traum, Andrew Jones, Kia Hays, Heather Maio, Oleg Alexander, Ron Artstein, Paul Debevec, Alesia Gainer, Kallirroi Georgila, Kathleen Haase, Karen Jungblut, Anton Leuski, Stephen Smith, and William Swartout. New Dimensions in Testimony: Digitally preserving a holocaust survivor's interactive storytelling. In Henrik Schoenau-Fog, Luis Emilio Bruni, Sandy Louchart, and Sarune Baceviciute, editors, *Interactive Storytelling*, pages 269–281, Cham, 2015. Springer International Publishing.

[160] National Institutes of Health U.S.National Library of Medicine. Plain language adaptation of biomedical abstracts (PLABA) at trec 2024. [Online; accessed 1-April-2025].

[161] Victoria Van Hyning, Lauren Algee, Mason Jones, Carlyn Osborn, Trevor Owens, Lauren Seroka, and Abby Shelton. By the people crowdsourcing datasets from the library of congress. *Journal of Open Humanities Data*, 8, 2022.

[162] Victoria Anne Van Hyning and Mason A Jones. Data's destinations: Three case studies in crowdsourced transcription data management and dissemination. *Startwords*, (2), 2021.

[163] Tom Vasich. What's next: The future of museums. https://news.uci.edu/2020/08/17/whats-next-the-future-of-museums/, 2020. [Online; accessed 28-March-2021].

[164] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA, 2017. Curran Associates Inc.

[165] E Voorhees, Narendra K Gupta, and Ben Johnson-Laird. The collection fusion problem. *NIST Special Publication*, pages 95–95, 1995.

[166] Hao Wang, Zhengdong Lu, Hang Li, and Enhong Chen. A Dataset for Research on Short-text Conversations. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 935–945, Seattle, Washington, USA, October 2013. Association for Computational Linguistics.

[167] Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Atharva Naik, Arjun Ashok, Arut Selvan Dhanasekaran, Anjana Arunkumar, David Stap, Eshaan Pathak, Giannis Karamanolakis, Haizhi Lai, Ishan Purohit, Ishani Mondal, Jacob Anderson, Kirby Kuznia, Krima Doshi, Kuntal Kumar Pal, Maitreya Patel, Mehrad Moradshahi, Mihir Parmar, Mirali Purohit, Neeraj Varshney, Phani Rohitha Kaza, Pulkit Verma, Ravsehaj Singh Puri, Rushang Karia, Savan Doshi, Shailaja Keyur Sampat, Siddhartha Mishra, Sujan Reddy A, Sumanta Patro, Tanay Dixit, and Xudong Shen. Super-NaturalInstructions: Generalization via declarative instructions on 1600+ NLP tasks. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5085–5109, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

[168] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2022. Curran Associates Inc.

[169] Paul Williams. *The personalization of loss in memorial museums*. Oxford University Press Oxford, 2017.

[170] Rebecca Wright and Alfred M. Worden. NASA Johnson Space Center Oral History Project Edited Oral History Transcript.`https://historycollection.jsc.nasa.gov/JSCHistoryPortal/history/oral_histories/WordenAM/WordenAM_5-26-00.htm`, 26 May 2000. [Online; accessed 26-January-2021].

[171] Yating Wu, William Sheffield, Kyle Mahowald, and Junyi Jessy Li. Elaborative simplification as implicit questions under discussion. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods*

in Natural Language Processing, pages 5525–5537, Singapore, December 2023. Association for Computational Linguistics.

[172] Qiongkai Xu, Chenchen Xu, and Lizhen Qu. ALTER: Auxiliary text rewriting tool for natural language generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*, pages 13–18, Hong Kong, China, November 2019. Association for Computational Linguistics.

[173] Wei Xu, Alan Ritter, Bill Dolan, Ralph Grishman, and Colin Cherry. Paraphrasing for style. In *Proceedings of COLING 2012*, pages 2899–2914, Mumbai, India, December 2012. The COLING 2012 Organizing Committee.

[174] Yi Yang, Wen-tau Yih, and Christopher Meek. WikiQA: A challenge dataset for open-domain question answering. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2013–2018, 2015.

[175] Juntao Yu, Bernd Bohnet, and Massimo Poesio. Named entity recognition as dependency parsing. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6470–6476, Online, July 2020. Association for Computational Linguistics.

[176] Wenhao Yu, Chenguang Zhu, Zaitang Li, Zhiting Hu, Qingyun Wang, Heng Ji, and Meng Jiang. A survey of knowledge-enhanced text generation. *arXiv preprint arXiv:2010.04389*, 2020.

[177] Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[178] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. BERTScore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019.

[179] Yinhe Zheng, Rongsheng Zhang, Minlie Huang, and Xiaoxi Mao. A pre-training based personalized dialogue generation model with persona-sparse data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9693–9700, 2020.