ABSTRACT

Title of Dissertation: TOWARDS AUTOMATED CONTRACT ANALYSIS: APPLYING LANGUAGE MODELS TO RISK IDENTIFICATION IN THE CONTEXT OF PUBLIC-PRIVATE PARTNERSHIPS

Yu Wang, Doctor of Philosophy, 2024

Dissertation directed by: Professor Qingbin Cui, Civil and Environmental Engineering

Risk management is critical to project success, especially in public-private partnerships (P3s) featuring long-term relationships, uncertainty, and complexity. Poorly handled risk management, especially regarding risk transfer, can lead to incentive distortion, disputes, or even project failure. The contract, serving as the formal and enforceable legal agreement binding on the public and private partners, plays a vital role in transferring risks associated with P3s. Risk identification is an important step in contract risk analysis, since overlooking specific risk clauses may cause detrimental consequences, such as revenue loss, unexpected financial liabilities, and legal disputes for contracting parties.

Previous research has extensively examined the identification and allocation of project risks between contracting parties, predominantly employing questionnaire surveys, interviews, or content analysis methods. These studies depict common practices of risk identification and

allocation, with some addressing risks stipulated in contracts. Nonetheless, there are notable limitations. Firstly, the findings derived from these traditional approaches often lack replicability. Secondly, given the typical lengthy nature of P3 contracts, conventional methods for analyzing risk-related contract content are labor-intensive and time-consuming. Thirdly, most of the studies do not offer a means to retrieve specific provisions for nuanced scrutiny. Addressing these limitations necessitates the adoption of innovative approaches to gain more granular and replicable results in risk-related contract analysis. The ideal solution should allow for the effortless and consistent extraction of specific contractual provisions related to project risks, providing a microscopic lens to risk allocation practices.

With the recent advancements in natural language processing (NLP), especially transformer-based pre-trained language models (PLMs) and cutting-edge large language models (LLMs), there has been a significant breakthrough in the efficiency of processing and extracting information from textual data. Motivated by both the pivotal yet complicated nature of contract documents and the increasingly mature NLP techniques that create new opportunities for text analysis, this research aims to utilize NLP to automate the identification of risk-related aspects in contract documents. Firstly, a risk-related framework of P3 contracts is developed based on a literature review and a contract review. Based on that, a series of NLP-based tools are developed for the automated identification of risks-related contract language, including 1) a rule-based model for contingency liability identification with a weighted F1-score of 88.9%, 2) a fine-tuned PLM (particularly the BERT family) for risk type and allocation identification with a weighted F1-score of 80.6% and 80.5%, and 3) a prompt design with an LLM (particularly GPT-3.5) for risk type and allocation identification with a weighted F1-score of 64.1% and 72.1%. Next, the effectiveness of these different approaches is compared. Finally, we apply the tools to real

contract documents to offer risk profiles of P3 contracts. The goal is to foster a more efficient, precise, and in-depth understanding of contract risks by leveraging the capabilities of NLP technologies.

TOWARDS AUTOMATED CONTRACT ANALYSIS: APPLYING LANGUAGE MODELS TO RISK IDENTIFICATION IN THE CONTEXT OF PUBLIC-PRIVATE PARTNERSHIPS

by

Yu Wang

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
2024

Advisory Committee:
Dr. Qingbin Cui, Chair
Dr. Gregory B. Baecher
Dr. Young Hoon Kwak
Dr. Peter A. Sandborn, Dean's Representative
Dr. Mirosław J. Skibniewski

Dedication

This dissertation is dedicated to my mentors, friends, and family.

Acknowledgements

I would like to thank the following people, without whom I would not have been able to complete this research or make it through my PhD journey.

First, I am profoundly grateful to my dissertation committee. In particular, I would like to express my deepest gratitude to Prof. Qingbin Cui, my advisor, whose expertise in both academia and practice has greatly enriched my learning experience. His sharp insights and relentless enthusiasm have encouraged me to expand my research perspectives, infusing my PhD journey with intellectual inspiration. I would like to thank Prof. Young Hoon Kwak, who has been a very supportive advisor and a great coauthor. The collaborations with him have been enjoyable and fruitful. I also would like to express my appreciation to Prof. Gregory B. Baecher, Prof. Peter A. Sandborn, and Prof. Mirosław J. Skibniewski, whose insightful feedback and support were invaluable not only in guiding my research through this process but also in positioning me for continual improvement in the future.

I am deeply indebted to Prof. Yongqiang Chen, my mentor and adviser back at Tianjin University, who has left an indelible mark on my values and profoundly influenced my professional development. He not only encouraged me to pursue a second PhD abroad but has also been continuously supporting me in every aspect.

My thanks also go to my fellow PhD students and friends, whose camaraderie and support went beyond our shared academic challenges. Their presence was essential in helping me navigate both my studies and personal highs and lows during this period. In particular, I owe a great deal of gratitude to Chenglong Xu. The discussions I had with him about large language models have benefited me a lot. I am also grateful to

Lingyao Li, whose enthusiasm for research and collaborative spirit have greatly impacted my research. Special thanks go to my friends at the University of Maryland and Purdue University, especially those from the badminton court. The laughter we shared brought a lot of joy and color to my PhD life.

I cannot express enough thanks to my family. My parents have unconditionally supported me in innumerable ways throughout this journey. Their love and encouragement have been my sustenance and solace during tough times. Last but not least, I want to thank my dearest partner in life, Wenqian, with whom I have shared over a decade, including the unforgettable years both in College Park and West Lafayatte. We have been each other's rock, comfort, and source of encouragement. I could not imagine how this journey would have been without your love, trust, company, and endless jokes.

Finally, I wish to thank everyone who has played a part, directly or indirectly, in the completion of this dissertation.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

Abbreviation	Definition
AP	Availability Payment
API	Application Programming Interface
BERT	Bidirectional Encoder Representations from Transformers
CNN	Convolutional Neural Network
DB	Design-Build
DBF	Design-Build-Finance
DBFOM	Design-Build-Finance-Operate-Maintain
DBOM	Design-Build-Operate-Maintain
FIDIC	Federation Internationale des Ingenieurs Conseils
FN	False Negative
FP	False Positive
GPT	Generative Pre-Training
LLM	Large Language Model
LSTM	Long Short-Term Memory
MLM	Masked Language Modeling
NEPA	National Environmental Policy Act
NSP	Next Sentence Prediction
P3	Public-Private Partnership
PLM	Pre-Trained Language Model
RNN	Recurrent Neural Network
TIFIA	Transportation Infrastructure Finance and Innovation Act
TN	True Negative
TP	True Positive

Chapter 1. Introduction

1.1. Research Needs and Purpose

1.1.1. Background and Research Needs

Contracts serve as a foundational interorganizational governance mechanism (Z. Cao & Lumineau, 2015; Williamson, 1985). They are crafted collaboratively to guide interactions and improve outcomes, particularly in large construction projects involving multiple parties with potentially conflicting interests. Risk management is critical to project success. Poorly handled risk management, especially risk transfer, may lead to incentive distortion, dispute, or even project failure. Contracts as the formal and enforceable legal tool binding on the involved parties, plays a vital role in transferring project risks (Lam et al., 2007).

However, contract risks sometimes cannot be well identified and understood. A failure to adequately understand and address contractual risks can lead to detrimental consequences, such as incentive distortion, non-cooperation among involved parties, and disputes (Ni, 2012; S. Zhang et al., 2016). The global construction market experienced an average dispute resolution cost of $52.6 million and a resolution timeline of 15.4 months in 2021 (Arcadis, 2022). Issues such as “errors and omissions in the contract document” or failing to understand and/or comply with its contractual obligations” are among the primary causes. These challenges underscore the importance of effective risk identification and allocation, which, if improperly managed, can adversely impact cooperation and project success.

In the specific context of public-private partnerships (P3s), the stakes are heightened. P3s are contractual relationships between public and private partners, designed to facilitate greater private sector involvement in infrastructure projects. Due to the long-term relationships, broader scope of work, significant risk transfer, additional considerations associated with financing, and the lack of standardized contract models (Heravi & Hajihosseini, 2012; Smith et al., 2019), contract risk management becomes even more crucial in this context. Risk identification is the first and a critical step in contract risks analysis, since overlooking specific risk clauses could have detrimental consequences, such as revenue loss or unexpected financial liabilities.

From the perspective of practical needs in the P3 context, the identification of contractual risks is crucial not only for providing a reference benchmark for dealing with risk events once they have occurred, but also for optimizing decision-making during project procurement in both the public and private sectors. When public entities initiate large-scale infrastructure projects, they engage in a competitive bidding process aimed at attracting capable private partners. For these potential private bidders, a draft version of the contract has already been prepared by the public sector, leaving risk identification and assessment of all potential risks as a primary step before entering into the contract (M.-T. Wang & Chou, 2003). A quick and comprehensive initial risk identification process could allow them to make well-informed decisions regarding their participation and to strategize effectively on risk mitigation. For the public sector, the reasonableness of risk allocation is not just procedural formalities but is essential in creating a competitive environment that attracts high-quality bids, thereby enhancing project viability and success. Overall, a

balanced risk allocation reassures the private sector of the public sector's commitment to fairness, transparency, and mutual benefit in the partnership. This, in turn, lays a solid foundation for successful, sustainable infrastructure development that benefits all stakeholders involved.

However, the complexity of P3 contracts, often detailed in documents spanning several hundred pages, presents a challenge in risk assessment. Traditional methods of contract analysis are labor-intensive and fraught with subjective interpretations and replication issues. Therefore, a tool that facilitates quick and accurate contract risk analysis would be invaluable. Thus, developing and utilizing such advanced tools would contribute to a more transparent, equitable, and efficient negotiation and project execution process for both public and private sectors.

1.1.2. Purpose of the Study

Since contracts fundamentally constitute textual data, natural language processing (NLP) presents an innovative avenue for the achievement of rapid, low-cost, and stable text analysis. Previous studies have already demonstrated NLP's capability to efficiently process project-related texts, including those of contracts. Yet, these studies primarily rely on rule-based or statistical NLP methods. With the recent advancements in NLP, especially pre-trained language models (PLMs) like BERT (Bidirectional Encoder Representations from Transformers) and state-of-the-art large language models (LLMs) like GPT (Generative Pre-Training) series models, the efficiency of processing and extracting information from textual data has significantly improved. The rapid advancement of LLMs marks a significant leap forward in NLP. As this field is still relatively new, the deployment of these models in vertical

domains has not been fully realized. Particularly, the application of these technologies to contract analysis remains underdeveloped. This opens up extensive opportunities for further customizing these models to meet the unique demands and challenges of specialized industries.

Motivated by both the important yet complicated nature of contract documents and the increasingly mature NLP techniques that create new opportunities for text analysis, this research aims to automate contract risk identification by utilizing using a series of language models. This involves the development of a series of tools using approaches under different NLP paradigms—including rule-based approaches, fine-tuning of PLMs, and prompt engineering with LLMs—with a variety of models and strategies.

Given the importance of comprehensive contract risk analysis for both public and private sectors, a tool that enables quick assessment of contract risk profiles would be highly beneficial. For the public sector, such a tool could help streamline the procurement process by ensuring that the contracts are structured in a way that is both fair and attractive to the private sector. Moreover, this would empower them to identify potential pitfalls and address them proactively, thereby minimizing delays and disputes during the execution phase of the project. For private bidders, this tool can help them conduct their own risk assessments with greater accuracy and efficiency, thereby facilitating their strategic planning. Overall, such a tool for contract risk analysis would contribute to a more transparent, equitable, and efficient bidding, negotiation, and project execution. Additionally, this research contributes to the NLP field by evaluating and comparing the performance of different language

models in the specific domain of P3 contract, providing evidence of the effectiveness of these approaches. Finally, this research deploys the models to reviewing risk profiles across different P3 contracts, showcasing the application the tools, and facilitating a holistic and fine-grained understanding of current practice in the transportation sector in the U.S.

1.2. Current Knowledge

1.2.1. P3s and P3 Risk Analysis

1.2.1.1. P3 Overview

Definitions of P3s

P3s are a collaborative approach in addressing infrastructure needs. While there is no unanimous definition, P3s generally refer to any contractual arrangement between a public partner and a private partner for delivering infrastructure or public services, characterized by shared responsibilities and risks, a combination of all project phases into one contract, substantial risk transfer to the private partner, and extended contract terms (Iossa et al., 2007). A P3 typically allows more private sector participation than is traditional in public infrastructure projects .

Rising public infrastructure demands like transportation have increased pressures on public budgets, prompting the need for P3s (Osei-Kyei et al., 2023). P3s are favored for their potential to reduce pressure on public budgets, enhance efficient management, and foster innovation through private sector involvement. However, they also face various challenges and controversies, such as structural pitfalls,

multiple stakeholder involvement (Ahmed & Garvin, 2022), protracted negotiations, and high financing costs (Osei-Kyei et al., 2023).

P3s in Transportation

The transportation sector exemplifies the application of P3s with 60%-70% of global investment from 2000 to 2016 (Yescombe & Farquharson, 2018). In the U.S., transportation P3s are agreements between state or local governments and private entities for the delivery of transportation projects like highways, bridges, tunnels, and transit, typically with the public partner retaining ownership while the private partner assumes more risks or have more decision-making authority (Congressional Research Service, 2021; U.S. Department of Transportation, 2004).

In the U.S., transportation P3s take various forms and sizes. These partnerships are characterized by the extent of private sector participation and can involve diverse combinations of responsibilities between the public and private entities, particularly often including financing along with long-term operations and maintenance (Congressional Research Service, 2021; Papajohn et al., 2011). Some commonly seen types of P3s include:

1) **Design-Build (DB):** Through a fixed-price contract, the private partner bears most or all the risk of design and construction, while the public partner is responsible for financing, operations, and maintenance.
2) **Design-Build-Finance (DBF):** This resembles a DB project, except that in DBF the private partner also needs to provide the necessary up-front capital and is typically reimbursed by a government agency over time through a series of payments.

3) **Design-Build-Operate-Maintain (DBOM):** The private partner takes on design, construction, operation, and maintenance, with the public partner responsible for financing and toll collection.
4) **Design-Build-Finance-Operate-Maintain (DBFOM):** Also known as a toll concession, this arrangement encompasses design, construction, financing, operation, maintenance, and toll collection by the private partner.
5) **Long-term Lease Agreements:** The private partner operates and maintains an existing facility for a specified time, collecting tolls or user fees, and paying a concession fee to the public partner.

P3 legislation in the U.S. for transportation P3s

The implementation of P3s in the U.S. for transportation projects requires legal and policy frameworks established through state enabling statutes. These laws authorize state and local government agencies to establish P3s and include provisions on the types of P3 arrangements allowed, project selection and approval processes, proposal reviews, public involvement, and information release requirements.

Currently, P3s have a short history in the US and there is no national P3 legislative framework. The federal government influence on P3s through funding, programs and regulations, but each state has its own set of laws and regulations governing P3s. As of August 2018, 36 states, the District of Columbia, and Puerto Rico had general P3 enabling legislation, resulting in 37 transportation P3s with long-term financing between 1993 and 2018 (Congressional Research Service, 2021).

1.2.1.2. P3 risks Analysis: Traditional Methods and Research Gaps

While P3s offer a promising avenue for addressing transportation infrastructure needs, they also call for careful management of their complexities. Key among these challenges is the effective management of risks, a topic of significant interest in P3 research and practice due to its pivotal role in the success of P3s (Osei-Kyei et al., 2023). These large-scale infrastructure projects often entail various risks related to economic, social, political, and environmental factors (Charoenngam & Yeh, 1999). A key advantage of P3s is the ability to transfer these risks, enhancing project outcomes such as cost and schedule performance, and leveraging private sector expertise. However, inadequate risk management in P3s, which is not uncommon, can lead to significant adverse consequences (Heravi & Hajihosseini, 2012; Ni, 2012). As P3s continue to be an important tool for infrastructure development, understanding their legal frameworks, risk management, and the balance of responsibilities and benefits between the public and private partners are of vital importance for their successful implementation.

Previous research has extensively identified the key risk categories and the allocation of these risks between contracting parties. Commonly used methods for project risk analysis include systematic literature review (Kumari & Kumar Sharma, 2017; Kwak et al., 2009; Le et al., 2019), surveys (Bing et al., 2005; Carbonara et al., 2015; Ke et al., 2010; J. Li & Zou, 2011; Thomas et al., 2003), case studies (Heravi & Hajihosseini, 2012; Hodge, 2004b), interviews or expert discussions (J. W. Brown et al., 2009; D. Chung et al., 2010), and content analysis (Nguyen et al., 2018). The insights from these studies carry significant practical importance, depicting how

contracting parties tend to identify and allocate risks in common practices. A P3 contract serves as a mechanism for allocating risks to the party most capable of handling and minimizing them (Smith et al., 2019). Nonetheless, there has been a lack of extensive utilization of contracts as a primary source for investigation despite the critical role of contracts in risk management within P3s (Nguyen et al., 2018). This indicates a potential area for further research, emphasizing the need to more closely examine the contractual aspects of risk management. Besides, there are notable limitations inherent to these conventional approaches.

Firstly, the findings derived from these traditional approaches lack replicability. The reliance on subjective data sources, influenced by respondents' perceptions, affects the consistency of the data. Even with objective data, the analysis methods heavily depend on the researchers' subjective judgments (Jahedi & Méndez, 2014). Consequently, attempts to replicate these studies with different respondents or by different researchers can yield inconsistent results.

Secondly, as P3 contracts tend to be lengthy—often spanning several hundred pages—contract analysis is never an easy task for scholars and contract managers. Traditional methods for risk-related contract content analysis can be labor-intensive and time-consuming.

Thirdly, most of the studies have primarily assessed risks at the overall risk level with limited scenarios, without offering a means to retrieve specifics for more nuanced scrutiny. In other words, the results often generalize risk allocation, simplifying it to statements like "Risk 1 is assumed by Party A or Party B or is shared" while leaving specific contractual language largely unexamined.

Addressing these limitations necessitates the development and adoption of innovative approaches to gain more nuanced and replicable insights into risk allocation in contracts. Such novel methods should enable the retrieval and scrutiny of specific contractual sentences and provisions, offering a granular view of risk allocation practices. By doing so, we can advance our understanding of this critical aspect of contract management and enhance its practical applicability for a wide range of contracts.

1.2.2. NLP and Its Application in the Construction Industry

1.2.2.1. Overview of NLP

NLP is an interdisciplinary field that combines elements of computer science, linguistics, and artificial intelligence to enable computers to understand, interpret, and respond to human language in a valuable way. The roots of NLP can be traced back to the 1950s, with the emergence of computational linguistics. One of the earliest milestones was Alan Turing's Turing Test (Turing, 1950), which was designed to evaluate a machine's ability to exhibit intelligent behavior indistinguishable from that of a human. The evolution of NLP can be broadly categorized into three stages:

1) Symbolic NLP (1950s–1980s)

In its early stages, NLP was primarily driven by a symbolic approach that relies on rule-based systems, requiring linguists and computer scientists to manually construct language rules and logic. These methods initiated the early attempts to use computers for extracting information from unstructured text data and automating tasks that were previously labor-intensive. However, this approach faced limitations due to its heavy

reliance on manual rule-setting and a lack of adaptability to the nuances of human language.

2) Statistical NLP (1980s–2010s)

Transitioning into the late 1980s, NLP entered the statistical stage, which gained prominence until the early 2010s. This stage represented a paradigm shift to statistical methods to learn language patterns from large corpora of text. (Manning & Schutze, 1999). The introduction of machine learning algorithms advanced NLP to enable more sophisticated text analysis and language understanding. This era witnessed the development of both unsupervised methods, like topic modeling and clustering, and supervised learning techniques utilizing manually annotated data with algorithms such as Naïve Bayesian, Linear Regression, Logistic Regression, Random Forests, and Support Vector Machines. Representative methods in this era include Bag of Words, Term Frequency-Inverse Document Frequency, n-grams, Latent Semantic Analysis, Hidden Markov Models, and more. The methods significantly advanced the field of text understanding and representation. They marked a shift towards machine learning models, though humans still played a crucial role in identifying data features. Although these methods offered less context-aware and semantically rich word representations compared to modern neural word embeddings later on, they were instrumental in paving the way for these more advanced techniques.

3) Neural NLP (2010s–Present)

With advances in computational power and deep learning techniques, NLP embarked on a new era characterized by neural networks. This current stage began in the early 2010s and continues to evolve rapidly. Neural NLP leverages deep learning to model

and understand language. It represents a shift towards models that can learn representations and semantic relationships from data.

Early deep learning approaches applied in NLP included convolutional neural networks (CNNs) and recurrent neural networks (RNNs). Word2Vec (Mikolov et al., 2013a, 2013b) represented the rise of word embeddings. Using a shallow neural network, Word2Vec maps words into a multi-dimensional vector space where semantically similar words are closely positioned. Other similar methods like GloVe (Pennington et al., 2014), FastText (Joulin et al., 2016) were proposed subsequently.

However, CNNs and RNNs had limitations, particularly in handling complex tasks and long-term dependencies. CNN uses a window-based approach to extract high-level features, and it cannot deal with long-term dependencies (Lauriola et al., 2022; Wolf et al., 2020). RNN is more suitable for processing sequential information that is a typical characteristic of human language, but its efficiency is often decreased by the exploding or vanishing gradients. Long Short-Term Memory (LSTM) networks, as a variety of RNNs, were developed to mitigate these issues by allowing the addition or removal of information through cell states. Based on LSTMs, ELMo (Embeddings from Language Models) (Peters et al., 2018) set the precedent for pre-training language models. Nonetheless, as RNNs and LSTMs are trained by passing words sequentially, they faced loss of context and challenges in parallel implementation.

A significant breakthrough came with the introduction of Transformer (Vaswani et al., 2017), utilizing a self-attention mechanism that enabled computation in parallel and extraction of global dependencies. These transformer-based models can generate contextualized word representations, a significant improvement over context-free

models like Word2Vec and GloVe that only generate a single representation for a word despite that the meaning of a word often varies depending on the context. The Transformer architecture has revolutionized NLP, leading to the development of models like Google's BERT (Devlin et al., 2019) and OpenAI's GPT (Radford et al., 2018), which are pre-trained on large datasets and can be fine-tuned for specific tasks. Since then, language models have continued to evolve into LLMs with billions of parameters and become unprecedentedly powerful, and various LLMs have been launched successively. In 2020, OpenAI released the GPT-3 model (T. Brown et al., 2020) that showed amazing capability of generating human-like text. The release of ChatGPT in November 2022 marked a significant milestone, demonstrating the immense capabilities of LLMs in a variety of applications. The current stage, with its focus on LLMs, has brought unprecedented capabilities in understanding and generating human language, setting a new standard in the field of NLP.

1.2.2.2. NLP Application in the Construction Industry

The rapid evolution of NLP has attracted considerable attention and led to its wide use in various fields. Consequently, the construction industry is also experiencing a surge in NLP applications designed to process textual data (Bilal et al., 2016; Ding et al., 2022; J. Liu et al., 2022, 2022; Pan & Zhang, 2021; Wu et al., 2022; Yan et al., 2020, 2022).

In terms of tasks and purposes, most common applications of NLP in the construction domain include information extraction and retrieval (Akanbi & Zhang, 2021; Kim et al., 2015; Moon, Chi, et al., 2022; N. Wang et al., 2022), compliance checking (J. Zhang & El-Gohary, 2016; R. Zhang & El-Gohary, 2021; Zhou & El-Gohary, 2017),

model comparison (G. Lee et al., 2011; Shi et al., 2018), automated reasoning (J. Zhang & El-Gohary, 2015), and opinion mining (Lv & El-Gohary, 2016; Park et al., 2022).

In terms of methods, various approaches and their combinations, as discussed in Section 1.2.2.1, have been applied to the field. For example, Zhang and El-Gohary (2016) extracted requirements from construction regulatory documents using rule-based NLP. Tixier et al. (2016) proposed an NLP system based on hand-coded rules and keywords to extract precursors and outcomes from construction injury reports. Zhong et al. (2020) developed a pipeline that combines multiple text mining techniques including Latent Dirichlet Allocation and CNN to automate the analysis of transportation project hazard records. Moon, Lee, et al. (2022) developed an automated system for reviewing construction specifications using Word2Vec and bi-directional long short-term memory (Bi-LSTM). Feng and Chen (2021) adopted the BiLSTM-CRF model to extract profile information from accident news reports, such as date, location, accident type and causes, and fatality and injury situations. F. Zhang (2022) adopted Word2Vec and machine learning methods for automatic classification of construction accident causes from accident reports. Gao et al. (2022) compared multiple approaches such as CNN, BERT and traditional machine learning methods, to extract risk narratives from news articles.

In terms of data sources, most commonly seen textual data in the construction domain include construction specifications and requirements (Moon, Lee, et al., 2022; J. Zhang & El-Gohary, 2015, 2016), legal documents (Moon, Chi, et al., 2022), safety-related reports and records (Baker et al., 2020; Feng & Chen, 2021; F. Zhang, 2022;

Zhong et al., 2020) , news articles (Gao et al., 2022), social media (Lv & El-Gohary, 2016; Park et al., 2022), and others.

Specifically related to this research, previous studies in the sub-domain of information extraction from legal documents have also applied different NLP and machine learning approaches to automate the process. Al Qady and Kandil (2010) applied a shallow parser to extract semantic knowledge from construction contracts. Li et al. (2015) utilized NLP and pattern matching methods to extract the early software requirements of new projects from project reports and user manuals. Lee et al. (2019) proposed an automatic model of contract-risk extraction that can detect predefined clauses of the contract, using a subject-object-verb tuple matching method. Lee et al. (2020) presented a rule-based model to identify missing contractor-friendly clauses in the owner's modified contract conditions. Hassan and Le (2020) utilized rule-based NLP and machine learning techniques to classify construction contract text into requirement and nonrequirement text. Padhy et al. (2021) developed a rule-based model for exculpatory clause identification from construction contracts. Choi et al. (2021) utilized rule-based NLP for contractor's risk identification in engineering, procurement, and construction contracts. Khalef and El-adaway (2021) adopted NLP techniques and multiple machine learning algorithms to identify contractual changes in airport projects. Candaş and Tokdemir (2022) classified contract clause topics in standard contracts issued by Federation Internationale des Ingenieurs Conseils (FIDIC) using multiple machine learning algorithms. Yang et al. (2022) also compared different machine learning algorithms such as support vector machine and logistic regression to classify contract language with three functions—control,

coordination, and adaptation—from FIDIC contracts. Fu et al. (2023) fine-tuned five BERT series models for contract function identification using a dataset of FIDIC standard contract forms and real contracts. Qi et al. (2024) also trained a deep learning-based model to analyze the structure of construction contracts from a multifunctional perspective. L. Zhang et al. (2023) conducted both questionnaire surveys and machine learning based contract analysis to compare subjective and objective measures of contract complexity. Pham and Han (2023) developed a multitask classification model using BERT to identify five risk factors, along with risk allocation and risk response from contract documents.

With regards to risk identification, current research primarily focuses on recognizing either a single type or only a limited range of risks (Pham & Han, 2023), while lacking comprehensive and detailed risk analysis. In terms of methodologies, the field predominantly relies on traditional paradigms—symbolic and statistical approaches—although recent studies have begun to incorporate transformer-based models. However, considering the recent emergence of LLMs, their application in analyzing legal documents such as construction contracts, especially in the P3 context, remains relatively insufficient.

A comparison between state-of-art NLP and its applications in the construction industry (S. Chung et al., 2023) highlighted the disruptive impact of recent advancements in NLP like ChatGPT across various fields, while also noted the challenges in applying these cutting-edge technologies to the construction industry such as “the lack of a validation dataset”. Therefore, this research aims to bridge this gap by focusing on the utilization of sophisticated NLP methods for processing

complex construction-related texts, especially in the context of transportation P3 projects. With the aid of the automated approaches, risk identification from a contract or comparison across multiple contracts becomes a quick, repeatable, reliable, and consistent process.

1.3. Research Design

This dissertation work employs three approaches by utilizing language models in contract risk identification. For each approach, one or more multi-label classification tasks are implemented to extract specific contractual languages related to risks.

1) **Rule-based approach** (Chapter 2)
2) **Fine-tuning PLMs to obtain a domain specific model** (Chapter 3)
    - Model selection:
        - BERT
        - RoBERTa
        - LegalBERT
        - XLM-RoBERTa
3) **Prompt engineering with LLMs** (Chapter 4)
    - Model selection: GPT-3.5 Turbo
    - Different prompting strategies:
        - Zero-shot prompting
        - Few-shot prompting
        - Few-shot with chain-of-thought prompting

1.4. Dissertation Outline

The rest of the dissertation is organized as follows:

Chapter 2 proposes the concept of contingency liability that determines if a specific circumstance or event occurs which the parties are obligated to take actions accordingly or they are entitled to a relief, compensation, or the right to terminate the contract. Five types of contingency liability contract languages are defined, including remedy entitlement, remedy obligation, liability waiver, mitigation, and termination. Using a rule-based approach, a domain-specific lexicon developed from U.S. transportation P3 contracts and a set of matching rules are developed to identify target sentences.

Chapter 3 develops two classification schemes tailored to risks stipulated in contracts: 1) Risk type classification: This categorizes the various types of risks commonly found in P3 contracts; 2) Risk allocation elements classification: This focuses on classifying various elements related to the allocation of risks, including rights, obligations, liabilities, and prohibitions of the contracting parties. Together, these two classification schemes provide a comprehensive tool for analyzing and understanding how risks are defined and allocated in P3 contracts. To implement these classifications, this chapter adopts the approach of fine-tuning PLMs.

Chapter 4 shifts focus from the methodology employed in Chapter 3 to an alternative approach: employing prompt engineering with large general-purpose language models for the same tasks, i.e., automated identification of risk type and allocation in contracts. In this chapter, various prompting techniques, such as zero-shot, few-shot, and chain-of-thought prompting, are explored and compared. The aim is to identify

effective prompt formats to elicit LLMs for classifying risk-related aspects in P3 contracts. Finally, the findings are discussed in the context of existing literature.

Chapter 5 compares the performance of the approaches and models employed in this Dissertation. Following that, this chapter applies the developed models on a corpus of P3 contracts to extract and systematically analyze their risk profiles. This application demonstrates the real-world viability of the models and provides insights into the current practices of how risks are defined and allocated in P3 contracts.

Chapter 6 summarizes the contributions, limitations, and future research directions.

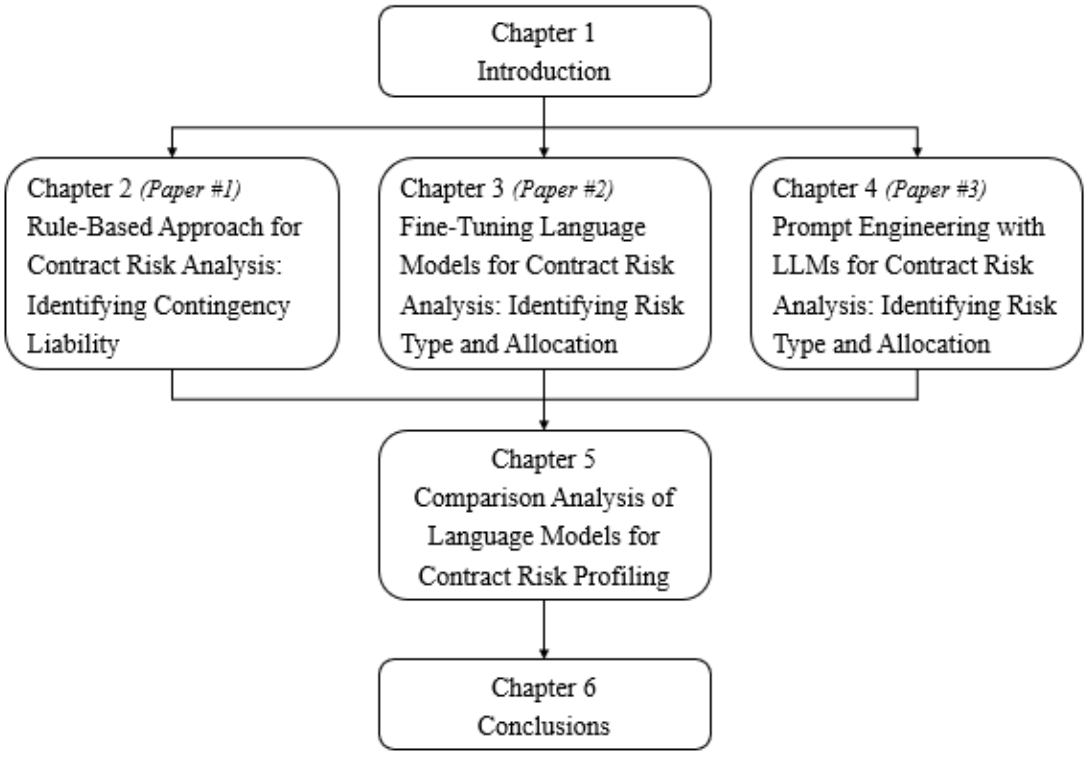


Figure 1-1 Dissertation structure

Chapter 2. A Rule-Based Model for Contract Analysis: Identifying Contingency Liability

2.1. Abstract

The COVID-19 pandemic has increased contractual concerns under contingencies for P3 projects. Conventional manual contract extraction is time-consuming and error-prone. Devising a method for automatic contract extraction can support contract management in this aspect. This chapter proposes a rule-based NLP approach to extracting contingency liabilities allocated between the public partner and the private partner in the contract. The model consists of a domain-specific lexicon developed based on 21 U.S. transportation P3 concession agreements and a set of matching rules to identify target sentences which fall into five classes, namely remedy entitlement, remedy obligation, liability waiver, mitigation, and termination. This automatic process can reduce the time and cost of the contract review process and help identify issues that the contracting parties should consider going forward in drafting new contracts, negotiation, or in amending existing contracts to avoid potential disputes, in response to consequences of contingencies, including the COVID-19 pandemic.

2.2. Introduction

The ongoing COVID-19 pandemic has profoundly impacted many construction and infrastructure projects, influencing all aspects from project planning, design, construction, to operation and maintenance (Baxter & Casady, 2020). Although many projects were not forced to shut down, the mobilization of workers and equipment was hindered in compliance with government restrictions and mandates. For instance,

the 14-day quarantine and travel restriction can delay on-site operations. Additionally, the global production halts across various industries caused a shortage of materials and equipment.

These disruptions led to an increase in claims and disputes. Contractors may seek cost relief based on contractual pricing mechanisms and escalation clauses, with potential disputes over insurance coverage for construction risks (Alsharef et al., 2021). The Purple Line project serves as a typical case of the broader pandemic-induced challenges in the construction sector. This project experienced significant cost overruns and delays, leading the Maryland Department of Transportation to incur substantial costs to resolve disputes. The original contractor withdrew from the project citing a contractual provision that allowed termination after prolonged delays. Subsequently, the Maryland Department of Transportation engaged a new contractor under a modified contract, which removed the right to unilateral termination due to extended delays from the agreement, to avoid a repeat of similar issues in the future.

In this situation, questions about contingency liabilities between contracting parties, specifically referring to the public partner and the private partner in P3 context, have become especially pertinent (Casady & Baxter, 2020). Resolving these issues requires a thorough contract review to determine relevant provisions and assess the extent of non-monetary and/or monetary compensation available to affected parties. It is vital to understand whether the contract relieves the parties of their obligations or allows for termination of the contract. In response to these complexities, practitioners and scholars advocate for a reevaluation of force majeure and excusable delay clauses in contracts to more effectively safeguard stakeholders' financial interests (Abubakar et

al., 2022). To mitigate delays and minimize economic losses, it is essential that parties identify and leverage contract provisions that address potential consequences and provide relief for such unforeseen risks.

Prompted by the impact of COVID-19 pandemic, this study aims to extract contingency liability information from P3 contracts. Traditionally, such contract information extraction has been performed through manual content analysis, a method that is time-consuming and labor-intensive. This conventional approach also struggles to accommodate the need for efficient, real-time updates. Another problem with manual analysis is the inconsistencies among individuals if the task is done by multiple coders. To address these issues, text mining approaches have been used to automate the reviewing process, making text analysis easy and consistent (Pan & Zhang, 2021).

This research reviewed 21 U.S. transportation P3 concession agreements to identify contingency liability issues and shed light on contract design. To this end, we developed a rule-based model for auto-contract extraction. With the lexicon and rules specific to the allocation of rights and obligations under contingencies, the model identifies the linguistic structures that people have laboriously coded by hand. This automatic process can help identify issues that the contracting parties should consider going forward in drafting new contracts or in amending existing contracts to avoid potential disputes, in response to contingency-related consequences.

2.3. Contract Provisions in Relation to Contingency

To deal with uncertainties during the contract term, construction contracts usually have provisions that define specific circumstances or events in which the parties are

obligated to take actions accordingly or they are entitled to a relief, compensation, or even a right to terminate the contract.

In construction contracts, the force majeure provision is almost a requisite that incorporates risk events that may affect the parties' ability to perform their obligations under the contract and even lead to an early termination. Force majeure events are typically defined as events beyond the parties' reasonable control, including wars, terrorism, riots, strikes, natural disasters, nuclear or other explosions, etc. It is worth noting that force majeure definitions and related provisions are not standardized in the contract world. Different contracts may give different definitions and interpretations of a contingency. Besides force majeure, other events that entail specific contract provisions related to risk allocation under contingencies are often defined as relief events, compensation events, delay events, uninsurable events, etc. According to the Federal Highway Administration's P3 concession contract guide, compensation events are those for which the public partner assumes the risk, whereas in delay events the private partner are better placed to assume the risk (FHWA, 2015). Some contracts use supervening events as an umbrella term to refer to various unforeseen circumstances (Hovy, 2015). For example, in the Central 70 project, Supervening Event means "any Relief Event, Compensation Event; or Appendix B Parcel Unexpected Hazardous Substances Event to the extent that it does not constitute a Compensation Event".

Table 2-1 is an overview of 23 U.S. transportation P3 concession agreements from 2006 to 2021. Regarding the P3 concession type, all the projects are in DBFOM structure except that the Indiana Toll Road project is a long-term lease. All the

contracts have force majeure provisions; 16 define the compensation event; 17 define the relief event; 5 define the delay event; 3 define the supervening event.

Table 2-1 Presence of major contingencies in U.S. transportation P3 project contracts

Project	Jurisdiction	Commercial close	Payment method	Force Majeure	Compensation event	Relief event	Delay event	Supervening event
Indiana Toll Road	IN	2006	Lease	√	√		√	
I-495 Capital Beltway Express	VA	2007	Tolled	√	√		√	
SH 130: Segments 5 and 6	TX	2007	Tolled	√	√	√		
North Tarrant Express Segments 1 and 2A	TX	2009	Tolled	√	√	√		
LBJ Express (IH 635 Managed Lanes)	TX	2009	Tolled	√	√	√		
Eagle Project	CO	2010	AP	√		√		
Midtown Tunnel	VA	2011	Tolled	√	√		√	
Presidio Parkway	CA	2011	AP	√		√		
I-95 Express Lanes	VA	2011	Tolled	√	√		√	
Ohio River Bridges East End Crossing	IN	2012	AP	√		√		
North Tarrant Express Segments 3A and 3B	TX	2013	Tolled	√	√	√		
US 36 Express Lanes (Phase 2)	CO	2013	Tolled	√	√	√		
I-4 Ultimate	FL	2014	AP	√		√		
I-69 Section 5	IN	2014	AP	√		√		
I-77 Express Lanes	NC	2014	Tolled	√	√	√		
Pennsylvania Rapid Bridge Replacement	PA	2015	AP	√	√	√		√

Southern Ohio Veterans Memorial Highway	OH	2015	AP	√	√	√		√
SH 288 Toll Lanes	TX	2016	Tolled	√	√	√		
Purple Line	MD	2016	AP	√		√		
Transform 66	VA	2016	Tolled	√	√			
I-395 Express Lanes	VA	2017	Tolled	√	√		√	
Central 70	CO	2018	AP	√	√	√		√
I-495 & I-270 Program (Phase 1)	MD	2021	Tolled	√		√		

(Note: “AP” refers to “availability payment.”)

A contingency may trigger a variety of relevant provisions, including but not limited to the abovementioned circumstances. These provisions reflect the contractual mechanism for risk allocation under such unusual events (Dewulf & Garvin, 2020), which can be classified into two types. The first type is definitions. These provisions clarify whether a specific incident is explicitly recognized in the relevant clauses, whether the incident constitutes a force majeure event, a relief event, a compensation event, and/or a delay event. For example, the contract may define "pandemic" as a force majeure, which applies to the COVID-10 disruption. The definitions are usually defined in the contract attachments (e.g., Exhibit) and are easy to identify. The second type is consequences. When an incident is recognized based on the definition, do the relevant clauses allow for non-monetary relief and/or monetary compensation, allow a party to be excused from performance of its obligations, or allow a party to terminate the contract? This type of provision addresses contingency liability and is core to risk allocation, as improper contingency liability assignment may create misaligned incentives. Moreover, the consequences provisions are more dispersedly distributed in the contract, which increases the difficulty for manual identification. In this research, the type of provisions is the target of our auto-contract extraction model.

2.4. Methodology

2.4.1. Data

The model was developed and tested using the concession agreements of 21 transportation P3 projects out of the 23 projects in Table 2-1. Note that North Tarrant Express Segments 3A & 3B was awarded, through negotiation, to the same developer as that of North Tarrant Express Segments 1 & 2A, and the two facility agreements were under the umbrella of the same predevelopment agreement. Similarly, I-395 Express Lanes was undertaken by the same

developer as that of I-95 Express Lanes, and the former's contract was amended based on the latter. Considering that North Tarrant Express Segment 3A & 3B and I-395 Express Lanes added hardly any new information to their counterparts regarding the contingency issue, they were excluded from the final data set. Among the 21 contracts, 19 were used for model development and two were used to evaluate the performance of the model.

2.4.2. Rule-Based Approach

As mentioned in Chapter 1, the prevailing approaches for text mining tasks generally fall into a rule-based category and a machine-learning category depending on whether domain experts define the rules, or a machine automatically acquires the rules. The rule-based or pattern-matching approach relies primarily on human insights rather than a large training set, since the extraction rules are generated by humans instead of machines. In general, the manually developed rules have high accuracy in extracting text information, but it is time-consuming and requires considerable effort to develop the rules. The machine learning approach can automatically generate extraction rules by training a model. However, the accuracy of the machine learning approach may not be satisfying if the size of training data is not big enough.

Although there are more state-of-the-art machine learning algorithms, they do not often perform well in the domain of construction industry due to a limited data size (Tixier et al., 2016; Zou et al., 2017). Considering the subtle differences in patterns in the contingency liability related provisions and limited training data based on which the machine learning approach would likely lose its advantage, a rule-based model was developed to extract risk accountabilities from the contracts.

2.4.3. Classification Scheme of Contingency Liability in P3 Contract

The first task was to define and categorize the text for extraction. The process starts with identifying key provisions from the P3 contract documents. We also drew on articles related to contractual concerns during the COVID-19 pandemic as a supporting and supplementary source of knowledge (Finkel et al., 2020; Hansen, 2020)

Based on the domain knowledge and the review of relevant literature, the target texts related to contingency liabilities were identified. More specifically, a party may seek monetary compensation for the adverse financial effects, and/or claim a non-monetary relief such as an extension or a relief from performance of its obligation, and/or have the right to terminate the contract. Meanwhile, a contract may also include certain languages to limit the rights and liabilities of a party, such as the maximum amount recoverable in a compensation event or a required release of the right to compensation. Based on the purpose and meaning, we classified the target text into five categories:

1) **Remedy Entitlement**: The affected party will have the right to be entitled to the increased costs or time extension incurred by a contingency event.
2) **Remedy Obligation**: One or both parties will have the duty to compensate for the damages of the contingency event once it has occurred.
3) **Liability Waiver**: The affected party will be excused from the performance of its obligations under the contract.
4) **Mitigation**: The affected party will use its reasonable effort to prevent or mitigate the effects of a contingency event.
5) **Termination**: One or both parties will have the right to terminate the contract (usually in an extended or significant force majeure event).

2.4.4. Model Development

The development of the automatic provision extraction model includes three steps detailed as follows.

Step 1: Identification of Text Features of Target Text

Now that the target text had been defined, we first identified the features that distinguish the target text from other parts of the contract. These features can be the unique words which are frequently or exclusively used in specific contexts, the unique structures of sentences or phrases, as discussed in Step 2 and 3.

Considering the characteristics of P3 contracts and the goals of the analysis, we chose to define the unit of analysis as a sentence ending with a period. Firstly, in these contracts, the precision in language usage ensures that each sentence is crafted carefully to convey a distinct and self-contained meaning. Furthermore, our text analysis aims to capture the nuanced expressions in contracts. By focusing on individual sentences rather than clauses, we ensure a more granular examination of the language related to contingency liabilities.

Step 2: Development of Extraction Rules

To represent these features, we defined extraction rules based on the context-specific structure of sentences. For instance, given the example sentence in Figure 2-1, we first classified it into the "Remedy Entitlement" type, since it states that if there is a contingency (in this case a relief event), one party will be entitled to compensation. The rule for this type of sentences consists of four components—Party_n modal_v entitled_n Impact_n—appearing in this particular order, where Party_n refers to *Concessionaire*, modal_v corresponds to *shall*, entitled_v corresponds to *be entitled to*, and Impact_n refers to *the Incurred Cost*. In other words, if a sentence mentions a

contingency event and matches this rule, then it will be identified as Remedy Entitlement. Note that a feature of contract language is that many long sentences are separated by semicolons, with the meaning of each semicolon-separated sentence being relatively complete and independent. Therefore, in the automatic extraction model, the above pattern-matching is processed within each semicolon- separated sentence.

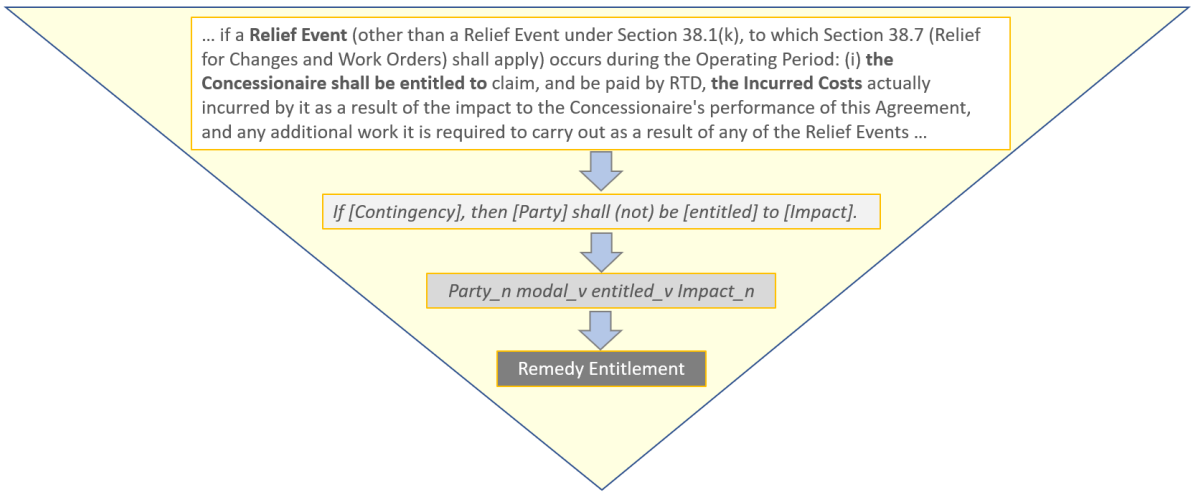


Figure 2-1 An example of sentence extraction

Except for the rule in Figure 2-1, more rules were defined to refer to Remedy Entitlement as well. For each of the five types of target sentences, the number of rules was 19 (Remedy Entitlement), 6 (Liability Waiver), 8 (Remedy Obligation), 1 (Mitigation), and 4 (Termination), respectively.

Step 3: Lexicon Development

The extraction model also consists of a domain-specific lexicon to provide instances, referred to as terms, for each component of the rule, so that the model will know which words and phrases should be recognized as Party_n, modal_v, entitled_v, Remedy_n, and so forth.

Table 2-2 summarizes the final lexicon. Firstly, the lexicon had a special type of terms named Contingency, representing a variety of contingency events. In this extraction task, the sentence will be a candidate sentence for subsequent pattern matching only if the sentence and/or its article title contains one or more Contingency terms. Besides Contingency, two other types of components comprised the matching patterns. Each of the components ending with _n contained a set of noun terms and each of the components ending with _v contained a set of verb terms. In total, 216 (65.7%) nouns and 113 (34.3%) verbs were added to the lexicon. Among the nine classes of nouns, the top classes as ranked by number of terms were Impact_n (87), Remedy_n (33), Party_n (32), Resource_n (25), and Obligation_n (23). Among the 15 classes of verbs, the top classes as ranked by number of terms were obligated_v (19), excused_v (15), remedied_v (13), modal_v (12), and remedy_v (10). For the sake of uniformity in the implementation process, all words were in their base forms.

Table 2-2 Lexicon profile

Class	Count	Example
Contingency	14	force majeure, relief event, compensation event, delay event, relevant incident, relevant event, uninsurability, uninsurable event, uninsurable risk, insurance unavailability, supervene event, extend event, extend delay
Exclusion_n	3	in no event, no election, no circumstance
Impact_n	87	impact, delay, effect, extra cost, increase in cost
Obligation_n	23	responsibility, performance, monetary obligation, timely payment
Party_n	32	(a) General names: (the) department, (the) concessionaire, (b)

		Project-specific names: txdot, fdot, ncdot, hpte
Remedy_n	33	compensation, remedy, amount, extension
Resource_n	25	cure period, time period, period of time, deadline date,
Right_n	7	right to, entitlement, right of recovery, any right
Risk_n	3	risk of loss, the risk, risk of delay
Termination_n	3	termination, election to terminate, suspension order
apply_v	8	apply, take effect, withdraw, be effective
excuse_v	3	excuse, waive, relieve
excused_v	15	be excuse, be waive, be relief from, be assess
entitle_v	2	entitle, permit
entitled_v	9	be entitle to, have the right to, elect to, be reimburse for
execute_v	9	execute, relinquish, assign, subrogate
limited_v	3	be limit, be reduce by, be exclude from
modal_v	12	shall (not), will (not), may (not), agree to
mitigate_v	4	mitigate, minimize, prevent, overcome
obligated_v	19	bear, responsible for, responsibility for, be liable for
remedy_v	10	compensate, pay, recover, owe, reimburse
remedied_v	13	be extend, be compensate, be pay, be modify
represent_v	4	represent, mean, refer to, constitute
suffer_v	1	suffer
terminate_v	1	terminate

The model development was a gradual and iterative process. At the outset, we selected three projects implemented in three different states: Midtown Tunnel (Virginia), Eagle Project (Colorado), and North Tarrant Express Segments 1 & 2A (Texas). Based on these contracts, we

built a preliminary lexicon and rules. Then the rules and lexicon were tested, augmented, and modified as more sentences were labeled (Figure 2-2). For example, the terms in entitle_v and entitled_v were considered as the same class at the beginning, but as the rules were refined, we found it was better to distinguish the active voice and the passive voice.

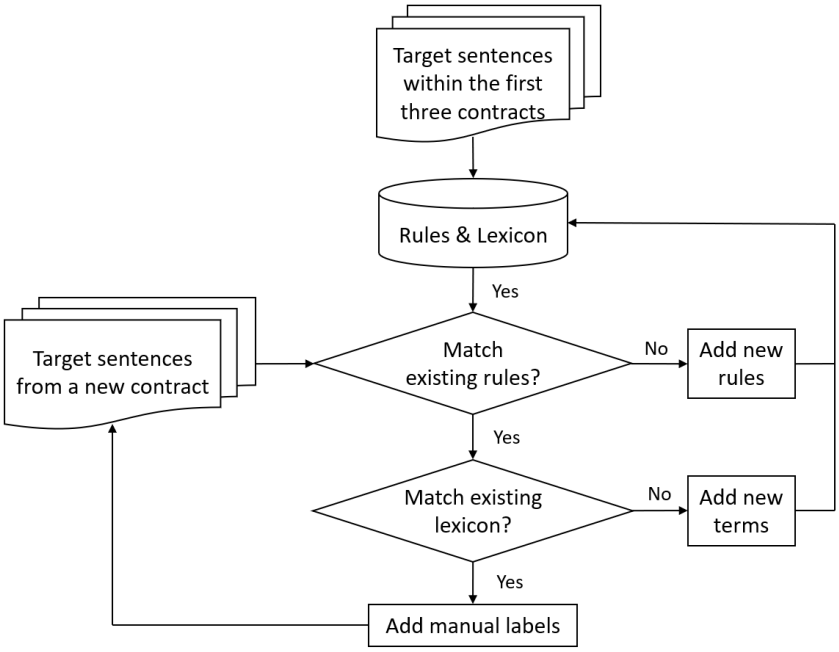


Figure 2-2 The process of extraction model development

To improve the extraction performance, we filtered the extraction output using a set of exclusion rules. These rules excluded the false positives that matched extraction rules but their lexicon collocations did not conform to semantic logic. For example, the sentence *"... the Concessionaire will implement and perform the Work in question as directed by the Department and the Department will make payments to the Concessionaire..."* was identified as Remedy Obligation since it included the terms that matched the rule Party_n modal_v remedy_v Impact_n. Clearly, the collocation of [remedy_v = *perform*, Impact_n = *payment*] did not make

sense in terms of semantic meaning, so this collocation was considered as an exclusion rule to filter out such false positives.

Table 2-3 illustrates how the pattern-matching mechanism works with a sample sentence for each type. For example, the first sentence was classified as Remedy Entitlement because it matched a rule in that class. Some sentences might be classified as more than one category because they matched multiple rules.

Table 2-3 Extracted sentence examples

Type	Sentence	Rule
Remedy Entitlement	For the Compensation Event described in clause (n) of the definition thereof, the Concessionaire will be entitled to recover the Net Cost Impact for such Compensation Event;	Party_n modal_v entitled_v Remedy_n
Remedy Obligation	Subject to any agreed scope of work and budget and to any rights of Developer in the case of a Compensation Event, Developer shall fully reimburse TxDOT for all costs and expenses, including TxDOT's Recoverable Costs, TxDOT incurs in providing such cooperation and assistance, including those incurred to conduct further or supplemental environmental studies, and in carrying out TxDOT actions.	Party_n modal_v remedy_v Impact_n
Liability Waiver	Developer shall be relieved from the performance of obligations, and Noncompliance Points shall not be assessed against Developer, as a result of Developer's inability to perform its obligations due solely and directly to, and during the duration of, the Relief Event.	Party_n modal_v excused_v Obligation_n

Mitigation	As soon as practicable following the notification referred to in Section 43.3, the Parties shall consult with each other in good faith and use reasonable endeavors to agree to appropriate terms to mitigate the effects of the Force Majeure Event and facilitate the continued performance of this Contract.	Party_n modal_v mitigate_v Impact_n
Termination	Either Party may deliver to the other Party written notice of its conditional election to terminate this Agreement under the following circumstances: 19.2.1.1 A Relief Event has occurred;	Party_n modal_v terminate_v

2.4.5. Model Implementation

Before implementing the model on our data set, text preprocessing was conducted using Python. Provisions of a contract are usually organized as articles, each comprised of multiple sections. To capture information in the titles, each contract document was split into sections which begin with articles titles (e.g., *ARTICLE 13. DELAY EVENTS* (I-95 Express Lanes)) and section titles (e.g., *Section 13.02 Delay Events During the Construction Period*). Regular expression was used to automatically search the start of each section and split the text. Then text cleaning was performed by removing some special characters and lowercasing the text. Python library NLTK was applied to lemmatize the text. In this step, all words were converted to the base or dictionary form (e.g., *was → be*, *contracts → contract*).

After all the text preprocessing steps, the data were ready for the extraction model. The workflow of model implementation includes the following: first, read in text data in the format of a list of section titles and section contents, tokenize each section into sentences. If the sentence and/or its article title contains a Contingency term, split each sentence by semicolons if there is

any. If the sentence is a conditional sentence, then any part beginning with *if, provided, unless, given, following, except that, without limiting, as a condition precedent to* is removed, so that only the terms in the remaining part that describe the consequence can be identified. For each of the five types of target sentences, and then for each of the rules, search every semicolon-separated sentence using the lexicon to determine if the sentence matches the rule. If the sentence matches any rule in a type, it will be extracted and labeled with the matched type, matched rule, and matched terms.

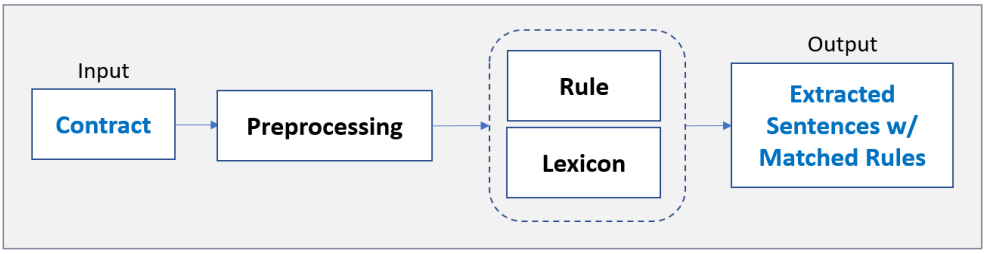


Figure 2-3 Data processing and extraction workflow

2.5. Results

For a binary classification, commonly used metrics are precision, recall, and F1-score (Tixier et al., 2016; Wimalasuriya & Dou, 2010).

$$Precision = \frac{TP}{(TP + FP)}$$

$$Recall = \frac{TP}{(TP + FN)}$$

$$F1\ score = \frac{2 \times Precision \times Recall}{(Precision + Recall)}$$

where true positives (TP) are the number of sentences the model correctly predicts the presence of the given label, false positives (FP) are the number of sentences that the model incorrectly

predicts the presence of that label, and false negatives (FN) are the number of sentences the model incorrectly predicts the absence of that label.

In multi-label classification, each data point can be assigned to one or multiple classes since the classes are not mutually exclusive. Common metrics for evaluating the overall performance of a multi-label classifier includes the macro (the unweighted mean of all the class-wise scores), micro (the mean of all score based on the sum of the TP, FN, FP of all classes), weighted (weighted mean of all the class-wise scores that consider each class's support, i.e., the number of actual occurrences of a class, when averaging the class-wise metrics). Since our dataset is imbalanced, we evaluate the model's overall performance using weighted precision, recall, and F1-score.

$$Weighted\ Precision = \sum_{i=1}^{N} \left(\frac{N_i}{N} \times Precision_i \right)$$

$$Weighted\ Recall = \sum_{i=1}^{N} \left(\frac{N_i}{N} \times Recall_i \right)$$

$$Weighted\ F1\ score = \sum_{i=1}^{N} \left(\frac{N_i}{N} \times F1\ score_i \right)$$

where N is the total sample size, N_i is the sample size of class i.

The performance of the extraction model was evaluated by two data sets: (1) 19 contracts that were used for model development, and (2) two contracts (I-495 & I-270 Program (Phase 1) and Presidio Parkway) reserved for validation. The result is shown in Table 2-4.

Table 2-4 Validation results

	Precision	Recall	F1-Score

19-project dataset			
Remedy Entitlement	0.899	0.944	0.921
Remedy Obligation	0.905	0.880	0.892
Liability Waiver	0.869	0.932	0.899
Mitigation	1.000	0.911	0.953
Termination	0.903	0.966	0.933
Overall (weighted)	0.896	0.934	0.914
2-project dataset (I-495 & I-270 Program (Phase 1) & Presidio Parkway)			
Remedy Entitlement	0.950	0.905	0.927
Remedy Obligation	1.000	0.800	0.889
Liability Waiver	0.700	0.824	0.757
Mitigation	0.800	1.000	0.889
Termination	1.000	1.000	1.000
Overall (weighted)	0.899	0.886	0.889

According to the results, the model achieves a higher weighted recall (0.934) than weighted precision (0.896) for the 19 contracts used in model development, resulting in a weighted F1-score of 0.914. For the two contracts reserved for validation, the model achieves a weighted recall of 0.899 and a weighted precision of 0.886, resulting in a weighted F1-score of 0.889. Overall, the model demonstrates reliable performance by capturing most of the common features among P3 contracts across different states through the past 15 years. Given the negative

relationship between precision and recall (Buckland & Gey, 1994), we needed to consider the tradeoff between these two metrics. In this task, recall was more important than precision because the cost of manually filtering out the false cases from the limited outputs is much lower than the cost of missing a positive case.

2.6. Conclusions

In light of the recent COVID-19 pandemic, we should not underestimate the danger and damage of occurrence of such unusual and unforeseeable contingency events. This chapter presents a rule-based model to automate the extraction of key sentences about contingency liabilities in P3 agreements. This model offered a tool to rapidly identify contingency provisions involving risk allocation. This is particularly helpful when a supervening event such as the COVID-19 pandemic happens which carries huge cost and schedule implications. In view of legal complexity, this automated approach does not substitute lawyers' job in reviewing legal implications of contracts. However, it provides a way of rapidly identifying key issues and relevant provisions. This automatic process also reduces human error. Additionally, more work can be done to add value. Firstly, machine learning approaches can be applied and compared with the rule-based model. Secondly, other types of contracts can be added into the dataset to develop a more generalized model.

Chapter 3. Fine-Tuning Language Models for Contact Analysis: Identifying Risk Type and Allocation

3.1. Abstract

Moving from Chapter 2, this chapter aims to identify broader risks associated with P3 contracts. To enable a precise and comprehensive identification of risk profiles within these contracts, two classification schemes tailored to analyze risks in P3 contracts are developed. With regards to NLP methods, the introduction of the Transformer architecture by Google researchers in 2017 revolutionized the NLP field. This boosts fine-tuning, a process of further training a PLM with task-specific data to allow the model to gain specialized understanding, to become a popular and effective way of solving domain-specific tasks. By leveraging the fine-tuning method, this chapter provides an efficient means of automated risk identification in P3 contracts.

3.2. Introduction

Contracts, with legal effects, are an important mechanism in project risk management. They formalize the agreement between contracting parties concerning the distribution of project risks. Despite extensive research on P3 risks and the acknowledged importance of contracts as a critical tool for project risk management, there is a lack of risk studies specifically addressing contract content analysis. The manner in which risks are defined and allocated within contracts is under-researched.

While contracts reflect most of the key project risks, their focus can be different from how risks are defined in other project documents, such as risk registers. The latter involves a wider range of project risks throughout the project lifecycle. In contrast, the primary role of a contract is to allocate risks, particularly those with the potential to lead to legal disputes among the contracting

parties. The distinction between the risks addressed in contracts and those covered in other project documentation is nontrivial. On the one hand, at the moment of negotiating or signing a contract, many early-stage project risks, such as “poor public decision-making process” (J. Li & Zou, 2011), “lack of a standard model for PPP agreements” and “low attraction of funding” (Le et al., 2019), should have been manifest or resolved and thus are not typically reflected in contracts. On the other hand, many important issues pertinent to contracts do not receive adequate attention in discussions on project risk, especially regarding the consequences of risks. Risks associated with contracts should involve those project risks that (1) focus on the allocation of rights and obligations and (2) can be explicitly articulated in contract language.

Previous studies have also examined risks outlined in the contract, with varying degrees of focus and comprehensiveness. On one hand, some studies have only focused on specific facets of contract risks. For example, Khalef et al. (2021) identified five risk categories under exculpatory clauses in construction contracts. On the other hand, some studies provided a comprehensive analysis of contract risks (e.g., Nguyen et al., 2018), with their risk breakdown framework offering valuable practical implications. However, such work is difficult to replicate in practice. It is an overwhelming task to scrutinizing risks in a contract even given a well-defined risk breakdown framework. This brings about another gap regarding the lack of an effective and replicable way of risk identification from contract documents. Due to the length of construction contracts, a thorough examination demands a considerable amount of work. Moreover, relevant content can be scattered throughout the contract, complicates the task of pinpointing specific risks.

With regard to methods, the rapid development in the field of NLP has provided unprecedented powerful tools for both researchers and practitioners. The introduction of the Transformer

architecture has significantly enhanced both computational efficiency and model accuracy. Since its release, many PLMs have been developed the approach of pre-training language models on large volumes of unlabeled data and then fine-tuning them for downstream tasks has become a widely adopted paradigm, which marked a significant shift in the NLP field (Han et al., 2021; J. Li et al., 2021; Xu et al., 2020). This has broadened the application of NLP techniques beyond traditional tasks, leading to new opportunities for a variety of domain-specific NLP tasks, including contract analysis.

In this context, this chapter aims to utilize PLMs for automated identification of risk within P3 contracts. Specifically, our objectives for employing these models are twofold: Firstly, to detect the presence of risk points within the contract documents; and secondly, to identify the allocation of these risks—for instance, determining whether the public or private partner is eligible for compensation under certain conditions.

For this purpose, this chapter first develops two specialized classification schemes tailored to analyze risks in P3 contracts:

1) **Risk Type Classification**. This scheme categorizes the various types of risks commonly found in P3 contracts. These include economic and financial risks, socio-political and legal risks, environmental and site risks, utilities, permits, and third-party risks, performance risks, changes by contracting parties, revenue risks, force majeure risks, disagreement and dispute risks, and others.
2) **Risk Allocation Elements Classification**. This focuses on classifying various elements related to the allocation of risks. This includes identifying and categorizing rights, obligations, liabilities, and prohibitions of the contracting parties.

These classification schemes provide a comprehensive perspective for understanding how risks are defined and allocated in P3 contracts. Following that, this chapter employs the approach of fine-tuning PLMs, specifically the BERT family of models, to realize automated risk type and allocation identification.

3.3. Literature Review and Theoretical Framework

3.3.1. Perspectives and Classifications of P3 Risks

This section offers an overview of the various ways risks are perceived and categorized in P3s as documented in existing literature. The literature review reveals that the perceptions of risks in P3s are multifaceted, often influenced by the nature of the project, the market environment, and the stakeholders involved. The literature broadly classifies these risks into several categories such as financial, operational, legal, environmental, and socio-political risks, each with its distinct characteristics and impact on project outcomes. Additionally, a temporal dimension is often considered, distinguishing between risks arising in the early project phases and those becoming evident during the construction or operational phases.

Table 3-1 presents the risks categories associated with P3s as identified in the literature from diverse global markets in chronological order, with a primary focus on the transportation sector, facilitating an understanding of the key risks that are widely recognized and have been consistently emphasized in P3 projects around the world.

Table 3-1 Overview of key risk categories in P3 literature

Authors	Project type	Targeted market	Risks category
Nicolini-Llosa (2002)	Transportation (toll road)	Argentina	Exchange rate, Growth in traffic demand, Political policies, Contract rescission, Renegotiation
Thomas et al. (2003)	Transportation (road)	India	1. Developmental phase (Pre-investment risk, Resettlement and rehabilitation risk, Delay in land acquisition, Permit/Approval risk, Delay in financial closure) 2. Construction phase (Technology risk, Design and latent defect risk, Completion risk, Cost overrun risk) 3. Operation phase (Traffic revenue risk, Operation risk, Demand risk, Debt servicing risk) 4. Project life cycle (Legal risk, Political risk (direct & indirect), Partnering risk, Regulatory risk, Financial risk, Environmental risk, Physical risk, Non-political force majeure risk)
Grimsey & Lewis (2004)	General	General	Site, Technical, Construction, Operating, Revenue, Financial, Political, Project default Asset
Hodge (2004)			1. Design and development (Design Suitability Development Problems Testing Problems, Design and Development variations, Delivery of Design)

			2. Construction (Fixed time and cost to complete, Delivery schedule, Planning approvals, Environmental issues, Disruption to existing services, Site preparation, Transport of assets to site, Design and construction variations, Industrial disputes) 3. Finance (Securing finance, Maintaining finance, Interest rate and tax amendments, Tax rulings, Price escalation) 4. Operation (Asset/service performance and availability, Maintenance cost variations, Security, Staff training, Change to client requirements, Cost of keeping existing assets operational, Latent defects in existing assets Changes in demand, Third party revenue) 5. Ownership (Uninsurable loss or damage to the assets, Technology change or obsolescence, Federal and state legislation/regulation changes, Public/third party liabilities, Force majeure, Realization of the residual value of assets)
Bing et al. (2005)	General	UK	1. Macro level (Political and government policy, Macroeconomic, Legal, Social, Natural) 2. Meso level (Project selection, Project finance, Residual risk, Design, Construction, Operation) 3. Micro level (Relationship, Third party)
Abednego & Ogunlana (2006)	Transportation (tollway)	Indonesia	Political, Construction, Operation and maintenance, Legal and contractual, Income, Financial, Force majeure
Lie & Zou	General	General	1. Feasibility study (Land acquisition and compensation problems, Planning deficiency, Poor public decision-making

(2008)			process, permit/ approval risk) 2. Financing (Interest rate volatility, Financial Legislation change, Poor financial market, Inflation rate volatility, Little financial attraction of project to investors, Ill capital structure) 3. Design (Lack design flexibility, Too many design changes, Design deficiency) 4. Construction (Capital materialized problem, Completion delay, Too many late design variation, Construction cost overrun, Poor quality workmanship, Safety risk, Inflation rate volatility, Construction force majeure events) 5. Operation (Legislation change, Operation / maintenance cost overrun, Fluctuating market demand, Environment pollution, Operator inability, public opposition because of high product/service price, Operation safety problems, interest rate volatility, Inflation rate volatility) 6. Transfer (Little residual value, Transmission failure)
World Bank (2008)	Transportation (road)	General	Design, Site, Construction, Force Majeure, Revenue, Operation and Maintenance, Performance, External, Other Market Risk, Political, Default, Strategic Risks
Kwak et al.,(2009)	General	General	1. Political (Expropriation, reliability and creditworthiness of the government; Change in law and government policies; Political opposition; Corruption; Delay in approvals; Political force majeure events) 2. Financial (Unfavorable economy in the host country; Rate of return restrictions; Lack of credit worthiness; Inability to service debt; Bankruptcy; Complex financial structure of PPP projects; Lack of guarantees; Financing risks; Loan ability; Fluctuation of the inflation rate, interest rate, foreign currency exchange rate; Unfavorable international

			economy) 3. Construction (Land acquisition and compensation; Construction cost overrun; Construction time delay; Material/labor availability; Project site conditions; Contractor's failure; Construction force majeure events) 4. Operation and Maintenance (Operation and maintenance cost overrun; Operator's incompetence and low operating productivity; Availability of material; Force majeure events) 5. Market and Revenue (Insufficient revenue; Government restriction of profit and tariff; Inaccurate pricing and demand estimate; Fall of demand; The competition risks; Force majeure events) 6. Legal Risks (Prejudiced and unfair process of awarding the project; Host-country's interference in choosing subcontractors; Overprotective control/supervision by the host government; Disapproval of guarantees by the government; Change of host-country's fiscal regime; Change of host-country's consideration of the project's scope; Non-cooperation between public agencies; Actions or omissions of the public authorities that prevent the project to be completed; Unsteady legal and regulatory framework; Poor legislation; Nonenforcement of legislation; Lack of a stable project agreement; Vague and inconsistent clauses and specifications and inaccurate phasing; Non-accordance between all contracts in the BOT framework; Language barrier for the contract; Breach of contract provisions; Revision of the contract clauses; Unanticipated change of the concessionaire scheme; Lack of confidentiality and trust in the concession company; Risks of early termination; Legal force majeure events)
Papajohn et al.	Transport ation	U.S.	Traffic demand, Right-of-way, Environmental issues, Operation and maintenance costs, Political and governmental issues, Loss of owner control, Delays because of legal issues

(2011)			
Cheung & Chan (2011)	Water and wastewater, power and energy, and transportation	China	government intervention, public credit, financing risk, poor public decision-making process, subjective project evaluation method, completion risk, government corruption, price change, operation cost overrun, imperfect law and supervision system, project/operation changes, inability of concessionaire, inflation, conflicting or imperfect contract, interest rate fluctuation, insufficient project finance supervision, delay in project approvals and permits, inadequate competition for tender, foreign exchange fluctuation, change in market demand
Heravi & Hajihosseini (2012)	Transportation (toll Road)	General	1. Political (Termination of concession by government, Influential economic events, Change in law, Sanction) 2. Financial (Limited capital Delays, Need for land appraisal Poor transparency, Financial problems due to environmental protection, Inflation risk, Change in value of granted lands due to inflation, Change in value of granted lands due to development) 3. Market (Tariff change, Market demand change, Insufficient income, Competition) 4. Legal (Lack of standard model for PPP agreements, Ownership assets) 5. Construction (Need for land acquisitions Delays, Need for environmental approval Delays, Inadequate study and insufficient data Additional design work, Improper design)

			6. Operation and maintenance (Operation cost overrun, Operator default) 7. Organization and coordination (Organization risk, Coordination risk) 8. Force majeure (Severe weather, war, natural disasters)
(FHWA, 2014)	Transportation	U.S.	1. Development phase (Planning and environmental process, Political will, Regulatory, Site, Permitting, Procurement, Financing 2. Construction phase (Engineering and construction, Changes in market conditions) 3. Operation phase (Traffic, Competing facilities, Operation and maintenance, Appropriation, Financial Default Risk to public agency, Refinancing, Political, Regulatory, Handback)
Kumari & Kumar Sharma (2017)	General	General	1. Commercial risks (Technical, Construction, Operational, Performance, Demand, Input) 2. Financial risks (Equity, Accounting and economic, Liquidity, Bankruptcy, Counterparty, Refinancing, Fluctuations in exchange rates, Incorrect expectations of inflation, currency/devaluation) 3. Country and community risks (Riots and domestic disturbances, Currency inconvertibility and transfer, Breaches of contracts, regulatory/political) 4. Force majeure risks (Natural disasters, Terrorism and war) 5. Other risks (Lack of fidelity, Theft, Residual value)
Nguyen et al. (2018)	Transportation	U.S.	1. General (Financing, Socio-political opposition and protesters, Change in law, Refinancing, Inflation)

	(highway)	2. Pre-financial close (Interest rates pre-financial close) 3. Construction (Design, Rights of way and Easements, Additional properties, Site geology, Environmental risks, Archaeology/fossils/protected species, Access and adjustment to utilities, Permits, Environmental permits, Commodity price Adjustments, Changes by the public authority, Revenue during construction, Performance) 4. Operation (Usage/demand risk, Network, Payment for services, Availability and service, Operation expenses, Maintenance, Latent defects, Transfer, Project company default, Termination by the public authority, Force majeure, Residual value)
Le et al. (2019)	Transport General ation	1. Within specific phases of the project life cycle: Identification, Procurement, Design and construction, Operation and maintenance, Transfer 2. Across the project life cycle: Commercial, Financial, Legal, Political, Economic, Force majeure, Other

3.3.2. Classification Schemes of Risk Types in P3 Contracts

While existing risk classifications provide valuable insights, most of them are not well suited for contract analysis. The majority of risk analysis in literature is centered around assessing the probability and consequences of risks, whereas contract risk analysis is more focused on the precise definition and allocation of risk events. Currently, research specifically focused on contract risks is still insufficient (Fathi & Shrestha, 2023). Drawing from existing literature, this study develops two classification schemes for risks in P3 contracts, tailored to align with their unique features.

Classification Scheme 1: Risk Type.

The first scheme classifies risk types in P3 contracts, synthesized from previous studies (Federal Highway Administration, 2014; Le et al., 2019; Nguyen et al., 2018).

1) Economic and financial risks
    - Financing risks: Risks related to securing financial investors or favorable financial terms, adjustments in financing during the project lifecycle (e.g., changes in public funds amount or concession fee due to review of initial base case financial model; changes in the rates for Transportation Infrastructure Finance and Innovation Act (TIFIA) financing, bonds, and bank debt), and refinancing risks related to changing financial structures or agreements (e.g., changes in interest rates; stricter agreement)
    - Economic and market risks: Risks including inflation, interest rate fluctuations, foreign exchange fluctuations, commodity price changes, etc.
2) Socio-political and legal risks

- Socio-political risks: Risks arising from social and political environments, such as opposition by government agencies or public groups, and changes in government policies.
- Legal risks: Risks related to applicable laws and regulations (e.g., change in laws) that could impact on the project.

3) Environmental and site risks
   - Environmental risks: Risks arising from environmental regulations and compliance (e.g., the National Environmental Policy Act (NEPA) process), unexpected environmental conditions (e.g., presence of hazardous materials, contamination, and impacts on archaeological, paleontological, cultural resources, and endangered species. Such risks may lead to delays in project approval and increased costs for environmental mitigation.
   - Site risks: Risks associated with the ground conditions at the project site, such as site and subsurface conditions.
4) Utilities, permits, and third-party risks

   Difficulties and delays in coordinating with third parties as utility providers, local governments, other contractors, or community groups for utility adjustments and relocations, getting permits and approvals, acquiring right-of-way and easements, etc.
5) Performance risks

   Risks related to the potential for either party to default on contractual obligations, such as project delays, quality issues (e.g., failure to meet standards or availability), or failure to make timely payments. This can include performance risks from both private and public partners:

- Private partner's failure, negligence, or misconduct, such as design flaws, construction defects, failure to meet milestones or standards, and discrepancies in financial models, and failure to pay concession fee.
- Increased costs, delays or other issues arising from the public party's actions or failures to act in a timely manner (e.g., the Department ceases to provide all or a material part of the electronic toll collection services and as a result the Developer incurs costs related to self-performing; A delay caused by Department's failure to notify the Developer on time).

6) Changes by contracting parties

   Risks emerging from changes initiated by either party, such as alterations in project scope or specifications, change of contractors or service providers, and amendment or modification to contracts, proposals, or orders.

7) Revenue risks (or Demand and tolling risks)

   Risks related to inaccurate demand projections, toll collection and enforcement issues (e.g., closures or suspension of tolling caused by emergencies affecting toll incomes), and competing facilities that can affect usage or performance of the project.

8) Force Majeure

   Unforeseeable events beyond the control of the contracting parties, such as severe weather conditions, natural disasters, fire or explosion, strike, terrorism or sabotage, war, riot, and epidemics and quarantine restriction.

9) Disagreement and dispute risks

   Risks of disagreements or disputes between parties regarding contractual terms and any project-related issues.

10) Other

There is no explicit risk mentioned in the contract language; or, any underlying risk not belonging to the abovementioned nine types, including but not limited to: intellectual property rights, insurance issues, subcontractor issues, unavailability of a highway section, lane closure needs, public sector intervention, key personnel replacement, changes in ownership, damage to health and safety, damage to project assets due to accidents, latent defects, reinstatement work, retender, voluntary termination of either party.

Classification Scheme 2: Risk Allocation Elements.

Contracts employ specific language to stipulate the assignment of rights and responsibilities between the contracting parties, particularly in relation to risk allocation. This linguistic accuracy serves to delineate the obligations, rights, liabilities, and prohibitions pertinent to the contractual agreement. Obligation language specifies the duties and responsibilities incumbent upon each party. Liability language delineates how risks are shared, specifies who is liable for various types of risks or damages, such as the legal consequences for any breaches of the contract by either party. Right language defines the entitlements and privileges granted to each party under the contract. Prohibition language identifies the actions that are expressly forbidden by the contract stipulations.

In legal terms, “obligation”, “responsibility”, “duty”, and “liability” are often used to describe the requirement to do something. Although they are sometimes used interchangeably in common language, these terms have different meanings. “Obligation” usually refers to a legal requirement to do something. In a construction contract, an obligation might involve completing the work on time or to a specified quality standard. The term “responsibility” is often used interchangeably with “obligation”, but it typically implies a broader sense of accountability or management. In a

construction contract, responsibility may refer to the overall management and oversight of the project. A “duty” is a specific obligation that is usually imposed by law. In a construction contract, a duty may involve compliance with building codes, labor laws, or environmental regulations. Given the similar meaning of “obligation”, “responsibility”, “duty”, this study uses “obligation” as the umbrella term for them.

“Liability” typically refers to the state of being legally responsible for something, particularly the responsibilities and risks assumed by the contracting parties for damages, injuries, losses or other legal penalties that may occur during the construction process (Bunni & Bunni, 2022). The differentiation between liability and obligation sometimes presents a nuanced challenge. For example, consider the scenario depicted by the following sentence regarding a public liability: “If Financial Close does not occur by the Financial Close Deadline Date and such failure is attributable to one of the matters in Section 2.6(c) then HPTE shall pay the Stipend to the Concessionaire.” This provision stipulates that if the Financial Close is not achieved by the specified deadline due to listed reasons, HPTE (the public partner) becomes liable to compensate the Concessionaire (the private partner) with a Stipend. Here, the financial obligation is triggered by a failure to fulfill a condition, resulting in the legal responsibility (liability) to make a specific payment. It does not constitute an obligation in the sense of a requirement to perform a task but rather specifies the consequences (liability) for not achieving a particular contractual milestone.

The language for “right” in a construction contract is relatively more straightforward to understand and refers to the entitlements, privileges, or permissions granted to a party under the contract. These rights define what each party is allowed to do, such as access to the property, entitlement to payments, entitlement to documentation copies, authority over decisions about the

construction process, amendments to the scope of work, inspection of the work, option for dispute resolution, extensions of time, and termination of the contract (Pellegrino et al., 2013).

In a P3 contract, these terms serve to define and allocate the various responsibilities, obligations, rights, and potential liabilities between the contracting parties. It is important to carefully review the specific language and context of each term within a contract, given that the legal implications can significantly influence risk allocation.

This chapter introduces a second classification scheme aimed at identifying the rights, obligations, liabilities, and prohibitions of contracting parties as they pertain to elements of risk allocation within P3 contracts. This scheme encompasses two dimensions, which are designed to provide a comprehensive framework for understanding risk allocation elements between the public and private partners:

1) The party in focus, i.e., *public* or *private*.
2) The nature of the risk element, which can be one of the followings:
    - **Obligation** (to perform): Responsibilities and duties of a party involved in the contract, such as the provision of necessary materials and labor, obtaining required permits, timely completion of work or payment, adherence to specifications and standards, compliance with applicable laws and regulations, and mitigation of impacts in unforeseen events.
    - **Liability** (to bear risk effect): Legal responsibility or obligation of a party for any damages and losses arising during the project, particularly regarding financial compensation for such occurrences. This can include liabilities arising from accidents, property damage, breaches of contract, or other claims related to the project. This is often articulated or implied in contract language such as “shall

warrant/be responsible for something", "bear the cost/burden of matters arising from a risk event", or "be liable to/indemnify, defend, and hold harmless someone".

- **Right**: A legal entitlement or permission granted to a party under the contract, including but not limited to the right to receive payment, access information, receive notifications, request change orders, elect to dispute resolution or termination, and seek relief or compensation for risk events.
- **Prohibition**: Explicit restrictions or actions that parties are forbidden to undertake as stipulated in the contract.
- **None**: Language not mentioning any one of the abovementioned types, such as definition language which defines key terms and concepts used in the contract to ensure that all parties have a clear understanding of the contract terms.

The combination of the two partners leads to nine classes of risk allocation elements, including "public obligation", "public liability", "public right", "public prohibition", "private obligation", "private liability", "private right", "private prohibition", and "none".

Together, these two classification schemes provide a comprehensive framework for analyzing and understanding how risks are defined and allocated in P3 contracts. Again, our unit of analysis as a sentence ending with a period. This allows our analysis to capture the nuanced expressions of contract risks inherent in these contracts.

Table 3-2 presents several examples to demonstrate the labeling of sentences under each classification scheme.

Table 3-2 Examples regarding two classification schemes

Sentence	Risk type	Risk allocation element
… the **Department** will **bear the risk** and **have the benefit** of: (1) 100% of the impact on the Public Funds Amount or Concession Fee (either positive or negative) arising from changes in Benchmark Rates solely with respect to TIFIA financing, bonds, private placement and bank debt assumed and indicated in the Initial Base Case Financial Model, … ; **provided, however, that this protection will be extended only to the lesser of** (aa) the amount of proceeds of TIFIA financing, bonds, private placement debt and bank debt assumed and indicated in the Initial Base Case Financial Model, and (bb) the amount of proceeds of TIFIA financing, bonds, private placement debt and bank debt issued or incurred at Financial Close; …	economic and financial	public liability, public right,
The **Concessionaire recognizes that the requirements** of the Program Fraud Civil Remedies Act of 1986, as amended, 31 U.S.C. § 3801 et seq. and the USDOT regulations, "Program Fraud Civil Remedies," 49 C.F.R. Part 31, apply to its actions hereunder.	socio-political and legal	private obligation
Except as provided in Section 16.02, the **Developer** will **bear all costs and expenses** of preparing and complying with any Environmental Management Plan, of complying with Law and obtaining and complying with Governmental Approvals pertaining to Hazardous Substances, and otherwise	environmental and site	private liability

of carrying out Remedial Actions.		
The **Department**, to the extent permitted by law, will **assume responsibility for third-party claims** against the Developer or any Developer Party for personal injury, damages or harm to property or business due to a sudden release of Hazardous Substances by a Person other than the Developer that first occurs on the GP Lanes and concurrently or immediately thereafter is found to be present on the Express Lanes, and all related penalties, fines and administrative or civil sanctions arising out of or related to such sudden release of Hazardous Substances, **except to the extent such claims are due to the negligence, recklessness, illegal conduct or willful misconduct of the Developer or any Developer Party**.	environmental and site	public liability,
The **Developer** will **be responsible** for, and will **bear the costs and schedule risk** related to, coordination with WMATA regarding the WMATA Work, the WMATA Easement Impacts and the WMATA Incidental Impacts.	utilities, permits, and third-party	private obligation, private liability
If the **Department** intends to change any Commonwealth interoperability or compatibility standards, requirements or protocols for toll collection systems, it will **coordinate with the Developer prior to the implementation of such change** so as to minimize the loss of Toll Revenues, disruption and cost to the Developer, but the Department will **not be liable in any event for any loss of Gross Revenues, disruption or cost attributable to such change**.	changes by contracting parties	public obligation, public liability

If the Developer fails at any time to provide snow and ice removal to the Express Lanes at a level of service comparable to what the Department provides on the GP Lanes, the **Department may make a written determination to that effect**.	performance	public right
The **Department** will **have no liability to the Developer for the loss of Toll Revenues** or the increase in costs and expenses attributable to any such order issued pursuant to Law by the Department or any other Governmental Authority, provided that the Department: (i) concurrently (A) suspends tolling on all other Department-operated tolled facilities that are located within the area designated for evacuation or facilitation of evacuation and (B) orders suspension of tolling on all other tolled facilities operated by others within such area and over which the Department has the authority to order such suspension; and (ii) lifts the order on the Express Lanes before or concurrently with the lifting of the order for all other designated tolled facilities within the area designated for evacuation or facilitation of evacuation, which the Department shall endeavor to do as promptly as possible (taking into account the relevant emergency and the duration thereof).	revenue	public obligation, public liability
The **time periods specified in Section 10.05(c) will be extended** for the duration of a Force Majeure Event that prevents the **Department** or the **Developer**, as applicable, from performing under this Section 10.05.	force majeure	private right, public right
If no agreement is reached within such 60-Day period as to any such matter, **either party may**	disagreement	private right,

submit the Dispute to the dispute resolution procedures set forth in Article 21.	and dispute	public right
Regardless of whether the Department's consent is required, the **Developer will provide the Department notice** of all such proposed amendments, supplements and modifications and **will pay the Department**, upon demand, for all the **Allocable Costs it incurs** to review and consider proposed amendments, supplements or modifications that are subject to the Department's approval.	changes by contracting parties	public right, private obligation, private liability

3.4. Methods

3.4.1. Transformer and PLMs

Transformer (Vaswani et al., 2017) is a neural network architecture that revolutionized NLP by introducing global attention mechanisms. Its self-attention mechanism does not rely on sequential data processing; instead, it handles all parts of the input simultaneously. This significantly addressed many major concerns of previous sequence-to-sequence models with global context understanding and executing parallel processing (Y. Cao et al., 2023).

Since then, numerous transformer-based language models have been released successively. Specifically, the Transformer has an encoder-decoder structure, and there are encoder-only models represented by BERT (Devlin et al., 2019), decoder-only models like GPT (Radford et al., 2018), and encoder-decoder models such as T5 (Raffel et al., 2020) and BART (Lewis et al., 2019). BERT captures deep bidirectional contexts by processing both the left and the right context in all layers, while GPT generates text in a left-to-right fashion. BERT is designed to process input data in a bidirectional manner, making it highly effective for tasks that require context understanding such as text classification. GPT focuses on generating text and is trained to predict the next word in a sequence, making it powerful in tasks involving human-like text generation. These models are characterized by millions of parameters, significantly larger than those of earlier deep neural networks. They can capture hierarchical linguistic information in a way that mirrors the traditional NLP pipeline steps from POS tagging, parsing, NER, semantic roles, to coreference (Tenney et al., 2019). Beyond handling common NLP tasks such as sentiment analysis and named entity recognition, these transformer-based models can solve complex sequence-to-sequence tasks more effectively, outperforming other deep learning

methods in tasks like question answering, machine translation, text summarization, and contextual understanding, by capturing long-range dependencies in data (Qiu et al., 2020).

The subsequent years saw the release of even larger models, also known as LLMs such as GPT-2 (Radford et al., 2019), T5 (Raffel et al., 2020), and GPT-3 (T. Brown et al., 2020), with parameter counts reaching into the billions, representing a major advance in the field. These models, trained on immense amounts of data comprising web pages, Wikipedia, GitHub repositories, books, academic papers, and more, acquired a formidable ability to generate language based on input prompts. Despite their size and capabilities, these LLMs initially had a limited impact compared to smaller PLMs due to the high computational demands of fine-tuning such large parameters. Fine-tuning LLMs with billions of parameters is very costly, which often places them beyond the reach of individuals and smaller organizations. Considering the intensive data requirements for effectively fine-tuning larger models, this study chooses smaller, yet robust PLMs.

3.4.2. Pre-Training and Fine-Tuning of PLMs

The process of pre-training followed by fine-tuning language models has become a successful and widely adopted paradigm since the introduction of the Transformer architecture (Han et al., 2021; W. X. Zhao et al., 2023). Pre-training such a model involves training on a large corpus of unlabeled text (Qiu et al., 2020). The model takes an embedding, i.e., a vector of real numbers, as the representation of each input data. The input embeddings are comprised of three embeddings: token embeddings that captures the meanings and relationships of tokens, segment embeddings that record the sentence number to help the model understand sentence boundaries, and position embeddings that encode the order of tokens within the input sentence since transformer-based models do not process text sequentially as RNNs do.

Particularly for BERT, two unsupervised tasks are conducted simultaneously during pre-training, namely Masked Language Modeling (MLM) and Next Sentence Prediction (NSP). In MLM, the model takes in the input sentence with some of its tokens masked randomly, and the goal is to predict these masked tokens based only on their context. In NSP, the model receives two sentences and is required to determine whether one sentence logically follows another. This pre-training process makes BERT gain a good understanding of language and context.

The pre-trained model can then be fine-tuned for various downstream tasks such as text classification and question answering. Fine-tuning involves initializing a task-specific output layer with the pre-trained parameters and then further training the parameters on a labeled, task-specific dataset. This process adapts the generalized language understanding developed during pre-training to the specifics of the target task.

3.4.3. Model Selection

We chose models from the BERT family, the most representative PLMs, for fine-tuning. Four BERT models are selected for the fine-tuning process:

1) **BERT-base**. This is the original BERT model developed by Google (Devlin et al., 2019). It features 12 transformer layers, 768 hidden units, 12 attention heads, and a total of 110 million parameters. It is pre-trained on a large corpus of text from the 800 million words of BookCorpus and 2,500 words of English Wikipedia for understanding language context and meaning. With its balance between size and performance, it serves as a benchmark for comparisons with other BERT family of models.
2) **RoBERTa-base**. RoBERTa, which stands for Robustly Optimized BERT Pretraining Approach, is an improved BERT version developed by Facebook AI to improve its performance (Y. Liu et al., 2019). With 12 layers, 768 hidden units, 12 attention heads, and

125 million parameters, RoBERTa-base outperforms BERT across a range of benchmark datasets. Its ability to better capture context and details may result in superior performance in identifying complex risk-related information in contracts.

3) **LegalBERT-base**. This is a domain-adaptation of BERT, designed to support legal NLP research and applications (Chalkidis et al., 2020). Pre-trained on 12 GB of English legal text, it is a light-weight model with only 6 layers, 512 hidden units, 8 attention heads, and 35 million parameters (only 33% the size of BERT-base). Given its enhanced understanding of legal language, this model is particularly well-suited for contract analysis.
4) **XLM-RoBERTa-base**. XLM-RoBERTa (Cross-lingual Language Model-RoBERTa) extends RoBERTa to multiple languages (Conneau et al., 2020). With 12 layers, 768 hidden units, 12 attention heads, and 720 million parameters, it is trained on 2.5 TB of CommonCrawl data in 100 languages.

3.4.4. Data

The contract of two P3 projects from two states in Table 2-1, US 36 Express Lanes (Phase 2) and Transform 66, are selected as the dataset for manual coding, resulting in the dataset comprising 2,114 sentences in total. Consistent with Chapter 2, the unit of content analysis is still a sentence ending with a period. Each contract is split into individual sentences. For each sentence, the risk type and risk element labels are assigned through a manual annotation process. Moreover, considering the significance of hierarchical titles within contracts—such as primary titles (e.g., ARTICLE 7. PROJECT FINANCING; FINANCIAL CLOSE; LENDER RIGHTS AND REMEDIES; REFINANCING) and secondary titles (e.g., Section 5.04 Suspension of Tolls)—in providing contextual information about the clauses under analysis, the study extends beyond the mere examination of sentences. For this purpose, an augmented dataset was prepared in addition

to the vanilla dataset that solely contains sentences. This augmented dataset integrates the article and section titles with the sentences, linking these three components with "#". An instance of the input data within this dataset is structured as follows:

"ARTICLE 14. COMPENSATION EVENTS; DEPARTMENT CHANGES; DEVIATIONS; NET COST SAVINGS # Section 14.02 Department Changes # (v) If the Department approved in writing the Change Proposal Estimate and the Developer delivered the Change Proposal then, within 30 Days following the delivery of the Change Proposal, the Department will pay to the Developer the lesser of (1) the Developer's Allocable Costs for preparation of the Change Proposal or (2) the amount of the Change Proposal Estimate."

In preparing our dataset for fine-tuning the BERT family of models, the dataset is randomly divided into two subsets: 80% for training (1,692 sentences) and 20% (422 sentences) reserved for testing. The training set was further divided into four equal parts to perform 4-fold cross-validation to enhance model robustness and mitigate overfitting risks (Arlot & Celisse, 2010; Ghojogh & Crowley, 2023). This allowed for the optimization of model hyperparameters and provided more reliable results across folds.

3.5. Experiments and Results

3.5.1. Hyperparameter optimization

The fine-tuning process will involve adjusting several critical hyperparameters to explore different configurations for optimal results. This typically includes adjusting the batch size, learning rate, and number of epochs.

- **Batch size** is the number of training samples that are fed to the model in one iteration of training, which in other words is the size of the subset of the training data that is used to compute the gradient and update the model weights during training.

- **Learning rate** is the step size that gradient descent takes to minimize the loss function during optimization. It controls the magnitude of the model's weight updates between epochs.
- An **epoch** is a complete pass through the entire training dataset. The number of epochs is the total rounds of these complete iterations over the training process, which impacts the model's opportunity to learn from the data.

Adjusting these hyperparameters can help find the most effective combination for training the model by achieving a balance between learning efficiency and model accuracy.

To determine the optimal combination of hyperparameters, we conducted a series of tests on the RoBERTa-base model using the vanilla dataset that only comprises sentences. Typically, batch sizes are chosen from 2, 4, 8, 16, 32, etc. Given the size of our training set, we find that 8 is the optimal batch size. Smaller batch sizes than that may result in unreliable performance in the updates, while larger batch sizes may not yield optimal performance in our scenario given the limited size of the training set.

Regarding the learning rate, an excessively small learning rate can lead to slow convergence, while a high learning rate might cause the model to overshoot the minimum or even diverge. We first test 1e-6, which takes too long to converge; when increasing to 1e-5, the optimal interval can be reached within 40 epochs. As we increasing learning rate to 2×1e-5, the convergence process become unstable. Similarly, 1e-4 is also excluded as the model is unable to find the optimal interval during the training process.

Regarding the number of epochs, the training process began to reach the optimal interval (loss < 0.01) after 25 epochs, therefore we select the number of epochs a bit larger as 35. The loss converges and stabilizes after approximately 25 epochs, reaches the lowest point around 35

epochs, and begins to fluctuate after 40 epochs, meaning more epochs will lead to overfitting. The model's weighted F1-score peaks at 35 epochs, as illustrated in Figure 3-1. Therefore, batch size = 8, learning rate = 1e-5, and number of epochs = 35 is selected as the optimal hyperparameter configuration applied to the fine-tuning of the selected BERT family of models.

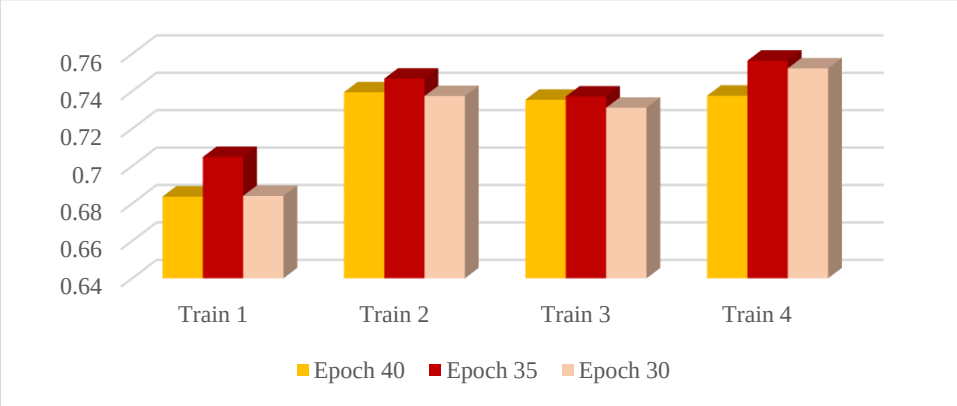


Figure 3-1 Change of weighted F1-score with different number of epochs

For each classification task, we conducted a 4-fold cross-validation for selecting the optimal hyperparameters and evaluating the consistency of model outcomes. Subsequently, we used the aggregated data from the four partitions to train the models. For each model from the BERT family that was selected, we obtained a final fine-tuned version, which was then applied to the test set to generate labels for performance evaluation.

3.5.2. Results

For this multi-label classification, we also evaluate the model's overall performance using weighted precision, recall, and F1-score. Additionally, we employ subset accuracy and Hamming Loss as other overall metrics. Subset accuracy is the proportion of data points for which the set of predicted labels exactly matches the set of true labels (Godbole & Sarawagi, 2004). Hamming Loss measures the fraction of incorrectly predicted labels to the total number of labels

(Tsoumakas & Katakis, 2007). The performance of each fine-tuned model on the test set is presented in Table 3-3.

$$Subset\ Accuracy = \frac{Number\ of\ perfectly\ matched\ samples}{N}$$

$$Hamming\ Loss = \frac{1}{N}\sum_{i=1}^{N}\frac{xor(y_i, \hat{y}_i)}{L}$$

where N is the total sample size, L is the number of classes, y_i is the true label of the ith sample, $\hat{y}_i$ is the predicted label of the ith sample. $xor(y_i, \hat{y}_i)$ is the XOR operation between y_i and $\hat{y}_i$, which equals 1 when $y_i = \hat{y}_i$ and 0 otherwise.

Regarding model selection, RoBERTa-base demonstrated the most superior performance among the selected models. Therefore, it is selected for further processing.

Table 3-3 Performance comparison across selected PLMs using the vanilla dataset

	Weighted Precision	Weighted Recall	Weighted F1-Score	Subset Accuracy	Hamming Loss
Classification 1: Risk type					
BERT-base	0.716	0.620	0.660	0.590	0.066
RoBERTa-base	**0.766**	**0.747**	**0.754**	**0.706**	**0.052**
LegalBERT-base	0.737	0.684	0.705	0.647	0.059
XLM-RoBERTa-base	0.774	0.722	0.743	0.680	0.054
Classification 2: Risk allocation					
BERT-base	0.693	0.615	0.648	0.445	0.111
RoBERTa-base	**0.791**	**0.763**	**0.775**	**0.607**	**0.075**

LegalBERT-base	0.708	0.684	0.693	0.517	0.102
XLM-RoBERTa-base	0.756	0.734	0.743	0.595	0.085

3.5.3. Enhanced model efficacy on augmented dataset

Up to now, the experimental evaluations were exclusively conducted on the vanilla dataset consisting of isolated sentences. Given the superior outcomes exhibited by the RoBERTa-base model on this dataset, we extended its application to the augmented dataset designed to integrate contextual information from article and section titles for potentially superior performance. The augmented dataset's inclusion of contextual cues was hypothesized to provide the model with a richer feature set for analysis, thereby improving its predictive accuracy.

The results, as shown in Figure 3-2 and Table 3-4, indicate the efficacy of incorporating contextual information into the dataset. The RoBERTa-base model demonstrated a significant improvement in its performance, achieving a weighted F1-score of 80.6% and 80.6% for risk type classification and allocation classification tasks, respectively, compared to the 73.5% and 78.5% scores using the vanilla dataset.

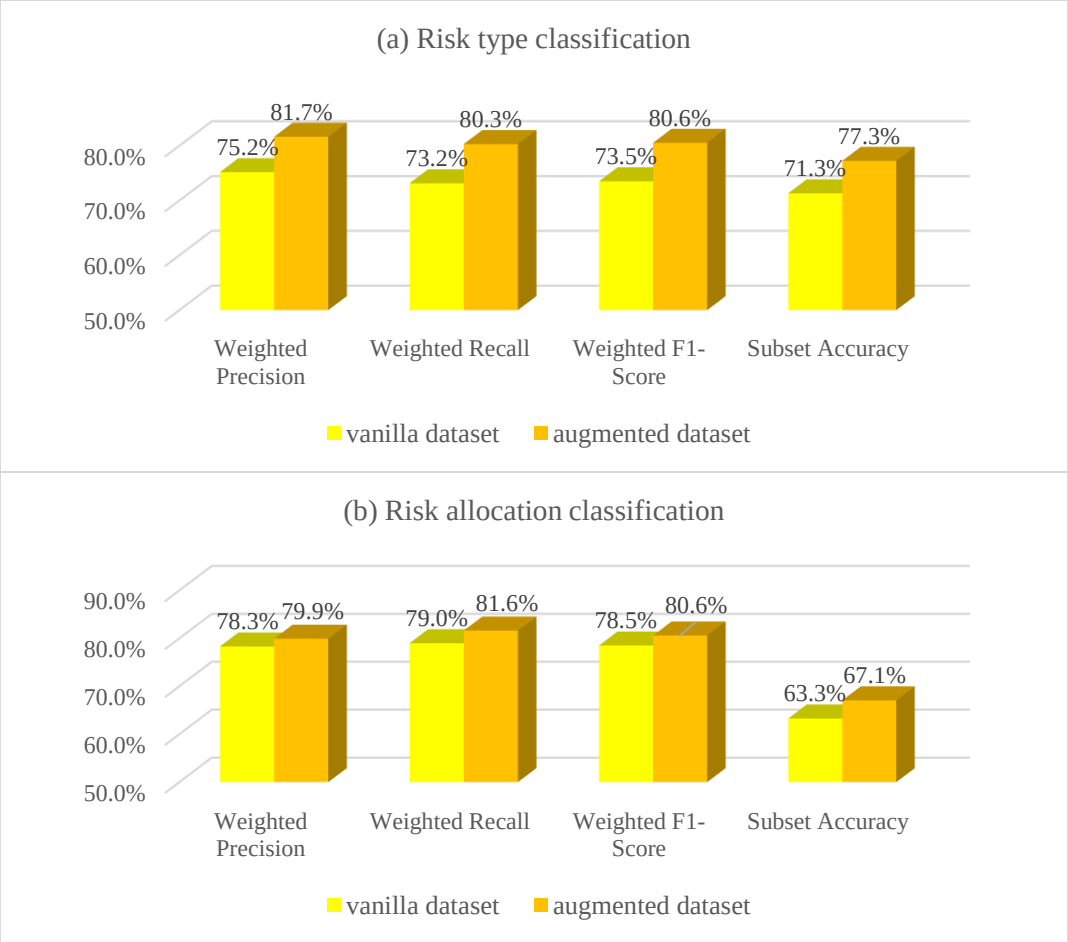


Figure 3-2 Performance of RoBERTa-base using different datasets. (a) Risk type classification. (b) Risk allocation classification.

Table 3-4 Performance of RoBERTa-base using the augmented dataset

	Precision	Recall	F1-Score	Support	Subset Accuracy	Hamming Loss
Classification 1: Risk type						
Economic and financial	0.781	0.781	0.781	32	-	-
Socio-political and legal	0.643	0.563	0.600	16	-	-

Environmental and site	1.000	0.938	0.968	16	-	-
Utilities, permits, and third-party	0.962	0.735	0.833	34	-	-
Performance	0.745	0.886	0.809	79	-	-
Changes by contracting parties	0.773	0.548	0.642	31	-	-
Revenue	0.750	0.714	0.732	21	-	-
Force majeure	0.667	0.500	0.571	4	-	-
Disagreement and dispute	0.947	0.783	0.857	23	-	-
Other	0.827	0.853	0.840	191	-	-
Overall (weighted)	**0.817**	**0.803**	**0.806**	-	**0.773**	**0.041**
Classification 2: Risk allocation						
Public obligation	0.844	0.783	0.812	83	-	-
Public liability	0.688	0.579	0.629	19	-	-
Public right	0.762	0.872	0.813	125	-	-
Public prohibition	0.333	0.250	0.286	4	-	-
Private obligation	0.883	0.883	0.883	188	-	-
Private liability	0.548	0.607	0.576	28	-	-
Private right	0.730	0.833	0.778	78	-	-
Private prohibition	1.000	0.923	0.960	13	-	-
None	0.785	0.718	0.750	71	-	-
Overall (weighted)	**0.799**	**0.816**	**0.805**	-	**0.671**	**0.063**

3.6. Applying the Fine-Tuned Model

In this section, we implement the RoBERTa-base model, distinguished as the most effective among the PLMs tested for the task of label prediction. With this application, we examine the risk profiles associated with these contracts.

First, the RoBERTa base model is fine-tuned using the optimal hyperparameter configuration and augmented input dataset format, and then the model is applied to perform classification tasks on unlabeled datasets. This unlabeled dataset consists of 12 contracts for P3 projects in eight states in the U.S. Additionally, incorporating the two contracts used for fine-tuning, the final dataset includes 14 contracts (20,493 sentences in total) from nine states, as shown in Table 3-5. These contracts include eight DBFOM toll concessions, five DBFOM availability payment concessions, and one long-term lease. Following that, the risk profiles of these contracts are generated and analyzed, as discussed in Section 5.2.

Table 3-5 Overview of P3 contracts analyzed via fined-tuned model

Project	Jurisdiction	Commercial close	Payment method
Indiana Toll Road	IN	2006	Lease
I-495 Capital Beltway Express	VA	2007	Tolled
SH 130: Segments 5 & 6	TX	2007	Tolled
Midtown Tunnel	VA	2011	Tolled
Presidio Parkway	CA	2011	AP
I-95 Express Lanes	VA	2011	Tolled
US 36 Express Lanes (Phase 2)	CO	2013	Tolled
I-4 Ultimate	FL	2014	AP

I-69 Section 5	IN	2014	AP
I-77 Express Lanes	NC	2014	Tolled
Pennsylvania Rapid Bridge Replacement	PA	2015	AP
Southern Ohio Veterans Memorial Highway	OH	2015	AP
Transform 66	VA	2016	Tolled
I-495 & I-270 Program (Phase 1)	MD	2021	Tolled

Chapter 4. Prompt Engineering with LLMs for Contact Analysis: Identifying Risk Type and Allocation

4.1. Abstract

In recent years, particularly since 2022, the field of LLMs has emerged and undergone exponential growth, with the GPT series developed by OpenAI standing out due to its broad applicability and advanced capabilities. This chapter explores the utilization of LLMs' in-context learning ability by examining various prompting techniques: zero-shot, few-shot, and chain-of-thought prompting. Moving forward from a baseline prompt design, more prompting strategies and model configurations are designed and implemented to test the capability of GPT-3.5 on the contract risk classification tasks and explore the paths for optimized performance. The goal is to design effective prompt formats for leveraging LLMs to classify risk-related sentences in P3 contracts. This application not only validates the adaptability of LLMs in the specific domain but also provides user-friendly automated tools for contract analysis.

4.2. Introduction

As PLMs continued to scale up in size, they have evolved into LLMs such as the OpenAI's GPT family, Meta AI's LLaMa-2 family, and Google's PaLM family, marked by their extensive datasets and enormous number of parameters. This scaling has led to LLMs' emergent abilities, i.e., extraordinary abilities not typically seen in smaller PLMs like BERT. A notable emergent ability is in-context learning, which allows an LLM to generate high-quality responses from a well-constructed prompt comprising instructions and example demonstrations, bypassing the need for further fine-tuning on domain-specific datasets to update the model's parameters (W. X. Zhao et al., 2023). In light of this new AI technology, this chapter shifts focus from the

methodology employed in Chapter 3 to the state-of-the-art approach—employing prompt engineering with large general-purpose language models—for the same tasks, i.e., automated identification of risk types and allocation in contracts.

4.3. Fundamentals of LLMs: Background Knowledge and Key Concepts

4.3.1. Evolution and Advancements in LLMs

As demonstrated in Chapter 3, a PLM can be further adapted to downstream tasks through the process of fine-tuning with domain-specific data. While this approach is efficient, it still requires considerable effort in data preparation and training. Therefore, researchers since 2019 have been exploring how to leverage the potential of LLMs without the need for fine-tuning on individual tasks (T. Brown et al., 2020; Radford et al., 2019). The period between 2019 and 2022 marked significant advancements in the field of LLMs. During this time, researchers focused on enhancing the adaptability and efficiency of LLMs, particularly in applying these models to diverse tasks. Released in 2019, GPT-2 (Radford et al., 2019) made a significant impact in the AI community with its advanced ability in text generation. The Text-to-Text Transfer Transformer (T5) (Raffel et al., 2020) reframed all NLP tasks as a text-to-text problem, simplifying the process of applying language models to a wide range of tasks. GPT-3 (Brown et al., 2020), with its 175 billion parameters, showcased exceptional ability in language understanding and generation. The model was able to perform various tasks without task-specific fine-tuning. It introduced the concept of few-shot learning, a technique where only a small number of examples are provided in the input to guide the model's response generation. The introduction of instruction tuning (H. W. Chung et al., 2022), a strategy to fine-tune a PLM on a dataset of prompts paired with desired responses, has significantly enhanced the performance and reduced the need for extensive task-specific fine-tuning. With the release of ChatGPT in November 2022,

LLMs rapidly attracted a large user base and entered into a breakthrough stage. As an application of the GPT-3.5 model, ChatGPT is designed specifically for conversational contexts. GPT-3.5, an iteration of the GPT-3 model, includes improvements in its understanding, context handling, and response generation compared to its predecessor. ChatGPT leverages the advanced capabilities of GPT-3.5 to effectively answer questions, simulate conversation, and perform various language-related tasks in a chat-based interface. In March 2023, GPT-4 (OpenAI, 2023) was released with further improvements. Particularly noteworthy is its multimodal understanding capabilities, which enable the model to process more types of data such as images. Table 4-1 shows the evolvement of GPT series models.

Table 4-1 GPT series models comparison

	GPT	**GPT-2**	**GPT-3**	**GPT-3.5**	**GPT-4**
Technical paper	Radford et al. (2018)	Radford et al. (2019)	Brown et al. (2020)	-	OpenAI (2023a)
Parameters	117 million	1.5 billion	175 billion	1.3/6/175 billion	>1.7 trillion
Training set size	-	40GB	570GB	-	-
Context window	512 tokens	1,024 tokens	2,048 tokens	4,096 tokens	-

Except for GPT models, popular general-purpose LLMs include LLaMA (Touvron, Lavril, et al., 2023), LLaMa-2 (Touvron, Martin, et al., 2023), PaLM (Chowdhery et al., 2022), BLOOM (Scao et al., 2022), Ernie 3.0 Titan (S. Wang et al., 2021), Claude (Anthropic, 2023), FLAN (J.

Wei, Bosma, et al., 2022), etc. Many task-based LLMs and their applications in diverse domains, such finance, biomedicine, law, have also been developed (Ling et al., 2023; Pahune & Chandrasekharan, 2023; Sun, 2023). Table 4-2 lists the 20 top models as ranked on three of the most popular LLM leaderboards as of March 21, 2024. These leaderboards evaluate and compare different LLMs across diverse tasks based on metrics such as crowdsourced preference votes and multiple-choice benchmark tests. More comparison across LLMs have been discussed in various survey articles (Minaee et al., 2024; Naveed et al., 2023; W. X. Zhao et al., 2023).

Table 4-2 Top ranked LLMs on leaderboards

	AlpacaEval 2.0		LMSYS Chatbot Arena		MMLU	
Rank	Model	Length-controlled win rate	Model	Arena Elo	Model	Average (%)
1	GPT-4 Preview	50.00%	GPT-4-1106-preview	1251	Gemini Ultra ~1760B	90
2	Claude 3 Opus	40.40%	GPT-4-0125-preview	1249	Claude 3 Opus	86.8
3	GPT-4	38.10%	Claude 3 Opus	1247	Leeroo	86.6
4	Qwen1.5 72B Chat	36.60%	Bard (Gemini Pro)	1202	GPT-4 ~1600B	86.5
5	GPT-4 0314	35.30%	Claude 3 Sonnet	1190	GPT-4	86.4
6	Claude 3 Sonnet	34.90%	GPT-4-0314	1185	Gemini Ultra	83.7
7	Mistral Large	32.70%	GPT-4-0613	1159	Flan-PaLM 2-L	81.2

8	Samba CoE v0.2 (best-of-16)	31.50%	Mistral-Large-2402	1155	Gemini Pro	79.1
9	GPT-4 0613	30.20%	Qwen1.5-72B-Chat	1146	Claude 2	78.5
10	Snorkel (Mistral-PairRM-DPO+best-of-16)	30.00%	Claude-1	1145	PaLM 2-L	78.3
11	Contextual AI (KTO-Mistral-PairRM)	29.70%	Mistral Medium	1145	Qwen1.5-72B	77.5
12	PairRM 0.4B+Yi-34B-Chat (best-of-16)	28.80%	Claude-2.0	1126	Claude 1.3	77
13	Mistral Medium	28.60%	Mistral-Next	1123	Leeroo	75.9
14	Claude 2	28.20%	Gemini Pro (Dev API)	1118	Camelidae-8×34B	75.6
15	Samba CoE v0.2	27.60%	Claude-2.1	1115	Flan-U-PaLM 540B	74.1
16	Claude	27.30%	Mixtral-8x7b-Instruct-v0.1	1114	Flan-PaLM 540B	73.5
17	Yi 34B Chat	27.20%	Claude-2.1	1115	Claude Instant 1.1	73.4
18	Snorkel (Mistral-PairRM-DPO)	26.40%	Mixtral-8x7b-Instruct-v0.1	1114	Flan-PaLM	72.2

19	Claude Instant 1.2	25.60%	GPT-3.5-Turbo-0613	1113	code-davinci-002 175B + REPLUG LSR	71.8
20	Claude 2.1	25.30%	Gemini Pro	1110	Gemini Pro	71.8

4.3.2. Emergent Abilities of LLMs

As PLMs scales in size and complexity, they began to exhibit unique abilities, known as emergent abilities, which are not present in smaller models like BERT (J. Wei, Tay, et al., 2022). These emergent abilities showcase the transformative potential of LLMs with their significantly improved language understanding and interaction. Two notable emergent abilities in LLMs that significantly broaden the potential applications of LLMs include:

- **In-context learning**: In-context learning refers to LLM's ability to learning from example demonstrations in the context (Dong et al., 2022; Min et al., 2022). It allows an LLM to generate high-quality responses based on a few input-label pairs provided in the input text. This ability eliminates the need for additional fine-tuning on domain-specific datasets to update the model's parameters. Moreover, it simplifies human intervention and enhances interpretability, given that the demonstrations are represented in natural language. In-context learning represents a paradigm shift in model utilization.
- **Multi-step reasoning**: LLMs display advanced capabilities in understanding and solving more complex through intermediate reasoning steps using evidence and logical deduction to reach conclusions (Huang & Chang, 2023; Qiao et al., 2023). This enables LLMs to generate step-by-step solutions for complex problems. This approach also facilitates a more transparent problem-solving approach to refining the LLM's performance, as

instructions can be incorporated to guide the model through the reasoning steps more effectively.

4.3.3. Prompt Engineering

In the context of LLMs, a "prompt" refers to the input text provided to the model to elicit desired output (White et al., 2023). Depending on the desired task or response, prompts can range from simple questions or statements to more complex and structured instructions. A prompt is typically comprised of specific instructions, contextual information, input data, and output indicators, all formulated in natural language. Prompt engineering involves crafting such prompts to enhance performance in specific, unseen tasks. Clarity and specificity of the prompt are proven to be able to enhance the effectiveness of LLMs in generating accurate and contextually appropriate responses. Prompt engineering, or the skillful crafting of these prompts, is key to leveraging the full capabilities of LLMs for various applications like text generation, conversation, and translation. Figure 4-1 illustrates the difference between the prompting engineering approach and the fine-tuning method in addressing NLP tasks.

Despite the great success of LLMs, there remains a gap in the thorough exploration of prompt engineering for complex tasks—those that demand multi-step reasoning and sophisticated semantic understanding (Fu et al., 2022; Santu & Feng, 2023). Tasks like inferring information from contracts require a high-level, domain-specific understanding. This chapter offers a use case that provides practical insights for crafting effective prompts specifically tailored for complex tasks within the domain of legal texts.

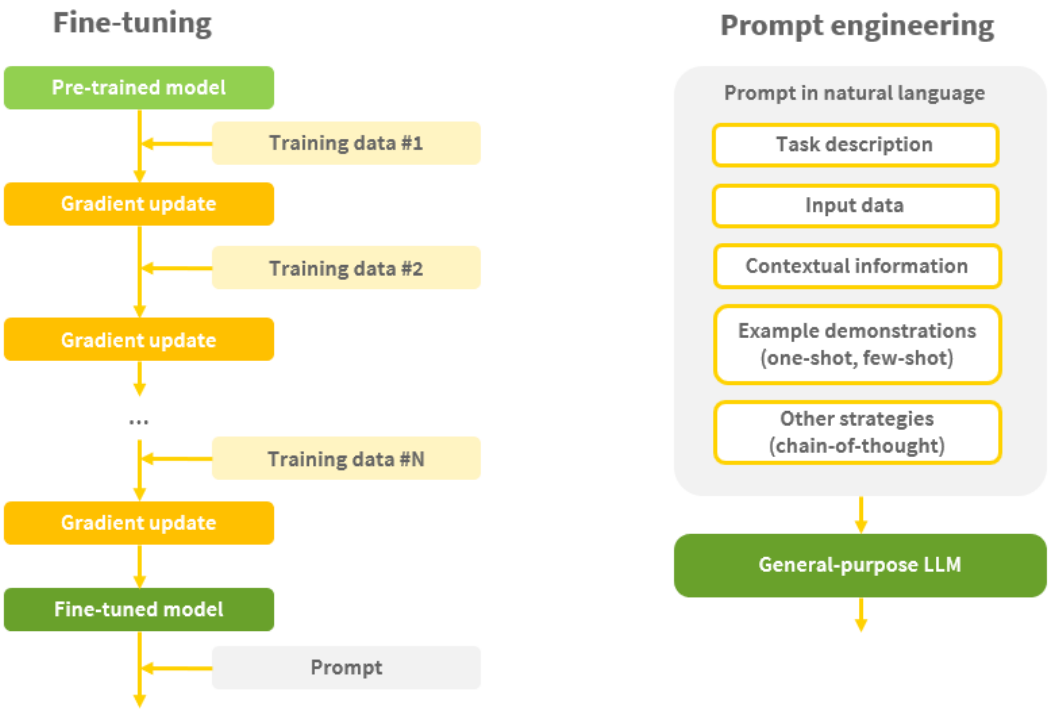


Figure 4-1 Contrasting approaches: Fine-tuning vs. prompt engineering

4.4. Research Methods

4.4.1. Prompting Strategies

This chapter conducts the following strategies for prompting engineering:

- **Few-shot prompting**. This is a primary strategy for interacting with LLMs (J. Wei, Tay, et al., 2022). This strategy involves providing the model with a small number of examples to guide its responses, which can be seen as a special case of in-context learning (P. Liu et al., 2023; W. X. Zhao et al., 2023). This enables a general-purpose LLM to understand and execute a wide range of downstream tasks without fine-tuning, therefore making LLMs more accessible and efficient for users to utilize these models to diverse applications. To evaluate the effectiveness of few-shot prompting, this study will undertake a comparative analysis that assesses the performance of LLMs under both few-shot and zero-shot prompting scenarios.

- **Chain-of-thought prompting**. This strategy utilizes LLMs' reasoning abilities to solve complex reasoning tasks using a step-by-step reasoning process (Dong et al., 2022). The model is guided to detail the logical steps leading to the output. This can help generate more accurate and interpretable responses. In assessing the efficacy of this strategy, this study will conduct a comparative analysis that assesses the performance of LLMs under scenarios with and without employing the chain-of-thought approach.

4.4.2. Data

The dataset consisting of sentences with risk type labels and risk element labels are the same as the dataset in Chapter 3.

4.4.3. Settings

The prompt engineering process is implemented on the GPT-3.5 Turbo model via OpenAI's Chat Completions API. Since these models update continuously, gpt-3.5-turbo-0125, the most up-to-date GPT-3.5 Turbo model as of March 10, 2024 (OpenAI, 2024b), were adopted for the implementation, as shown in Table 4-3. As mentioned in the OpenAI documentation, in simple tasks, GPT-4's improvement over GPT-3.5 is minor, but for complex reasoning, GPT-4 significantly outperforms earlier models.

Table 4-3 GPT models available in the OpenAI API

Model	**Family**	**Context window (tokens)**	**Training data up to**	**Feature**
gpt-4-turbo-preview (Currently points to	GPT-4 Turbo	128,000	Dec 2023	Latest GPT-4 model

gpt-4-0125-preview)				
gpt-4-vision-preview (Currently points to gpt-4-1106-vision-preview)	GPT-4 Turbo	128,000	Apr 2023	Multimodal—Can understand images
gpt-4 (Currently points to gpt-4-0613)	GPT-4	8,192	Sep 2021	
gpt-3.5-turbo (Currently points to gpt-3.5-turbo-0125)	GPT-3.5	16,385	Sep 2021	Latest GPT-3.5 Turbo model

In OpenAI's API, the chat completion object offers developers a suite of adjustable parameters to optimize performance:

- **temperature**: It influences the randomness of the output. Increase in temperature leads to more varied and creative responses, while decreasing it results in more predictable and deterministic outputs. For this specific task, temperature is set to 0 to ensure more consistent results.
- **top_p** (nucleus sampling): It controls the diversity of the responses by allowing the model to only consider the most likely next words within a certain probability range. The default value of 1 is maintained, as it is suggested against modifying this and the temperature parameter simultaneously (OpenAI, 2024a).
- **frequency_penalty**: Positive values will reduce the likelihood of the model repeating the content that already appears. The default value 0 is used.

- **presence_penalty**: Positive values will increase likelihood of the model to discusss new topics. The default value 0 is used.

4.5. Implementation and Results

4.5.1. Investigating Optimal Prompting Strategies for Enhanced Model Performance

This section systematically explores the optimal prompting techniques through the implementation of three prompt designs aimed at refining the efficiency and accuracy of model responses in contract risk classification tasks:

1. **Zero-shot Prompting (Baseline)**. This foundational approach employs a zero-shot prompt comprising a system prompt, task description, contextual information, output indicator, and the input data. It establishes a benchmark for subsequent strategy assessments. Table 4-4 showcases the format and elements of this basic prompt designed for the contract risk classification tasks, along with an example input and the resulting response in the stipulated format. In the illustrated example, the model successfully generates the response of correct classifications in the desired format.

2. **Few-Shot Prompting (Enhanced)**. Building on the zero-shot prompting, this approach introduces selected example sentences with manually annotated labels into the prompt to offer an enhanced strategy to improve model understanding and response accuracy. To examine the impact of the number of examples on performance, three experiments are conducted using 6, 12, and 18 examples, respectively. These 18 examples, derived from exploratory experiments, aim to rectify model deficiencies in task comprehension, particularly in distinguishing nuances of certain aspects such as obligations versus liabilities—a distinction readily apparent to humans

but sometimes challenging for language models. The demonstration format is presented in Table 4-4, with the demonstrations' details and their labels listed in Appendix A.

3. **Few-shot with Chain-of Thought Prompting (Advanced)**. This method extends the few-shot technique by integrating a chain-of-thought process to leverage the reasoning capabilities of LLMs for superior performance outcomes. This involves breaking down the analysis process into smaller, manageable sequential reasoning steps, which are added at the end of the task description, as presented in Table 4-4. Each step tackles a portion of the tasks by guiding the model through a structured analytical sequence to enhance contextual comprehension, pertinent information identification, and more accurate classifications.

Table 4-4 Prompt and response for contract risk classifications

Prompt element	**Role**	**Content**
		Zero-shot prompting (baseline) elements
System prompt (specifying a particular role for the model)	system	You are a legal expert specializing in construction contracts within the domain of transportation public-private partnership projects.
Task description (what the model is expected to do)	user	**Objective**: Analyze a sentence extracted from a contractual agreement of a transportation public-private partnership project. Your goal is to complete two distinct multi-label classification tasks on this sentence. Input Format: Your input is in the format of a JSON object containing a single key-value pair, where the key is 'sentence' and the value is the sentence you want to analyze. Here is an example of the expected input format: {"sentence": "Your sentence here."} **Tasks Description**: Task 1: Identify Risk Type Goal: Identify all types of risks mentioned in the sentence.

		Classes: There are 10 risk classes. Each class represents a different type of risk, ranging from economic and financial risks to no risk at all. Definitions of these classes are provided below. Task 2: Identify Allocation of Rights, Obligations, and Liabilities Goal: Determine the rights, obligations, and liabilities indicated in the sentence. The categories are defined below. Specifically, you need to identify whether they pertain to the public sector (typically referred to as "the Department" or the state DOT like "TxDOT", or a government-owned business like "HPTE" (High Performance Transportation Enterprise)) or the private sector (typically referred to as "the Developer" or "the Concessionaire"). Classes: This task involves 9 classes, including obligation, liability, right, and prohibition for both public and private sectors: i.e., "public obligation", "public liability", "public right", "public prohibition", "private obligation", "private liability", "private right", "private prohibition", plus a class "none" for sentences not mentioning any of these.
Contextual information (necessary background or setting relevant	user	**Risk Type Classes and Definitions**: 1. economic and financial: Risks related to uncertainties in market conditions, including inflation, interest rate or foreign exchange fluctuations, commodity price changes, etc., and risks related to securing and adjusting financing (e.g., failure of the PABs issuer or TIFIA lender), and refinancing. 2. socio-political and legal:…

to the task, such as key concepts)		… **Rights, Obligations, and Liabilities Definitions:** 1. obligation: Responsibilities and duties of a party involved in the contract, such as the provision of necessary materials and labor, obtaining required permits, timely completion of work or payment, adherence to specifications and standards, compliance with applicable laws and regulations, and mitigation of impacts in unforeseen events. 2. liability: … …
Output indicator (how the response should be structured or formatted)	user	Output Format: Your output should be a JSON object with two key/value pairs: risk_type and risk_allocation. Each key should be mapped to a list containing the identified classes based on the analysis of the sentence. Here is the expected structure for the output: {"risk_type": list_of_identified_risk_types, "risk_allocation": list_of_identified_rights_obligations_liabilities]} Example Output: {"risk_type": ["economic and financial", "performance"], "risk_allocation": ["public obligation", "private right"]} Important: Please strictly generate the output in the specified JSON format only, without including any additional information or commentary.

		Ensure your analysis is thorough, accurately identifying the relevant classes for each task based on the sentence provided in the input.
Input data (relevant data on which the model needs to generate response)	user	Below is the sentence to conduct the classification tasks: The Developer will be responsible for, and will bear the costs and schedule risk related to, coordination with WMATA_regarding the WMATA Work, the WMATA Easement Impacts and the WMATA Incidental Impacts.
Response	assistant	'risk_type': ['utilities, permits, and third-party'], 'risk_allocation': ['private obligation', 'private liability']
		Additional element for few-shot prompting
Demonstrations (using few-shot prompting strategies to provide the model with	user, assistant (as shown on the right column)	{'role': 'user', 'content': 'Here are a few example demonstrations:\n'}, {'role': 'user', 'content': "{'sentence': 'Subject to Section 10.04, if the Developer must submit a submittal or request to the Department for review and Response more than twice due to the Developer's failure to comply with the requirements of this Agreement, the Developer will pay the Department for the Department's Allocable Costs incurred thereafter in reviewing any portions of such submittal or request.'}"}, {'role': 'assistant',

human annotations)		'content': "{'risk_type': ['performance'], 'risk_allocation': ['private liability', 'public right']}"}, {'role': 'user', 'content': …, {'role': 'assistant', 'content': …}, …
		Additional element for few-shot with chain-of thought prompting
Chain-of Thought strategy	user	Read the sentence carefully: Start by understanding the content and context fully. Identify risk types: Analyze for phrases or keywords indicative of the 10 defined risk types. If no risk is explicitly mentioned or any of the miscellaneous risks defined in "other" is mentioned, consider it as "other". Determine allocation of rights, obligations, and liabilities: Search for terms specifying obligations, rights, liabilities, and prohibitions for the public or private sector. Combine findings: Reason through your findings by categorizing them into the appropriate classes. Formulate your output: Compile a JSON object with your classifications.

In the prompt, each element may consist of one or more JSON objects in the following format:

```
{
  'role': 'Insert the role here',
  'content': 'Insert your content here'
}
```

In this structure, the “role” refers to a specific directive that guides the model’s behavior or response style. It can represent different functions or perspectives within the interaction. These roles can direct the conversation to ensure that the model adopts specific expertise, or the output aligns with the intended tone. Typical roles include “system” for system-level commands or settings, “user” for simulating the end-user’s input, and “assistant” for simulating the model’s responses. The example below illustrates this concept:

```
{
  'role': 'system',
  'content': 'You are a scientist.'
},
{
  'role': 'user',
  'content': What is the process of photosynthesis?'
},
{
  'role': 'assistant',
  'content': 'Photosynthesis is the process through which plants, algae,
and some bacteria transform sunlight into chemical energy in glucose, usi
ng carbon dioxide (CO2) and water (H2O). Chlorophyll aids in converting t
```

```
hese inputs into glucose and oxygen (O2), vital for Earth's oxygen supply
 and as a foundation for most ecosystems' food chains.'
}
```

The “content” part of the JSON object specifies the actual input (e.g., a statement, a question, or a scenario) for the model to respond to. Part of the purpose of the prompt engineering experiment here is to craft the content in a way that leverages the assigned role’s expertise and ensures the desired responses. Finally, these elements form a complete prompt that is sent to the model for a response. By carefully designing the role and content, the aim is to tailor the interaction to achieve more specific, relevant, and contextually appropriate responses.

4.5.2. Prompt Engineering Results

In an exploration of zero-shot prompting techniques, multiple initial trials have indicated a lack of accuracy and stability of the outcomes. This inconsistency suggests that the complexity of the tasks may exceed the capacity of a simple prompt to effectively guide the generation of high-quality responses. The challenge appears to reside in the limitations of zero-shot prompting when applied to intricate tasks. To address this, it is imperative to enrich the prompting strategy by integrating example-based demonstrations following a few-shot learning approach. This method would potentially provide a more detailed context and clearer expectations, thereby enhancing the model’s ability to generate more accurate and reliable outputs. By leveraging examples as part of the input, the model can draw on concrete instances that illustrate the desired outcome, significantly improving its performance on complex tasks. This adaptation aims to refine the prompting mechanism, ensuring a more precise understanding of the task, thereby improving the quality of the generated responses to meet the demands of complex and specialized queries.

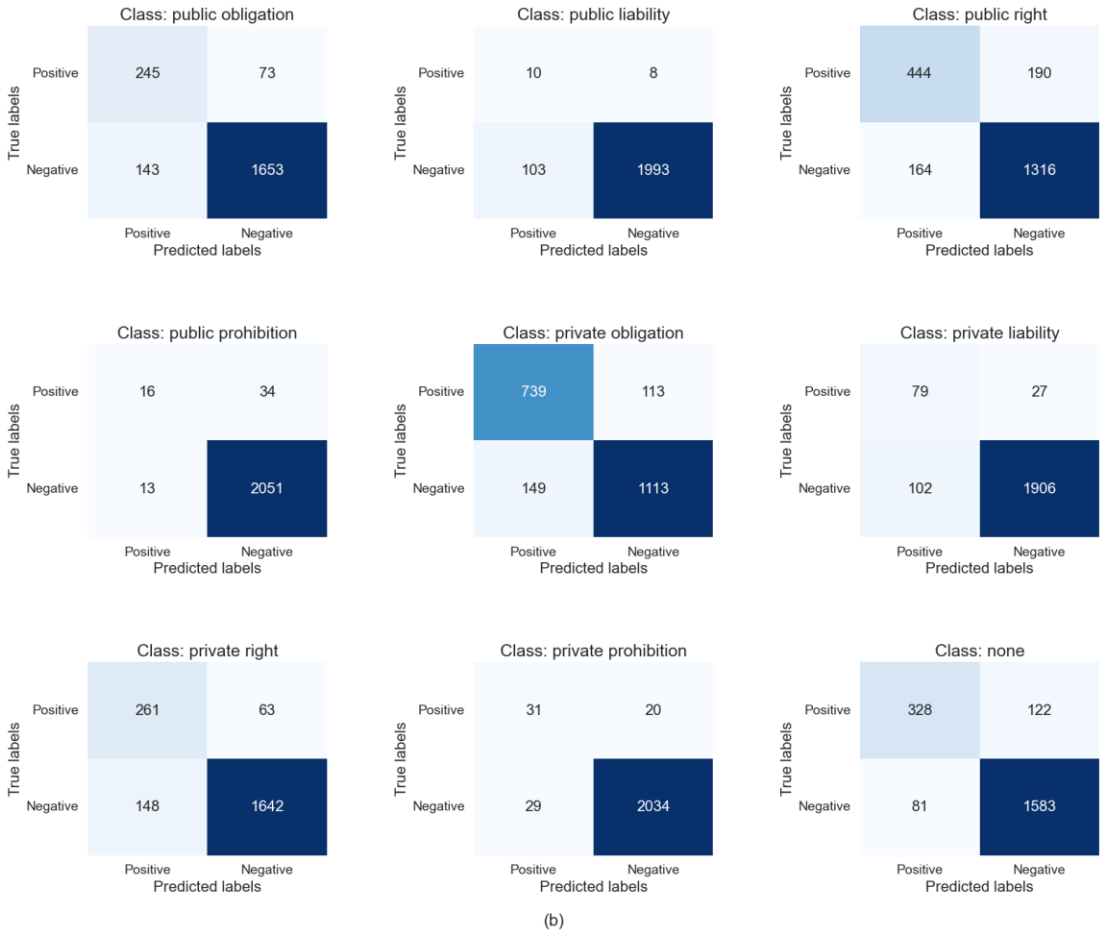
Table 4-5 presents the evaluation metrics to demonstrate the performance of each prompting strategy on the tasks. Overall, the performance of the few-shot approach and the few-shot approach augmented with chain-of-thought reasoning are very close, with the latter demonstrating a slight improvement. The zero-shot approach underperforms significantly. We also observe that zero-shot learning can exhibit high instability in responses across different experiments. The results suggest that in-context learning significantly outperforms zero-shot inference, which is consistent with most of previous findings (Min et al., 2022).

Table 4-5 Performance of prompting strategies

	Hamming Loss	Subset Accuracy	Weighted Precision	Weighted Recall	Weighted F1-Score
Risk type classification					
Zero-shot	0.123	0.259	0.484	0.308	0.332
6-shot	0.072	0.620	0.680	0.646	0.645
12-shot	0.070	0.631	0.698	0.652	0.649
18-shot	0.069	0.642	0.713	0.657	0.647
18-shot + Chain-of-thought	0.069	0.64	0.713	0.655	0.641
Risk allocation classification					
Zero-shot	0.187	0.276	0.543	0.296	0.314
6-shot	0.117	0.485	0.659	0.664	0.644
12-shot	0.092	0.572	0.730	0.707	0.706
18-shot	0.087	0.588	0.756	0.686	0.709

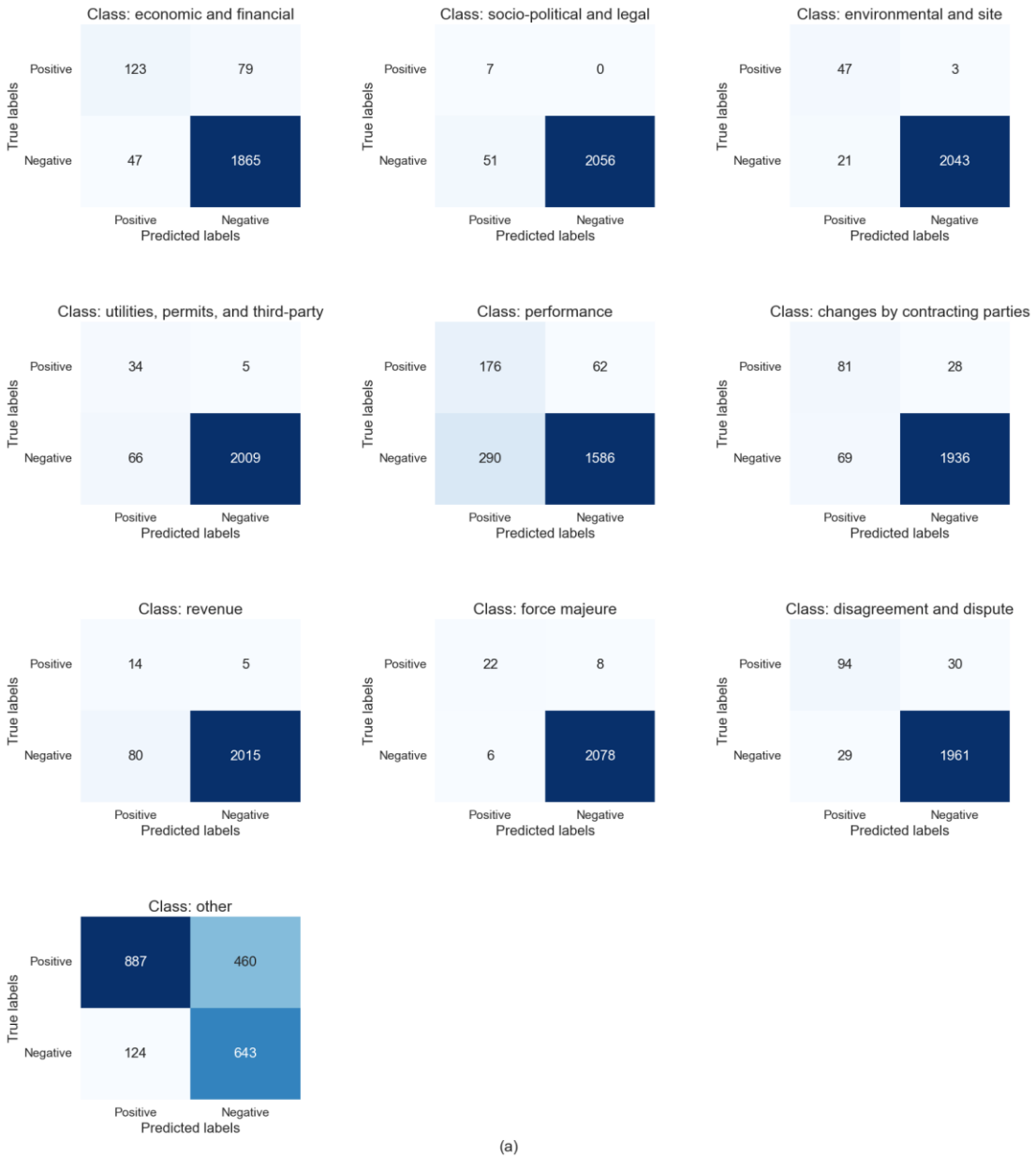
18-shot + Chain-of-thought	0.083	0.602	0.767	0.698	0.721

Based on the results generated by the few-shot approach augmented with chain-of-thought, confusion matrices corresponding to each class were calculated for both risk type and risk allocation classification, presented as heatmaps in



(b)

Figure 4-2.


Class: economic and financial
Positive
Negative
True labels
123
79
47
1865
Positive
Negative
Predicted labels
Class: socio-political and legal
Positive
Negative
True labels
7
0
51
2056
Positive
Negative
Predicted labels
Class: environmental and site
Positive
Negative
True labels
47
3
21
2043
Positive
Negative
Predicted labels
Class: utilities, permits, and third-party
Positive
Negative
True labels
34
5
66
2009
Positive
Negative
Predicted labels
Class: performance
Positive
Negative
True labels
176
62
290
1586
Positive
Negative
Predicted labels
Class: changes by contracting parties
Positive
Negative
True labels
81
28
69
1936
Positive
Negative
Predicted labels
Class: revenue
Positive
Negative
True labels
14
5
80
2015
Positive
Negative
Predicted labels
Class: force majeure
Positive
Negative
True labels
22
8
6
2078
Positive
Negative
Predicted labels
Class: disagreement and dispute
Positive
Negative
True labels
94
30
29
1961
Positive
Negative
Predicted labels
Class: other
Positive
Negative
True labels
887
460
124
643
Positive
Negative
Predicted labels

(a)

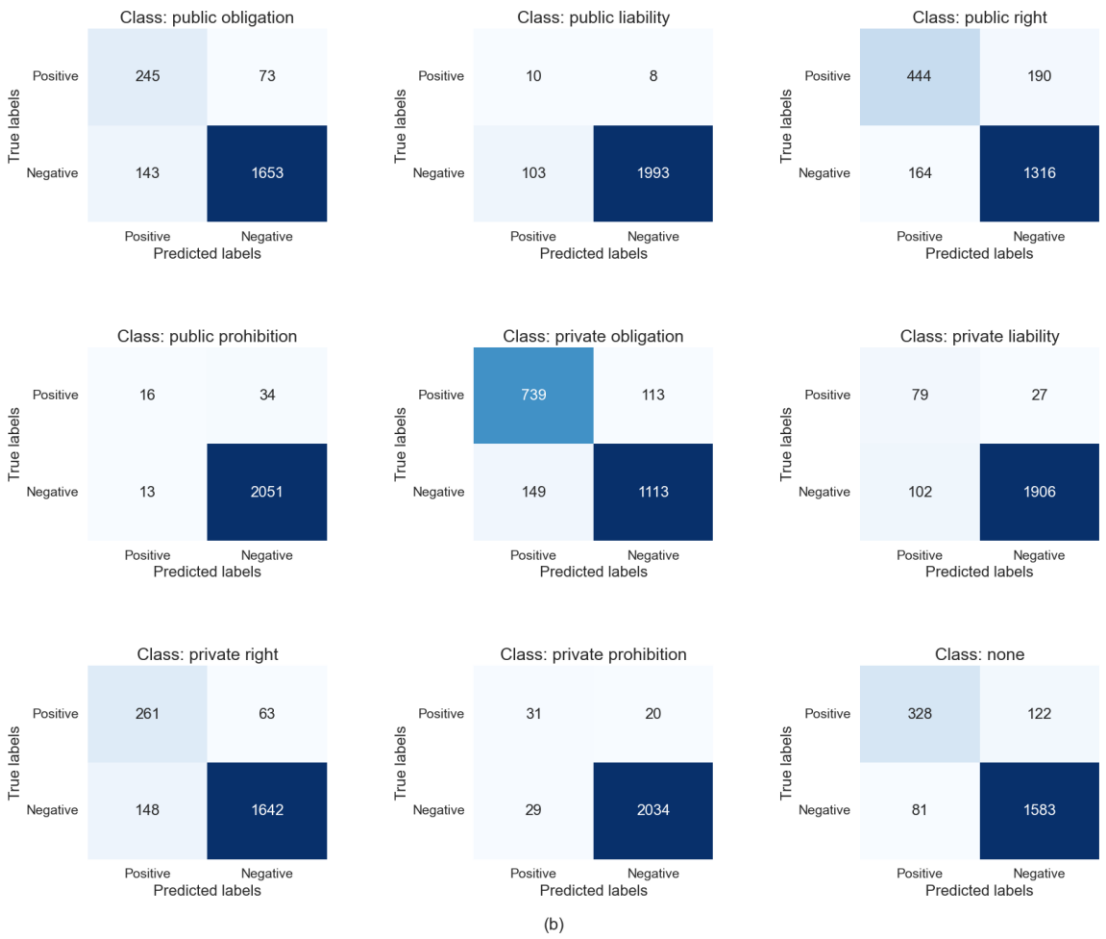


Figure 4-2 Heatmap of confusion matrices of classification results using the few-shot with chain-of thought prompting strategy. (a) Classification of risk type. (b) Classification of rights, obligations, and liabilities.

4.6. Conclusion

This chapter develops prompts using different strategies for the automation of risk identification using LLMs. It evaluates the effectiveness of LLMs in the specialized area of contract risk analysis. It also illustrates the development of targeted prompts, offering a practical example of strategic prompt engineering for specific domains.

Our comparative experiments of zero-shot and few-shot in-context learning reveal that zero-shot approaches can often generate inaccurate responses. This may be due to the complexity involving nuanced definitions and concepts, where words can fall short when explaining the underlying logic to an LLM through mere descriptions and instructions. In contrast, using prompt completion pairs can better impart the necessary intuition. By introducing a few examples, few-shot learning significantly enhances the quality of responses. This improvement suggests that, despite being trained on extensive datasets, LLMs can struggle with complex task description or when handling terminology that carries different meanings across various training sources. Few-shot in-context learning extends LLMs' capabilities to specific domains or tasks by improving their contextual understanding, thereby increasing accuracy in niche areas without the need for additional training. This approach leverages minimal examples to guide the model to enhance its precision in handling domain-specific terms.

Chapter 5. Comparison Analysis of Language Models for Contract Risk Profiling

This research develops multiple tools for automated contract risk identification using different approaches, as illustrated in Figure 5-1. This chapter first presents a comprehensive evaluation and comparison of the NLP approaches in Chapter 2 to 4 that are tailored for the identification of risks from P3 contracts. Through a detailed parallel comparison, this chapter dissects the performance, efficiency, and applicability of the selected approaches for analyzing contract documents. Following the model evaluation, the chapter explores the practical application of these NLP techniques on a corpus of real P3 contracts. By employing the top-performing models, we extract the risk-related content of these contracts and systematically analyze their risk profiles. Finally, this chapter discusses the risk profiles in relation to the findings of the existing literature.

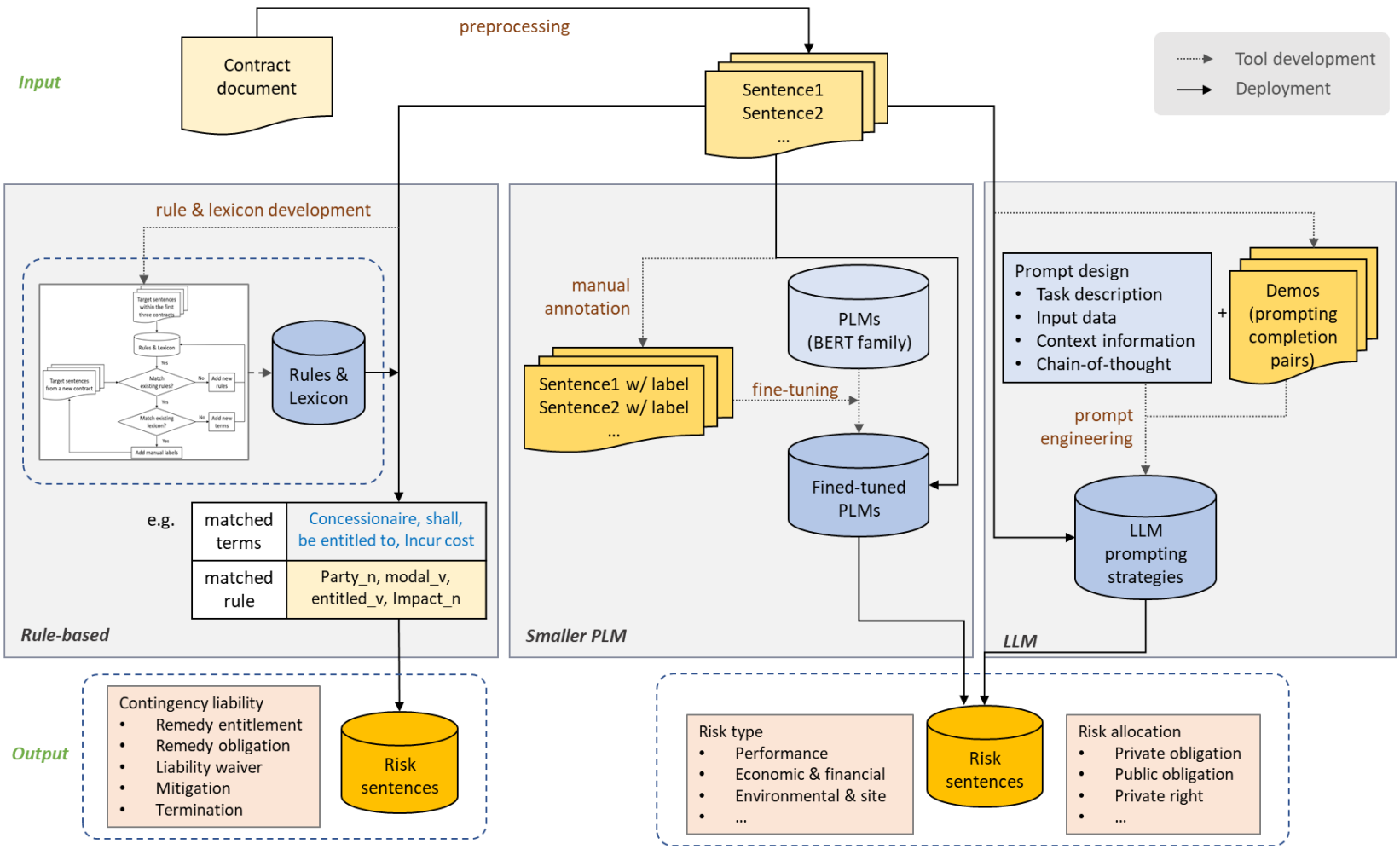


Figure 5-1 Overview of the NLP tools for automated contract risk identification

5.1. Evaluation and Comparison of NLP Approaches for Risk Identification in P3 Contracts

5.1.1. Comparison of NLP approaches

This dissertation work develops risk-related classification schemes in P3 contracts and evaluates various NLP approaches in performing the classification tasks. Through an array of approaches—rule-based models, fine-tuning PLMs, and prompt engineering with LLMs—this research aims to advance the automated analysis of contract documents. Based on the model development process and experiment results, each approach has its own advantages and limitations, as summarized in Table 5-1.

Table 5-1 Comparison of three NLP approaches in current study

	Pros	**Cons**
Rule-based	• Highly interpretable • Ease to customize rules and offer precise control • Generating consistent results	• Development and maintenance demand significant effort and expertise • Limited scalability • Reliance on predefined patterns can miss unknown terms
Fine-tuning PLMs	• High accuracy • Contextual understanding • Scalability for processing potentially large volumes of contracts	• Depending on potentially large, annotated datasets • Interpretability issues due to the models' "black box" nature

	• Open source	
Prompt engineering with LLMs	• Flexibility across a wide range of tasks • User-friendly input in natural language • Rapid deployment; No training required	• Usage can be costly for large datasets; resource-intensive even if open-source • Privacy and security concerns with sensitive data • Limited customization for highly specialized tasks

Table 5-2 further compares the two approaches—fine-tuning PLMs and prompt engineering with LLMs—applied to the same classification tasks.

Table 5-2 Comparison of fine-tuning PLMs and prompt engineering with LLMs

	Fine-tuning PLMs	Prompt Engineering with LLMs
Models used	BERT family	GPT-3.5
Open source	Yes	No
Training required	Yes	No
Prompt design require	No	Yes
Weighted F1-score of the optimal model on two classification tasks	0.806/0.805	0.641/0.721

As demonstrated by the weighted F1-score, the fine-tuned PLM outperforms prompt engineering with LLMs on the classifications. This model achieves results comparable to those multilabel or multitask classification tasks of contract documents in previous studies, such as a 0.75 macro F1-score in Candaş and Tokdemir (2022) and a 0.89 weighted F1-score in Pham and Han (2023). One reason for this performance difference is the substantial length of the example sentences and the need for multiple examples to help the model differentiate between various classes. Moreover, the detailed definitions of risk categories contribute to an even lengthy input context, often extending between 1600 to 2000 tokens. As observed by existing studies, there is a variation or even a decline in model performance as the input context lengthens (Fatemi & Hu, 2023; N. F. Liu et al., 2023). This finding may explain why prompt engineering may not always yield superior performance, particularly when the input text becomes excessively long.

5.1.2. Analysis and Discussion of Fine-Tuning PLMs

Regarding the performance of PLMs, we have tested four models from the BERT family: BERT-base, RoBERTa-base, LegalBERT-base, and XLM-RoBERTa-base. RoBERTa-base demonstrated the most superior performance among the selected models. Despite the large number of parameters in XLM-RoBERTa-base, its multilingual capability does not align with the objectives of this study. Instead, the excessive parameter count complicates the training process, resulting in diminished performance relative to RoBERTa-base. This indicates that the selection of a model is not merely a matter of size. It is contingent upon a case-by-case evaluation where the most appropriate model is chosen in accordance with the unique attributes of the task and the dataset.

Regarding the dataset, we have compared the performance of the selected model on a vanilla dataset containing sentences and an augmented dataset containing both sentences and the titles of the article and section. The model demonstrates superior performance on the augmented dataset. This outcome highlights the importance of context in understanding and analyzing contract sentences. It suggests that models equipped with a broader perspective of the document structure can more accurately identify and classify risk elements. This further emphasizes the role of context information in advancing model accuracy and reliability in practical applications.

5.1.3. Analysis and Discussion of Prompt Engineering with LLMs

Comparing with the fine-tuning approach, a significant advantage of prompt engineering with LLMs is that they do not require additional training on domain-specific data, or only require a small amount of examples to guide the model to generate desired response. In addition, LLMs can not only deliver classification results, but also can provide detailed, step-by-step explanations on the process by modifying the output indicator. This ability is particularly useful for contract analysis since contract language often contains implicit content and complex relationships.

Through the experiments of different prompting strategies, we find that zero-shot has a poor performance, which can be significantly improved using few-shot and chain-of thought. This finding is in line with previous studies (e.g., Fatemi & Hu, 2023; Manikandan et al., 2023). Although a lot of studies demonstrate the remarkable zero-shot capabilities of LLMs (e.g., Kojima et al., 2022; Sainz et al., 2022; Shen et al., 2023; X. Wei et al., 2023), the majority of these experiments utilized shorter prompts and focused on more generalized use cases such as named entity recognition or logical reasoning involving arithmetic and commonsense reasoning.

However, there are a few things to keep in mind when using this method. A big issue with LLMs is hallucination, which makes the model generate grammatically correct but factually incorrect or irrelevant response. The model may give its own interpretation despite a detailed definition, description, and examples given in the prompt. This study has also encountered such problems. For example, the model may generate new labels that do not align with the predefined classes and output formats, such as producing “public none” for risk allocation classification. Yet, these errors constitute only less than 1% of the generated results.

This tendency could be because LLMs are trained on a large corpus. This empowers them with their own “thoughts” or “perspectives”, which may not always be adjusted using a prompt. For example, in our classification tasks, the model tends to classify “revenue” risks as “economic and financial” for tolling related provisions. Although tolling is highly related to financing, its direct relation to revenue suggests that it is more appropriate classified as a revenue risk.

Other challenges include unstable outputs due to the generative nature of AI, which gives different interpretations each time even when a low temperature setting is used to facilitate consistency. Therefore, a future direction involves enhancing result consistency. Additionally, the model’s performance may degrade when processing long sentences, as it can struggle to extract all pertinent information from lengthy input texts.

Finally, it is important to acknowledge that, while LLMs possess impressive capabilities, their complex nature also brings challenges concerning transparency and interpretability (H. Zhao et al., 2024). These challenges highlight the importance of carefully adapting LLMs to the unique characteristics and needs of downstream tasks. The continuous advancement in the rapidly evolving field of LLMs suggests that we are on the cusp of a major revolution in discovering innovative applications and methodologies for their use. As LLMs become increasingly

integrated into various sectors, it is crucial to develop strategies that leverage the strengths of LLMs more effectively and ensure they serve as powerful tools tailored to particular challenges and opportunities within specific domains.

5.2. Practical Application of the Model on Real P3 Contracts

Based on the classification result of Section 0, this section examines the proportion of contract language pertaining to each risk type and how these risks are allocated between the public and private sectors. These results provide insights into the risk management strategies employed in each project.

Overall, the most described risk types identified in P3 contracts include “performance” (18.8%), “disagreement and dispute” (8.4%), “economic and financial” (7.3%), and “changes by contracting parties” (6.5%), as indicated in Table 5-3. In terms of risk allocation, these contracts frequently assign a greater portion of obligations to the private sector (“private obligation”, 35.1%) and more rights to the public sector (“public right”, 27.2%), followed by “private right” (17.3%) and “public obligation” (16.7%). The result also shows that contract language related to liability and prohibition predominantly concerns the private sector, exceeding those pertaining to the public. This pattern aligns with the logical expectation that P3 contracts, typically drafted by the public sector, would favor its interests.

Table 5-3 Distribution of risk-related sentences in the P3 contracts

Classification 1: Risk type	Volume	Proportion	*Classification 2: Risk allocation*	Volume	Proportion
Other	10539	51.4%	Private obligation	7202	35.1%
Performance	3844	18.8%	Public right	5571	27.2%
Disagreement and dispute	1724	8.4%	None	5130	25.0%
Economic and financial	1492	7.3%	Private right	3550	17.3%
Changes by contracting parties	1336	6.5%	Public obligation	3427	16.7%
Utilities, permits, and third-party	958	4.7%	Private liability	1547	7.5%
Environmental and site	575	2.8%	Public liability	953	4.7%
Revenue	508	2.5%	Private prohibition	457	2.2%
Socio-political and legal	479	2.3%	Public prohibition	197	1.0%
Force majeure	110	0.5%			

In addition, comparative analysis is conducted in terms of jurisdictions and payment methods. The distributions of sentences pertaining to risk type and risk allocation are shown in Figure 5-2 and Figure 5-3, respectively. The y-axis "volume" refers to the total number of sentences assigned to a class per contract in a subset, while "proportion" represents the total number of sentences assigned to a class divided by the total number of sentences.

As shown in Figure 5-2, although the total sentence volume varies by project, "performance" remains the most described risk type in all contracts. "Economic and financial", "disagreement and dispute", and "changes by contracting parties" are among the top ranked risks. The distribution of risk types significantly varies by the payment method. "Performance", "environmental", and "changes by contracting parties" risks are more frequently discussed in AP concessions and tolled concessions compared to lease projects. The proportion of "revenue" risk is higher in lease and tolled concessions compared to AP concessions.

In terms of risk allocation, the ratio of each category of risk allocation language are relatively consistent across different projects, as illustrated in Figure 5-3. Again, "private obligation" and "public right" receive much more attention than others. Among the three payment methods, toll concessions describe "private obligations" most.

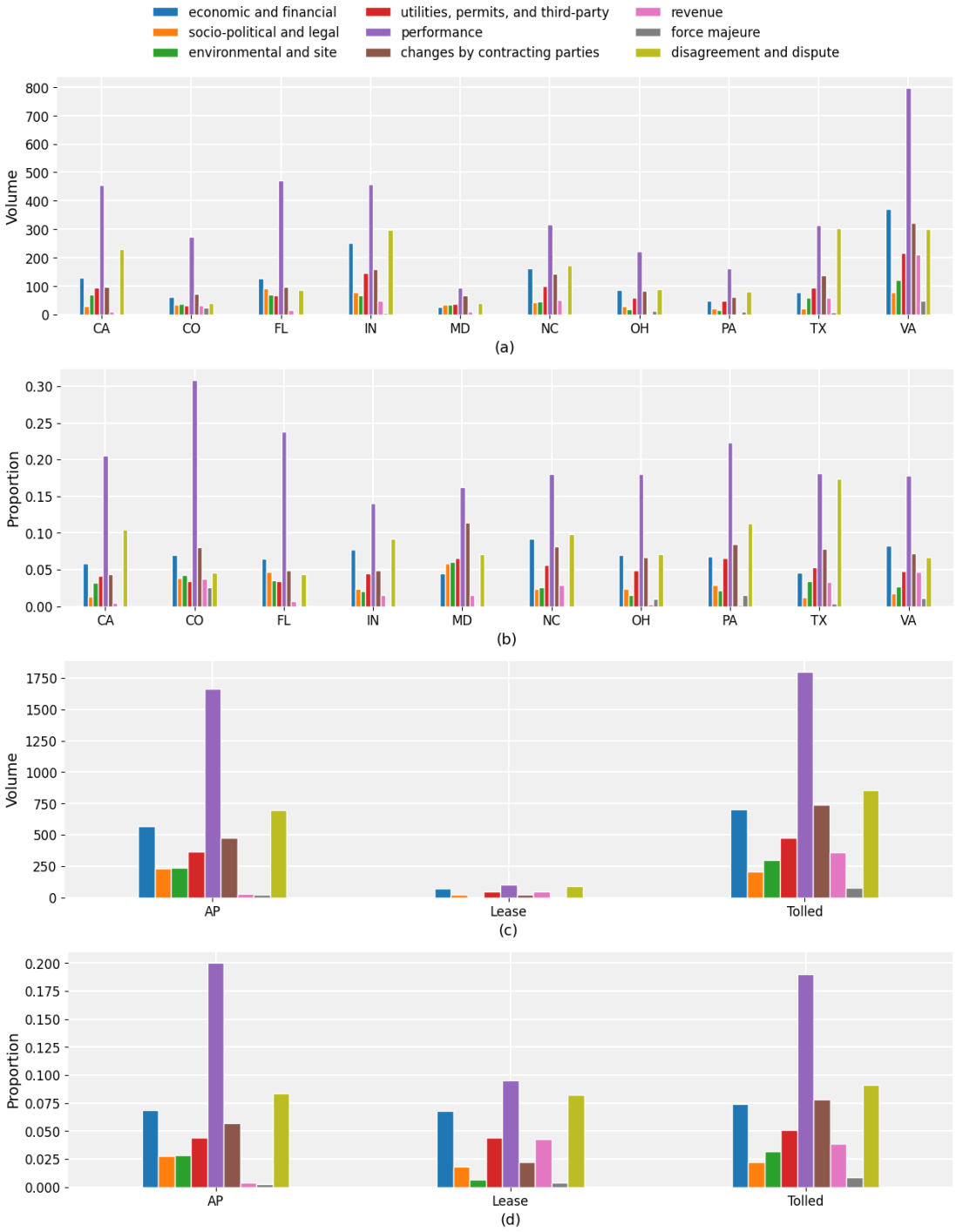


Figure 5-2 Distribution of risk type-related sentences in the real P3 contracts. (a) Sentence volume by risk type across different jurisdictions. (b) Proportional distribution of risk types

across different jurisdictions. (c) Sentence volume by risk type based on payment method. (d) Proportional distribution of risk types based on payment method.

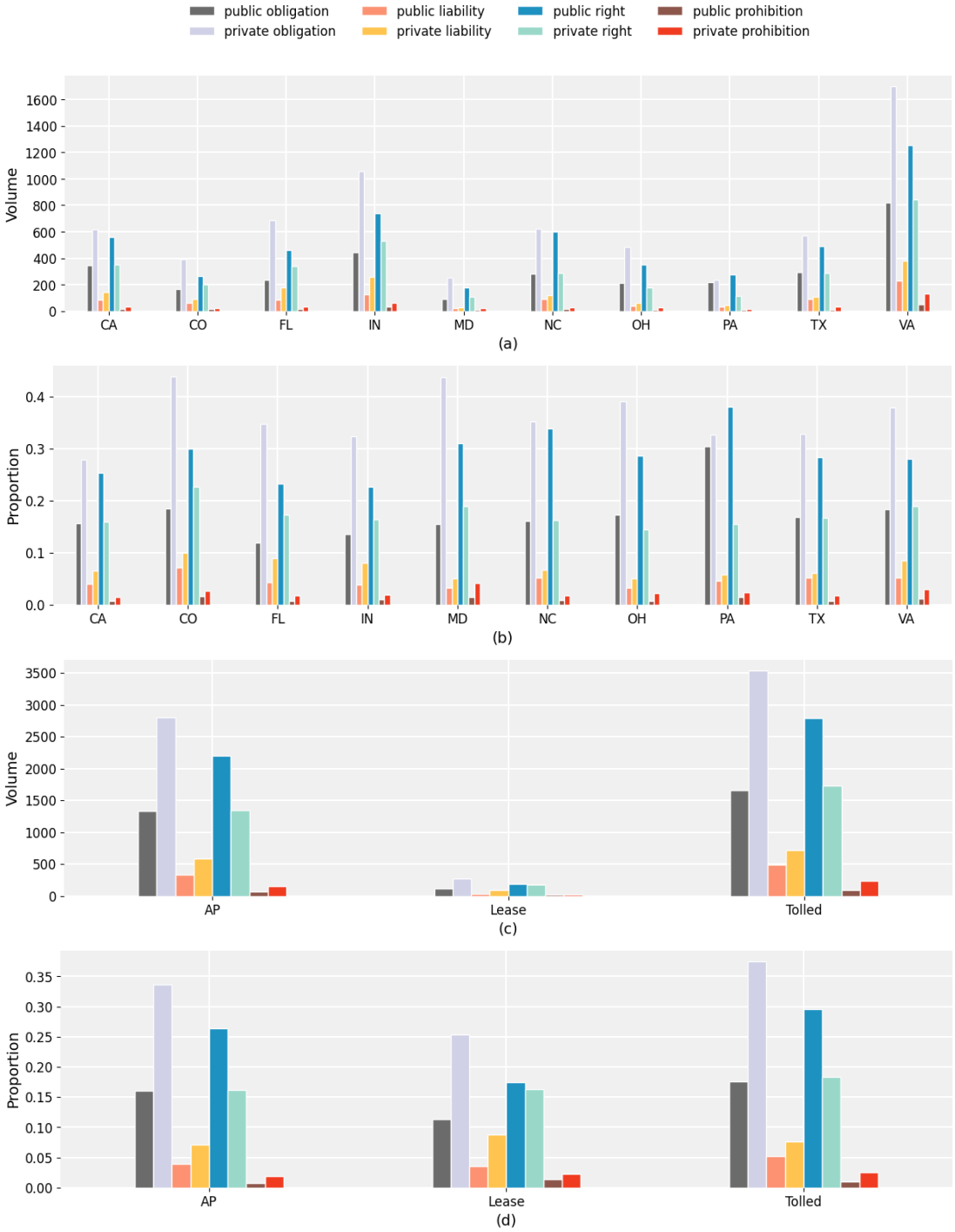


Figure 5-3 Distribution of risk allocation-related sentences in the real P3 contracts. (a) Sentence count by risk allocation across different jurisdictions. (b) Proportional distribution of risk allocation across different jurisdictions. (c) Sentence count by risk allocation based on payment method. (d) Proportional distribution of risk allocation based on payment method.

Figure 5-4 shows the distribution of risk allocation for each risk type. For the first six risks: "economic and financial", "socio-political and legal", "environmental and site", "utilities, permits, and third-party", "performance", and "changes by contracting parties", the private party bears much more obligation than the public partner. The public partner is entitled to more rights. For these risks, the contract tends to transfer more risks to the private party.

In comparison, for "revenue", "force majeure", and "disagreement and dispute" risks, the sentence volume about the two parties' obligations and right are much closer. The contract tends to use a shared risk mechanism for these risks.

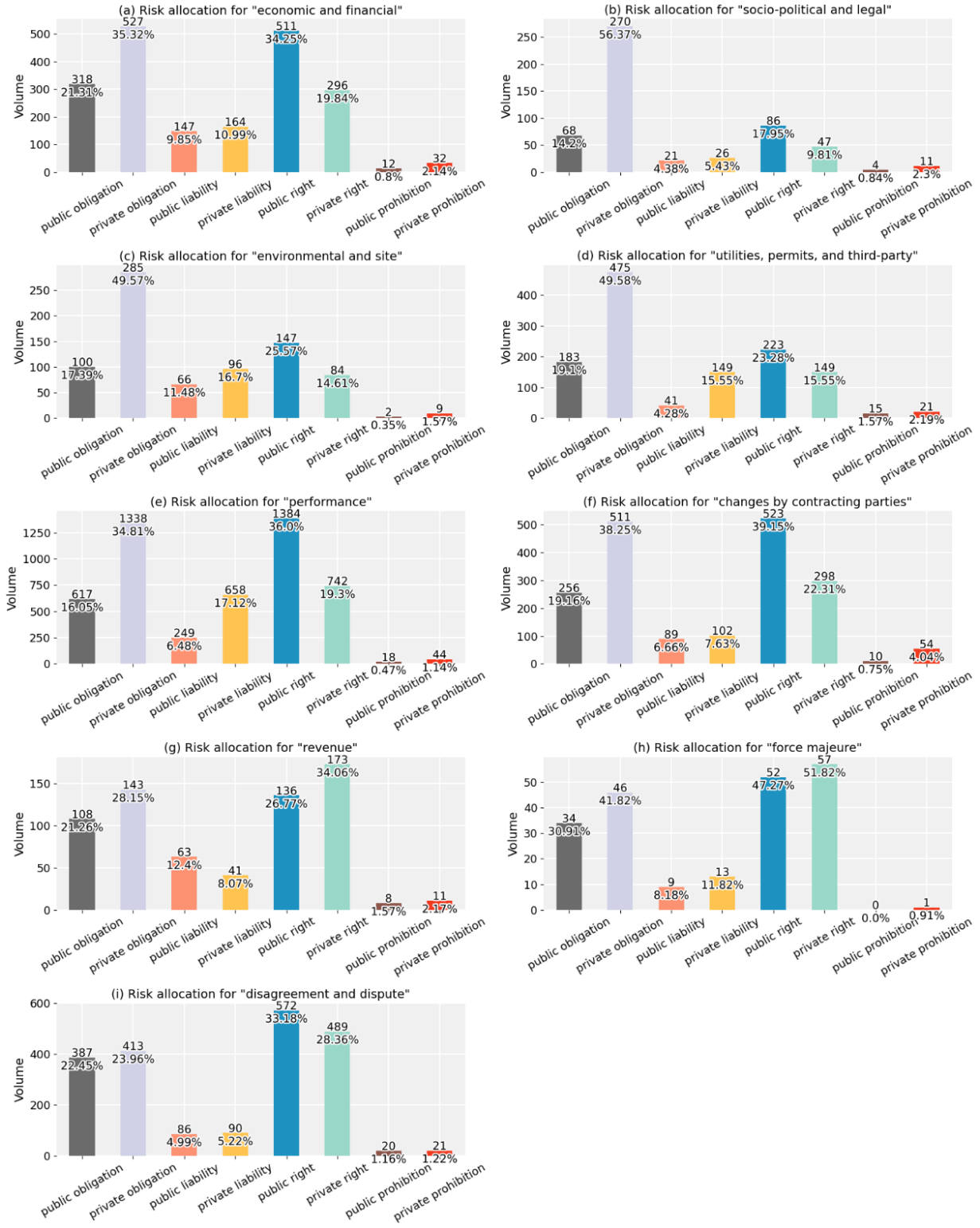


Figure 5-4 Distribution of risk allocation-related sentences for each risk type

The automated information extraction of multiple contracts simultaneously implies the potential of promoting contract standardization in the P3 industry. With the widely recognition that contracts can have significant impact on project success (William & Ashley, 1987), high transaction costs and the complexity of contracts have been barriers to quick P3 agreement settlements (van den Hurk & Verhoest, 2017). Interviews with industry professionals also emphasize the challenges of contract preparation, such as ensuring regulatory compliance, managing risks, ensuring project delivery, addressing changes, and negotiating terms. In contrast to the European countries where their documents are simple and standard as they do not run into some of the complexity in the U.S., where P3 legislation varies across the states, resulting in different contracts. As the whole P3 industry could benefit from a set of standard agreements to start with, there is an effort to try to standardize the process, aiming to drive the market toward a more efficient contracting process. By comparing P3 contracts across different states, the findings of this study can serve as guidance to aid project sponsors in designing more effective contracts with fair and balanced risk allocation, moving towards standardization that can lower costs and simplify negotiations. Although difference exist across projects, and standard contracts only "serves as a starting point for the negotiation process between the principal" (van den Hurk & Verhoest, 2017), the tools delivered in this research align with efforts to standardize the contracts with more fair and balanced risk allocation, potentially making procurement more efficient and accessible.

5.3. Comparison of Risk Profiles with Existing Literature Insights

This subsection compares the analyzed risk profiles with insights derived from existing literature. It discusses how the findings from the application of NLP techniques to P3 contracts

relate to, complement, or diverge from risk profiles in standard contract forms or the preferred risk allocation in P3s gained through surveys in previous studies.

In general, risk allocation follows the basic criteria and principle that risks should be allocated to the party that is in a better position to anticipate, assess, control, bear the consequences, and manage the risk (Iossa et al., 2007; Lam et al., 2007). However, there is no uniform standard for determining the specific allocation of each risk. Previous research mainly employed interviews, surveys, case studies, manual content analysis to obtain data about risk allocation in real projects or contract standards. These studies have investigated a variety of markets such as European, Australia, China, and India, while the U.S. P3 market has been under-discussed. An overview of risk allocation in current studies for P3 and highways projects is summarized from a comprehensive literature review, as shown in Table 5-4. For example, Ke et al., (2010) investigated the preferred risk allocation in multiple P3 markets using questionnaire surveys. Their findings suggest that public sector tends to transfer more risks to the private sector in the U.K., compared with China mainland, Hong Kong, and Greece. Our classification results offer a comparison of the U.S. transportation P3 market with these earlier studies.

Based on our classification result, there are more descriptions about the obligation and liability and less description about the right of the private partner compared to those of the public partner, indicating that the private partner bears more risks. The overall ratio of public obligation to private obligation is 44.1%. This finding is in consistent with Agrawal et al. (2021). Using a rule-based NLP model, they examined the ratio of obligation or right allocation in the general conditions of contract in India. They found that the number of tokens of client obligations are less than that of contractor obligations across different contract forms, with a ratio ranging from 0.32 to 0.66.

Table 5-4 Summary of risk allocation in P3 and highway projects

Authors	Method	Project type and targeted market	Risks mainly borne by public (owner)	Risks mainly borne by private (contractor)	Risks typically shared	Undecided/ Controversial
M.-t. Wang & chou (2003)	Case study	Transportation (highway). Taiwan	Changed orders (from legislative changes, accommodating public works, political pressure, owner changes), increased costs to crash activity time	Property losses resulting from natural disasters, delay or cost overrun owing to site discrepancy/lack of information/errors/unqualified work, public claims for arising from construction, faulty design not detected by contractor		Change in law, inflation, delay caused by faulty design/inefficient owner's supervisors, suspension of work before weather events
Bing et al. (2005)	Survey	General. UK	Political opposition, land acquisition, unstable government	Tax regulation change , late design, staff crises, third party tort liability, economic events, financing risk, demand risk, industrial regulation change,	Force majeure, legislation change	Public opposition, approvals and permits, excessive

				inflation/interest rate volatility, organization and co-ordination risk, weather, environment, geotechnical conditions, cost overrun, availability of labor/material, design deficiency, construction delay		contract variation,
Abedneg o & Ogunlana (2006)	Case study	Transportation (tollway). Indonesia		Delay and increased cost resulting from using preliminary design, unpredicted project site condition, construction quality, legal and contractual risks, financial risks, force majeure	Delayed project construction initiation, contractor's failure	
Lam et al. (2007)	Prototypic quantitative model		Conflicts in document, third-party delay, ground condition, rights of access to site, public disorder	Design, subcontractor's failure, quality, safety, obtaining approval or consent, inflation, labor and equipment availability, inclement weather, labor dispute and strikes	Change in law	
Iossa et	Content	DBFO	Statutory/planning risk,	Design, inefficient construction	Demand, exchange	

al. (2007)	analysis	contracts. General	misspecification of output requirements, adverse environmental conditions, delays in obtaining approvals, changes in output specifications/public needs, residual value	practices, inadequate cost management, increases in materials and labor costs, operation, legislative/regulatory	and interest rate fluctuations, financing cost increases
		Lease contracts. General	Misspecification of output requirements, changes in output specifications/public needs, discriminatory or specific changes in law, exchange and interest rate fluctuations, financing cost increases, residual value	Operation, demand, general changes in law	
J. W. Brown et al. (2009)	Interview and expert discussio	Transportation (highway). Portugal		Design, land acquisition, construction, O&M, latent defects, change in law, competing facilities	Environmental compliance, market/demand, force majeure,

	ns	Transportation (highway). Spain	Land acquisition, change in law, force majeure	Design, environmental compliance, construction, O&M, latent defects	Market/demand, competing facilities	
		Transportation (highway). United kingdom	Land acquisition, force majeure,	Design, environmental compliance, construction, O&M, market/demand, change in law, competing facilities	Latent defects	
		Transportation (highway). Australia	Land acquisition, change in law, force majeure	Design, environmental compliance, construction, O&M, market/demand, latent defects, competing facilities		
Ke et al. (2010)	Survey	General. Mainland China	Political, legislation change, public opposition, approvals and permits	Industrial regulation change, environment, interest rate volatility, geotechnical conditions, weather, staff crises	Relationship risks, third-party risks, force majeure, excessive contract variation, financial market risks,	Change in tax regulation, land acquisition, late design change, demand risk, inflation

					economic events	
		General. Hong Kong	Legislation change, political opposition, land acquisition, change in tax regulation	Demand, environment, two design risks, financial market risks, economic event, availability of finance, labor/material availability, delay, cost overrun	Relationship risks, force majeure, excessive contract variation, weather, staff crises, excessive contract variation, public opposition, residual risk, geotechnical conditions, inflation	Approvals and permits, late design changes, industrial regulation change, third party, influential economic
Heravi & Hajihosseini, (2012)	Case study	Transportation (highway). Iran	Limited capital, market risk, land acquisitions, O&M cost overrun,	Termination by government, economic events, change in law, land appraisal, environmental approval, improper design, operator default	Inflation, construction cost overrun, severe weather, war, natural disasters	
Nguyen et al.	Content analysis	Transportation	Usage-demand and network risks in availability payment	Usage-demand risks in tolled projects, financing, design,	Permits, environmental	Interest rate, inflation, right

(2018)		(highway). U.S.	projects, changes by the public authority. .	performance, availability, operating expenses, maintenance, transfer, residual value	risks, archaeology fossils and protected species, site geology, project company default and termination by public authority, force majeure, social-political opposition, change in law	of way, access/adjustment to utilities
Castelblanco et al. (2020)	Content analysis and case study	Transportation (tollway). Chile	Political opposition, change in law, disposal of surplus land, demand risk, termination by the contracting authority.	Financing, Planning and permits, ground condition, archeology and fossils, access, rights, and easements, price adjustments, design, insurable risks, delay by construction subcontractor, performance, availability and service, O&M, Transfer, residual	Inflation, site acquisition, connections to the site,	Interest rates, foreign exchange rate, protesters, force majeure

value

Chapter 6. Conclusions

6.1. Contributions

The major contribution of this research is the development of tools using different NLP approaches, including a rule-based model, a fine-tuned BERT model, and prompt designs. These tools offer an automated process for industry practitioners to leverage NLP for rapid and consistent contract risk identification. These extracted pieces of information could facilitate a quick response in risk-handling.

Besides, the study enriches the risk management literature by developing a risk framework customized for contract language about project risks. This serves as a basis of the current research and fills the gap in the literature by providing an analytical tool for detailed contract analysis.

Additionally, regarding language models, this research provides evidence using domain-specific data and tasks for the evaluation of the efficacy of different models. Particularly, a comparative analysis is conducted between the performance of traditional PLMs and the more advanced LLMs. PLMs had their highlights between 2018 to 2021. Despite the perceived obsolescence of earlier PLMs in the face of the rapid development of LLMs recently, the provided in this study challenges this view. It demonstrates that BERT, despite its significantly smaller parameter size and training dataset compared to GPT-3.5, does not universally underperform in all tasks. As smaller models are often faster and less expensive to train and use than larger models, there are trade-offs when deciding on the most suitable model type and size for each specific application.

This exploration serves as a case study for selecting and implementing the most suitable NLP models for tasks that require nuanced understanding and categorization of complex text data in

the legal domain. The effective performance of BERT family of models in this context can be attributed to the standardized nature of contractual language, which often follows standardized terminology and encompasses a smaller vocabulary relative to other datasets. This specificity reduces the necessity for massive training data, allowing BERT to achieve competitive results, as evidenced by achieving a weighted F1-score of 80.6% and 80.5% compared to 64.1% and 72.1% achieved by prompt engineering with GPT-3.5 in this study.

Finally, the research summarizes contract risk profiles using the model developed. Based on sentence-level risk identification from P3 contracts, we offer valuable perspectives for depicting current practices in risk definition and allocation in U.S. transportation P3s. This provides guidance for practitioners to understand and compare contract stipulation of risks across different states and payment methods at a fine-grained level.

6.2. Limitations and Future Research Opportunities

6.2.1. Limitations

Despite the tools developed and insights provided by this research, there are several limitations. These limitations come from data availability, methods choices, and the challenges associated with interpreting complex contract language.

Firstly, echoing the point of view in previous research, the limited availability of textual data in the construction field poses a significant challenge to the application of state-of-the-art NLP methods (Baek et al., 2021; S. Chung et al., 2023). This study faces similar obstacles. The dataset of contracts is limited, and the annotation process demands substantial manual effort. Given the limited research resources, the size of the current dataset is insufficient to fully exploit the potential of these advanced language models. Moreover, the focus of this study is exclusively

on P3 contracts within the U.S., limiting its applicability to other contract types without additional work.

Secondly, due to limited computational resources, only a select number of PLMs and LLMs were employed in this research. While the feasibility of the methods has been validated, there remains significant room for improvement and expansion in future studies. The rapid evolution of AI technologies promises the emergence of superior models and methodologies, which can offer more opportunities for advancing this line of research.

Thirdly, it is acknowledged that the risk framework introduced by this research, while customized for contract risk analysis, still simplifies the rich and precise meanings of contract language to a certain degree. While this simplification facilitates a quicker and more manageable analysis, it also introduces the concerns of overlooking subtler pieces of information. This necessitates a more sophisticated approach to capture a whole variety of information in contract documents. For example, regarding risk type identification, consider the sentence "Performance of any of the Work by a Contractor will satisfy the obligation of the Developer to perform such Work; provided that any such Work performed will be binding on the Developer and the foregoing will not relieve the obligation of the Developer to manage such Contractor." While this sentence does not explicitly mention performance risk, it implicitly addresses aspects related to performance risks. As the Developer is responsible for the Contractor's performance, such risks are relevant and need to be managed by the Developer. Regarding risk allocation, while the current models can identify the responsibilities associated with risk borne by a party, they have not been trained to extract information regarding the proportionality and extent of such risk allocations. Such simplification might mask the nuanced aspects of risk allocation. This

highlights the ongoing need for a more sophisticated approach to capture the more details conveyed in contract documents.

6.2.2. Future Research Opportunities

Building on these limitations, there are several future research directions that present an opportunity to delve deeper into this topic.

Firstly, to overcome the challenge of the availability of text data in the construction sector, future research can focus on 1) Enlarging the dataset by augmenting the amount of annotated data and incorporating a wider variety of contract types beyond P3 agreements. Given the complexity of P3 contracts and the demonstrated feasibility of analysis in the P3 context, similar approaches could be applied to other contractual forms. 2) Investigating other unsupervised learning approaches to leverage the full potential of LLMs, thus reducing the dependency on annotated datasets.

Secondly, the current application of LLMs, limited by resource and computational capacity, has focused on prompt engineering and fine-tuning with GPT-3.5 Turbo. Future research could attempt to develop specialized LLMs tailored for contract analysis by integrating more specialized knowledge. This would potentially overcome the limitations faced by general-purpose models in understanding complex legal documents, thus allowing for a more customized analysis of contract specifics.

Thirdly, future research could improve the tool both in terms of the risk framework design and model development for a more nuanced identification of risk features. 1) For risk type identification, exploring more output formats beyond classifications could capture the intercorrelation between risks. For example, both “direct” and “indirect” risks can be involved in a sentence, as an “indirect” one can be triggered by a “direct” one. For the sake of clarity and

conciseness, the current binary classification system is effective for identifying direct risks. An enhanced risk framework could introduce mechanisms to score the level of mention of each risk, providing a richer risk profile. 2) In terms of risk allocation, delving deeper into the subtleties of contract language could reveal more detailed information, particularly the inclination towards the public or private partner's interests. This could be achieved by a more fine-grained division of current categories to include additional ones such as obligation exemptions (Circumstances under which a party may be relieved of their obligations), liability exclusions (Limiting or excluding a party's liability under certain circumstances), and restricted rights (A right that is limited or removed under the contract terms). For example, most obligation languages specify the work to be done, but there are a small portion of cases where exemptions from the obligation are provided for. Although preliminary attempts to categorize obligations have been made, the current models and methodologies have not yet achieved satisfactory performance in classifying these detailed distinctions. This requires an advanced approach that can effectively differentiate between these fine-grained categories.

Finally, this research assumes the existence of pre-designed contracts, focusing solely on text recognition rather than generation. Contract generation is a much more challenging task, as construction contracts' complexity demands a large corpus for effective model training. Future research might start by exploring the generation of specific contract segments with more standardized content. The potential for leveraging LLMs in contract generation remains an area for exploration.

Appendices

Appendix A. Demonstrations for Prompt Engineering

ID	Sentence	Risk type label	Risk allocation label
1	Subject to Section 10.04, if the Developer must submit a submittal or request to the Department for review and Response more than twice due to the Developer's failure to comply with the requirements of this Agreement, the Developer will pay the Department for the Department's Allocable Costs incurred thereafter in reviewing any portions of such submittal or request.	['performance']	['private liability', 'public right']
2	The Developer will be responsible for coordinating and scheduling the Work with other separate contractors working in the Project Right of Way in accordance with the Technical Requirements.	['other']	['private obligation']
3	The Developer will provide appropriate oversight, management and reporting of all phases of the Project and its Contractors such that the Project is delivered, operated and maintained in accordance with this Agreement.	['other']	['private obligation']
4	The Department's review of any submittal will comply with the submittal and review procedures set forth in Section 10.05.	['other']	['public obligation']
5	This Agreement does not grant to the Developer any fee title, leasehold estate, easement or	['other']	['private right']

	other real property interest of any kind in or to the Project Assets or the Project Right of Way.		
6	(c) In consideration of the Permit granted to the Developer by the Department pursuant to this Section 4.01, the Developer will perform the Work in accordance with the terms hereof at its own expense except as otherwise provided herein and pay (to the extent required) to the Department the Permit Fee in accordance with the Permit Fee calculation attached as Exhibit J.	['other']	['private obligation', 'public right']
7	If no agreement is reached within such sixty (60)-day period as to any such matter, either Party may submit the Dispute to the Dispute Resolution Procedure.	['disagreement and dispute']	['public right', 'private right']
8	If HPTE issues a notice under Section 17.5(a) the Concessionaire shall bear its own costs and pay to HPTE on demand all reasonable costs and expenses incurred by or on behalf of HPTE in relation to the costs of the increased level of monitoring.	['performance']	['private liability', 'public right']
9	(b) HPTE shall be responsible or shall ensure that CDOT will be responsible in accordance with Sections 9.4(c), 9.4(d) and 9.4(e) for matters arising out of HPTE Hazardous Substances Circumstances.	['environmental and site']	['public obligation', 'public liability']
10	To the extent Liquidated Damages and/or sums due and payable to HPTE in accordance with Section 20.2(a) are not deducted from any amount owed by HPTE to the Concessionaire, HPTE may send the Concessionaire an invoice, and the Liquidated	['performance']	['public right', 'private obligation', 'private liability']

	Damages and/or such sums shall be payable by the Concessionaire to HPTE within ten (10) Business Days after the Concessionaire's receipt of the invoice.		
11	(a) The Concessionaire shall be responsible for obtaining all Necessary Consents and for arranging any necessary amendments to any Necessary Consents and such responsibility shall not be in any way diminished by any Law placing responsibility for the same upon HPTE or another Person.	['utilities, permits, and third-party']	['private obligation', 'private liability']
12	(b) If for any reason any individual in a Key Personnel role resigns, retires, dies, becomes disabled, or is terminated for cause, then the Concessionaire shall designate a replacement with equivalent expertise and experience to the unavailable individual and provide notice to HPTE setting out the identity, expertise and experience of the proposed replacement and such supporting information or evidence as HPTE may reasonably require in relation to such matters.	['other']	['private obligation', 'public right']
13	If it is agreed or determined by HPTE that Section 40.2 applies, then HPTE shall waive the Concessionaire's obligations in Section 37 and/or Schedule 17 (Required Insurances) in respect of that particular Insurance Term, and the Concessionaire shall not be considered in breach of its obligations regarding the maintenance of insurance pursuant to this Contract as a result of the failure to maintain insurance incorporating such Insurance Term for so long as the relevant circumstances described in this Section 40 continue to apply to such	['other']	['private obligation']

	Insurance Term.		
14	Change to the Financial Close Deadline Date (a) At any time prior to the then-current Financial Close Deadline Date the Concessionaire may submit a notice in writing to HPTE asking HPTE to extend the Financial Close Deadline Date.	['changes by contracting parties', 'economic and financial']	['private right']
15	(i) The Developer will pay to the Department 50% of any Refinancing Gain from a Refinancing that is not an Exempt Refinancing.	['economic and financial']	['private liability', 'public right']
16	(a) The Developer will furnish all design, construction and other services, provide all materials, equipment and labor to perform the Work reasonably inferable from this Agreement and perform the Work in accordance with this Agreement.	['other']	['private obligation']
17	If the need for such supplements or additional NEPA documents arises from a Department Change or a Compensation Event, the Department will bear the costs of preparing the necessary documentation, and the Department will pay the Developer’s Allocable Costs associated with the preparation of the data and other information provided by the Developer; in all other cases, the Developer will bear the costs of preparing the necessary documentation and will pay the Department for its Allocable Costs incurred in the preparation of such documentation.	['changes by contracting parties']	['public liability', 'private right', 'private liability', 'public right']

18	Notwithstanding the foregoing, if at any time prior to the Department assessing Project Completion Liquidated Damages, the Developer fails to achieve both the P&R Milestone and the Route 28 Signalization Milestone under Section 8.10(b), the Department will be entitled to assess Milestone Liquidated Damages on both Intermediate Milestones concurrently.	['performance']	['public right', 'private liability']

Bibliography

Abednego, M. P., & Ogunlana, S. O. (2006). Good project governance for proper risk allocation in public–private partnerships in Indonesia. *International Journal of Project Management*, *24*(7), 622–634. https://doi.org/10.1016/j.ijproman.2006.07.010

Abubakar, M. E., Hasan, A., & Jha, K. N. (2022). Delays and Financial Implications of COVID-19 for Contractors in Irrigation Projects. *Journal of Construction Engineering and Management*, *148*(9), 05022006. https://doi.org/10.1061/(ASCE)CO.1943-7862.0002329

Agrawal, A. K., Jagannathan, M., & Delhi, V. S. K. (2021). Control Focus in Standard Forms: An Assessment through Text Mining and NLP. *Journal of Legal Affairs and Dispute Resolution in Engineering and Construction*, *13*(1), 04520040. https://doi.org/10.1061/(ASCE)LA.1943-4170.0000441

Ahmed, M., & Garvin, M. (2022). Review of Critical Success Factors and Key Performance Indicators in Performance Assessment of P3 Transportation Projects. *Journal of Management in Engineering*, *38*(5), 04022045. https://doi.org/10.1061/(ASCE)ME.1943-5479.0001070

Akanbi, T., & Zhang, J. (2021). Design information extraction from construction specifications to support cost estimation. *Automation in Construction*, *131*, 103835. https://doi.org/10.1016/j.autcon.2021.103835

Al Qady, M., & Kandil, A. (2010). Concept Relation Extraction from Construction Documents Using Natural Language Processing. *Journal of Construction Engineering and Management*, *136*(3), 294–302. https://doi.org/10.1061/(ASCE)CO.1943-7862.0000131

Alsharef, A., Banerjee, S., Uddin, S. M. J., Albert, A., & Jaselskis, E. (2021). Early Impacts of the COVID-19 Pandemic on the United States Construction Industry. *International*

Journal of Environmental Research and Public Health, *18*(4), Article 4. https://doi.org/10.3390/ijerph18041559

Anthropic. (2023). *Introducing Claude*. https://www.anthropic.com/news/introducing-claude

Arcadis. (2022). *2022 Global Construction Disputes Report*. Amsterdam, Netherlands: Arcadis. https://www.arcadis.com/en-us/knowledge-hub/perspectives/global/global-construction-disputes-report

Arlot, S., & Celisse, A. (2010). A survey of cross-validation procedures for model selection. *Statistics Surveys*, *4*(none), 40–79. https://doi.org/10.1214/09-SS054

Baek, S., Jung, W., & Han, S. H. (2021). A critical review of text-based research in construction: Data source, analysis method, and implications. *Automation in Construction*, *132*, 103915. https://doi.org/10.1016/j.autcon.2021.103915

Baker, H., Hallowell, M. R., & Tixier, A. J.-P. (2020). Automatically learning construction injury precursors from text. *Automation in Construction*, *118*, 103145. https://doi.org/10.1016/j.autcon.2020.103145

Baxter, D., & Casady, C. B. (2020). A Coronavirus (COVID-19) Triage Framework for (Sub)National Public–Private Partnership (PPP) Programs. *Sustainability*, *12*(13), Article 13. https://doi.org/10.3390/su12135253

Bilal, M., Oyedele, L. O., Qadir, J., Munir, K., Ajayi, S. O., Akinade, O. O., Owolabi, H. A., Alaka, H. A., & Pasha, M. (2016). Big Data in the construction industry: A review of present status, opportunities, and future trends. *Advanced Engineering Informatics*, *30*(3), 500–521. https://doi.org/10.1016/j.aei.2016.07.001

Bing, L., Akintoye, A., Edwards, P. J., & Hardcastle, C. (2005). The allocation of risk in PPP/PFI construction projects in the UK. *International Journal of Project Management*, *23*(1), 25–35. https://doi.org/10.1016/j.ijproman.2004.04.006

Brown, J. W., Pieplow, R., Driskell, R., Gaj, S., Garvin, M. J., Holcombe, D., Saunders, M., Seiders Jr., J., & Smith, A. (2009). *Public-Private Partnerships for Highway Infrastructure: Capitalizing on International Experience* (FHWA-PL-09-010). Federal Highway Administration, US Department of Transportation, Washington, DC. https://international.fhwa.dot.gov/pubs/pl09010/

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., … Amodei, D. (2020). Language Models are Few-Shot Learners. *Advances in Neural Information Processing Systems*, *33*, 1877–1901.

Buckland, M., & Gey, F. (1994). The relationship between Recall and Precision. *Journal of the American Society for Information Science*, *45*(1), 12–19. https://doi.org/10.1002/(SICI)1097-4571(199401)45:1<12::AID-ASI2>3.0.CO;2-L

Bunni, N. G., & Bunni, L. B. (2022). *Risk and Insurance in Construction* (3rd ed.). Routledge. https://doi.org/10.1201/9781003222514

Candaş, A. B., & Tokdemir, O. B. (2022). Automating Coordination Efforts for Reviewing Construction Contracts with Multilabel Text Classification. *Journal of Construction Engineering and Management*, *148*(6), 04022027. https://doi.org/10.1061/(ASCE)CO.1943-7862.0002275

Cao, Y., Li, S., Liu, Y., Yan, Z., Dai, Y., Yu, P. S., & Sun, L. (2023). *A Comprehensive Survey of AI-Generated Content (AIGC): A History of Generative AI from GAN to ChatGPT* (arXiv:2303.04226). arXiv. https://doi.org/10.48550/arXiv.2303.04226

Cao, Z., & Lumineau, F. (2015). Revisiting the interplay between contractual and relational governance: A qualitative and meta-analytic investigation. *Journal of Operations Management*, *33–34*, 15–42. https://doi.org/10.1016/j.jom.2014.09.009

Carbonara, N., Costantino, N., Gunnigan, L., & Pellegrino, R. (2015). Risk Management in Motorway PPP Projects: Empirical-based Guidelines. *Transport Reviews*, *35*(2), 162–182. https://doi.org/10.1080/01441647.2015.1012696

Casady, C. B., & Baxter, D. (2020). Pandemics, public-private partnerships (PPPs), and force majeure | COVID-19 expectations and implications. *Construction Management and Economics*, *38*(12), 1077–1085. https://doi.org/10.1080/01446193.2020.1817516

Castelblanco, G., Guevara, J., Mesa, H., & Flores, D. (2020). Risk Allocation in Unsolicited and Solicited Road Public-Private Partnerships: Sustainability and Management Implications. *Sustainability*, *12*(11), Article 11. https://doi.org/10.3390/su12114478

Chalkidis, I., Fergadiotis, M., Malakasiotis, P., Aletras, N., & Androutsopoulos, I. (2020). *LEGAL-BERT: The Muppets straight out of Law School* (arXiv:2010.02559). arXiv. https://doi.org/10.48550/arXiv.2010.02559

Charoenngam, C., & Yeh, C.-Y. (1999). Contractual risk and liability sharing in hydropower construction. *International Journal of Project Management*, *17*(1), 29–37. https://doi.org/10.1016/S0263-7863(97)00064-1

Cheung, E., & Chan, A. P. C. (2011). Risk Factors of Public-Private Partnership Projects in China: Comparison between the Water, Power, and Transportation Sectors. *Journal of*

Urban Planning and Development, *137*(4), 409–415. https://doi.org/10.1061/(ASCE)UP.1943-5444.0000086

Choi, S. J., Choi, S. W., Kim, J. H., & Lee, E.-B. (2021). AI and Text-Mining Applications for Analyzing Contractor's Risk in Invitation to Bid (ITB) and Contracts for Engineering Procurement and Construction (EPC) Projects. *Energies*, *14*(15), Article 15. https://doi.org/10.3390/en14154632

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H. W., Sutton, C., Gehrmann, S., Schuh, P., Shi, K., Tsvyashchenko, S., Maynez, J., Rao, A., Barnes, P., Tay, Y., Shazeer, N., Prabhakaran, V., … Fiedel, N. (2022, April 5). *PaLM: Scaling Language Modeling with Pathways*. arXiv.Org. https://arxiv.org/abs/2204.02311v5

Chung, D., Hensher, D. A., & Rose, J. M. (2010). Toward the betterment of risk allocation: Investigating risk perceptions of Australian stakeholder groups to public–private-partnership tollroad projects. *Research in Transportation Economics*, *30*(1), 43–58. https://doi.org/10.1016/j.retrec.2010.10.007

Chung, H. W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, Y., Wang, X., Dehghani, M., Brahma, S., Webson, A., Gu, S. S., Dai, Z., Suzgun, M., Chen, X., Chowdhery, A., Castro-Ros, A., Pellat, M., Robinson, K., … Wei, J. (2022). *Scaling Instruction-Finetuned Language Models* (arXiv:2210.11416). arXiv. https://doi.org/10.48550/arXiv.2210.11416

Chung, S., Moon, S., Kim, J., Kim, J., Lim, S., & Chi, S. (2023). Comparing natural language processing (NLP) applications in construction and computer science using preferred

reporting items for systematic reviews (PRISMA). *Automation in Construction*, *154*, 105020. https://doi.org/10.1016/j.autcon.2023.105020

Congressional Research Service. (2021). *Public-Private Partnerships (P3s) in Transportation* (R45010). Congressional Research Service. https://sgp.fas.org/crs/misc/R45010.pdf

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., & Stoyanov, V. (2020). *Unsupervised Cross-lingual Representation Learning at Scale* (arXiv:1911.02116). arXiv. https://doi.org/10.48550/arXiv.1911.02116

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding* (arXiv:1810.04805). arXiv. https://doi.org/10.48550/arXiv.1810.04805

Dewulf, G., & Garvin, M. J. (2020). Responsive governance in PPP projects to manage uncertainty. *Construction Management and Economics*, *38*(4), 383–397. https://doi.org/10.1080/01446193.2019.1618478

Ding, Y., Ma, J., & Luo, X. (2022). Applications of natural language processing in construction. *Automation in Construction*, *136*, 104169. https://doi.org/10.1016/j.autcon.2022.104169

Dong, Q., Li, L., Dai, D., Zheng, C., Wu, Z., Chang, B., Sun, X., Xu, J., Li, L., & Sui, Z. (2022, December 31). *A Survey on In-context Learning*. arXiv.Org. https://arxiv.org/abs/2301.00234v3

Fatemi, S., & Hu, Y. (2023). *A Comparative Analysis of Fine-Tuned LLMs and Few-Shot Learning of LLMs for Financial Sentiment Analysis* (arXiv:2312.08725). arXiv. https://doi.org/10.48550/arXiv.2312.08725

Fathi, M., & Shrestha, P. P. (2023). Public–Private Partnership Contract Framework Development for Highway Projects: A Delphi Approach. *Journal of Legal Affairs and Dispute Resolution in Engineering and Construction*, *15*(1), 04522046. https://doi.org/10.1061/(ASCE)LA.1943-4170.0000575

Federal Highway Administration. (2014). *Risk Assessment for Public–Private Partnerships: A Primer* (FHWA-OIPD-13-004). Federal Highway Administration, U.S. Department of Transportation. https://www.fhwa.dot.gov/ipd/p3/toolkit/publications/primers/risk_assessment/ch_1.aspx

Federal Highway Administration. (2015). *Availability Payment Concessions Public-Private Partnerships Model Contract Guide*. https://www.fhwa.dot.gov/ipd/pdfs/p3/apguide.pdf

Feng, D., & Chen, H. (2021). A small samples training framework for deep Learning-based automatic information extraction: Case study of construction accident news reports analysis. *Advanced Engineering Informatics*, *47*, 101256. https://doi.org/10.1016/j.aei.2021.101256

Finkel, R. M., Trenor, J. A., & Soiffer, A. (2020, April 14). *COVID-19: Drafting Force Majeure Clauses in Light of the COVID-19 Pandemic*. https://www.wilmerhale.com/en/insights/client-alerts/20200413-drafting-force-majeure-clauses-in-light-of-the-covid-19-pandemic

Fu, Y., Peng, H., Sabharwal, A., Clark, P., & Khot, T. (2022, September 29). *Complexity-Based Prompting for Multi-step Reasoning*. The Eleventh International Conference on Learning Representations. https://openreview.net/forum?id=yf1icZHC-l9

Fu, Y., Xu, C., Zhang, L., & Chen, Y. (2023). Control, coordination, and adaptation functions in construction contracts: A machine-coding model. *Automation in Construction*, *152*, 104890. https://doi.org/10.1016/j.autcon.2023.104890

Gao, N., Touran, A., & Wang, Q. (2022). *Mining and Visualizing Cost and Schedule Risks from News Articles with NLP and Network Analysis*. 314–324. https://doi.org/10.1061/9780784483961.034

Ghojogh, B., & Crowley, M. (2023). *The Theory Behind Overfitting, Cross Validation, Regularization, Bagging, and Boosting: Tutorial* (arXiv:1905.12787). arXiv. https://doi.org/10.48550/arXiv.1905.12787

Godbole, S., & Sarawagi, S. (2004). Discriminative Methods for Multi-labeled Classification. In H. Dai, R. Srikant, & C. Zhang (Eds.), *Advances in Knowledge Discovery and Data Mining* (pp. 22–30). Springer. https://doi.org/10.1007/978-3-540-24775-3_5

Grimsey, D., & Lewis, M. K. (2004). *Public Private Partnerships: The Worldwide Revolution in Infrastructure Provision and Project Finance*. Edward Elgar Publishing.

Han, X., Zhang, Z., Ding, N., Gu, Y., Liu, X., Huo, Y., Qiu, J., Yao, Y., Zhang, A., Zhang, L., Han, W., Huang, M., Jin, Q., Lan, Y., Liu, Y., Liu, Z., Lu, Z., Qiu, X., Song, R., … Zhu, J. (2021). *Pre-Trained Models: Past, Present and Future* (arXiv:2106.07139). arXiv. https://doi.org/10.48550/arXiv.2106.07139

Hansen, S. (2020). Does the COVID-19 Outbreak Constitute a Force Majeure Event? A Pandemic Impact on Construction Contracts. *Journal of the Civil Engineering Forum*, *6*, 201–214. https://doi.org/10.22146/jcef.54997

Hassan, F. ul, & Le, T. (2020). Automated Requirements Identification from Construction Contract Documents Using Natural Language Processing. *Journal of Legal Affairs and*

Dispute Resolution in Engineering and Construction, *12*(2), 04520009. https://doi.org/10.1061/(ASCE)LA.1943-4170.0000379

Heravi, G., & Hajihosseini, Z. (2012). Risk Allocation in Public–Private Partnership Infrastructure Projects in Developing Countries: Case Study of the Tehran–Chalus Toll Road. *Journal of Infrastructure Systems*, *18*(3), 210–217. https://doi.org/10.1061/(ASCE)IS.1943-555X.0000090

Hodge, G. A. (2004a). Risks in Public-Private Partnerships: Shifting, Sharing or Shirking? *Asia Pacific Journal of Public Administration*, *26*(2), 155–179. https://doi.org/10.1080/23276665.2004.10779291

Hodge, G. A. (2004b). The risky business of public–private partnerships. *Australian Journal of Public Administration*, *63*(4), 37–49. https://doi.org/10.1111/j.1467-8500.2004.00400.x

Hovy, P. (2015, September 10). *Risk Allocation in Public-Private Partnerships: Maximizing value for money*. International Institute for Sustainable Development. https://www.iisd.org/publications/report/risk-allocation-public-private-partnerships-maximizing-value-money

Huang, J., & Chang, K. C.-C. (2023). *Towards Reasoning in Large Language Models: A Survey* (arXiv:2212.10403). arXiv. https://doi.org/10.48550/arXiv.2212.10403

Iossa, E., Spagnolo, G., & Vellez, M. (2007). *Contract Design in Public-Private Partnerships*. https://doi.org/10.13140/RG.2.2.19597.79847

Jahedi, S., & Méndez, F. (2014). On the advantages and disadvantages of subjective measures. *Journal of Economic Behavior & Organization*, *98*, 97–114. https://doi.org/10.1016/j.jebo.2013.12.016

Joulin, A., Grave, E., Bojanowski, P., Douze, M., Jégou, H., & Mikolov, T. (2016). *FastText.zip: Compressing text classification models* (arXiv:1612.03651). arXiv. https://doi.org/10.48550/arXiv.1612.03651

Ke, Y., Wang, S., & Chan, A. P. C. (2010). Risk Allocation in Public-Private Partnership Infrastructure Projects: Comparative Study. *Journal of Infrastructure Systems*, *16*(4), 343–351. https://doi.org/10.1061/(ASCE)IS.1943-555X.0000030

Khalef, R., & El-adaway, I. H. (2021). Automated Identification of Substantial Changes in Construction Projects of Airport Improvement Program: Machine Learning and Natural Language Processing Comparative Analysis. *Journal of Management in Engineering*, *37*(6), 04021062. https://doi.org/10.1061/(ASCE)ME.1943-5479.0000959

Khalef, R., El-adaway, I. H., Assaad, R., & Kieta, N. (2021). Contract Risk Management: A Comparative Study of Risk Allocation in Exculpatory Clauses and Their Legal Treatment. *Journal of Legal Affairs and Dispute Resolution in Engineering and Construction*, *13*(1), 04520036. https://doi.org/10.1061/(ASCE)LA.1943-4170.0000430

Kim, H., Lee, H.-S., Park, M., Chung, B., & Hwang, S. (2015). Information Retrieval Framework for Hazard Identification in Construction. *Journal of Computing in Civil Engineering*, *29*(3), 04014052. https://doi.org/10.1061/(ASCE)CP.1943-5487.0000340

Kojima, T., Gu, S. (Shane), Reid, M., Matsuo, Y., & Iwasawa, Y. (2022). Large Language Models are Zero-Shot Reasoners. *Advances in Neural Information Processing Systems*, *35*, 22199–22213.

Kumari, A., & Kumar Sharma, A. (2017). Infrastructure financing and development: A bibliometric review. *International Journal of Critical Infrastructure Protection*, *16*, 49–65. https://doi.org/10.1016/j.ijcip.2016.11.005

Kwak, Y. H., Chih, Y., & Ibbs, C. W. (2009). Towards a Comprehensive Understanding of Public Private Partnerships for Infrastructure Development. *California Management Review*, *51*(2), 51–78. https://doi.org/10.2307/41166480

Lam, K. C., Wang, D., Lee, P. T. K., & Tsang, Y. T. (2007). Modelling risk allocation decision in construction contracts. *International Journal of Project Management*, *25*(5), 485–493. https://doi.org/10.1016/j.ijproman.2006.11.005

Lauriola, I., Lavelli, A., & Aiolli, F. (2022). An introduction to Deep Learning in Natural Language Processing: Models, techniques, and tools. *Neurocomputing*, *470*, 443–456. https://doi.org/10.1016/j.neucom.2021.05.103

Le, P. T., Kirytopoulos, K., Chileshe, N., & Rameezdeen, R. (2019). Taxonomy of risks in PPP transportation projects: A systematic literature review. *International Journal of Construction Management*, *22*(2), 166–181. https://doi.org/10.1080/15623599.2019.1615756

Lee, G., Won, J., Ham, S., & Shin, Y. (2011). Metrics for Quantifying the Similarities and Differences between IFC Files. *Journal of Computing in Civil Engineering*, *25*(2), 172–181. https://doi.org/10.1061/(ASCE)CP.1943-5487.0000077

Lee, J., Ham, Y., Yi, J.-S., & Son, J. (2020). Effective Risk Positioning through Automated Identification of Missing Contract Conditions from the Contractor's Perspective Based on FIDIC Contract Cases. *Journal of Management in Engineering*, *36*(3), 05020003. https://doi.org/10.1061/(ASCE)ME.1943-5479.0000757

Lee, J., Yi, J.-S., & Son, J. (2019). Development of Automatic-Extraction Model of Poisonous Clauses in International Construction Contracts Using Rule-Based NLP. *Journal of*

Computing in Civil Engineering, *33*(3), 04019003. https://doi.org/10.1061/(ASCE)CP.1943-5487.0000807

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., & Zettlemoyer, L. (2019). *BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension* (arXiv:1910.13461). arXiv. https://doi.org/10.48550/arXiv.1910.13461

Li, J., Tang, T., Zhao, W. X., & Wen, J.-R. (2021). *Pretrained Language Models for Text Generation: A Survey* (arXiv:2105.10311). arXiv. https://doi.org/10.48550/arXiv.2105.10311

Li, J., & Zou, P. X. W. (2011). Fuzzy AHP-Based Risk Assessment Methodology for PPP Projects. *Journal of Construction Engineering and Management*, *137*(12), 1205–1209. https://doi.org/10.1061/(ASCE)CO.1943-7862.0000362

Li, Y., Guzman, E., Tsiamoura, K., Schneider, F., & Bruegge, B. (2015). Automated Requirements Extraction for Scientific Software. *Procedia Computer Science*, *51*, 582–591. https://doi.org/10.1016/j.procs.2015.05.326

Lie, J., & Zou, P. (2008). Risk identification and assessment in PPP infrastructure projects using fuzzy Analytical Hierarchy Process and life-cycle methodology. *Construction Economics and Building*, *8*(1), 32–46. https://doi.org/10.5130/AJCEB.v8i1.2996

Ling, C., Zhao, X., Lu, J., Deng, C., Zheng, C., Wang, J., Chowdhury, T., Li, Y., Cui, H., Zhang, X., Zhao, T., Panalkar, A., Cheng, W., Wang, H., Liu, Y., Chen, Z., Chen, H., White, C., Gu, Q., … Zhao, L. (2023). *Domain Specialization as the Key to Make Large Language Models Disruptive: A Comprehensive Survey* (arXiv:2305.18703). arXiv. https://doi.org/10.48550/arXiv.2305.18703

Liu, J., Luo, H., & Liu, H. (2022). Deep learning-based data analytics for safety in construction. *Automation in Construction*, *140*, 104302. https://doi.org/10.1016/j.autcon.2022.104302

Liu, N. F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., & Liang, P. (2023, July 6). *Lost in the Middle: How Language Models Use Long Contexts*. arXiv.Org. https://arxiv.org/abs/2307.03172v2

Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., & Neubig, G. (2023). Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing. *ACM Computing Surveys*, *55*(9), 195:1-195:35. https://doi.org/10.1145/3560815

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). *RoBERTa: A Robustly Optimized BERT Pretraining Approach* (arXiv:1907.11692). arXiv. https://doi.org/10.48550/arXiv.1907.11692

Lv, X., & El-Gohary, N. (2016). Text Analytics for Supporting Stakeholder Opinion Mining for Large-scale Highway Projects. *Procedia Engineering*, *145*, 518–524. https://doi.org/10.1016/j.proeng.2016.04.039

Manikandan, H., Jiang, Y., & Kolter, J. Z. (2023). Language Models are Weak Learners. *Advances in Neural Information Processing Systems*, *36*, 50907–50931.

Manning, C., & Schutze, H. (1999). *Foundations of Statistical Natural Language Processing*. MIT Press.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013a). *Distributed Representations of Words and Phrases and their Compositionality* (arXiv:1310.4546). arXiv. https://doi.org/10.48550/arXiv.1310.4546

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013b). *Distributed Representations of Words and Phrases and their Compositionality* (arXiv:1310.4546). arXiv. https://doi.org/10.48550/arXiv.1310.4546

Min, S., Lyu, X., Holtzman, A., Artetxe, M., Lewis, M., Hajishirzi, H., & Zettlemoyer, L. (2022). *Rethinking the Role of Demonstrations: What Makes In-Context Learning Work?* (arXiv:2202.12837). arXiv. https://doi.org/10.48550/arXiv.2202.12837

Minaee, S., Mikolov, T., Nikzad, N., Chenaghlu, M., Socher, R., Amatriain, X., & Gao, J. (2024). *Large Language Models: A Survey* (arXiv:2402.06196). arXiv. https://doi.org/10.48550/arXiv.2402.06196

Moon, S., Chi, S., & Im, S.-B. (2022). Automated detection of contractual risk clauses from construction specifications using bidirectional encoder representations from transformers (BERT). *Automation in Construction*, *142*, 104465. https://doi.org/10.1016/j.autcon.2022.104465

Moon, S., Lee, G., & Chi, S. (2022). Automated system for construction specification review using natural language processing. *Advanced Engineering Informatics*, *51*, 101495. https://doi.org/10.1016/j.aei.2021.101495

Naveed, H., Khan, A. U., Qiu, S., Saqib, M., Anwar, S., Usman, M., Akhtar, N., Barnes, N., & Mian, A. (2023). *A Comprehensive Overview of Large Language Models* (arXiv:2307.06435). arXiv. https://doi.org/10.48550/arXiv.2307.06435

Nguyen, D. A., Garvin, M. J., & Gonzalez, E. E. (2018). Risk Allocation in U.S. Public-Private Partnership Highway Project Contracts. *Journal of Construction Engineering and Management*, *144*(5), 04018017. https://doi.org/10.1061/(ASCE)CO.1943-7862.0001465

Ni, A. Y. (2012). The Risk-Averting Game of Transport Public-Private Partnership. *Public Performance & Management Review*, *36*(2), 253–274. https://doi.org/10.2753/PMR1530-9576360205

Nicolini-Llosa, J. L. (2002). Toll Road Concessions in Argentina: What Can Be Learned. *Transportation Research Record*, *1812*(1), 10–21. https://doi.org/10.3141/1812-02

OpenAI. (2023). *GPT-4 Technical Report* (arXiv:2303.08774). arXiv. https://doi.org/10.48550/arXiv.2303.08774

OpenAI. (2024a). *OpenAI Platform—API reference*. https://platform.openai.com/docs/api-reference

OpenAI. (2024b). *OpenAI Platform—Models*. https://platform.openai.com/docs/models

Osei-Kyei, R., Jin, X., Nnaji, C., Akomea-Frimpong, I., & Wuni, I. Y. (2023). Review of risk management studies in public-private partnerships: A scientometric analysis. *International Journal of Construction Management*, *23*(14), 2419–2430. https://doi.org/10.1080/15623599.2022.2063013

Padhy, J., Jagannathan, M., & Kumar Delhi, V. S. (2021). Application of Natural Language Processing to Automatically Identify Exculpatory Clauses in Construction Contracts. *Journal of Legal Affairs and Dispute Resolution in Engineering and Construction*, *13*(4), 04521035. https://doi.org/10.1061/(ASCE)LA.1943-4170.0000505

Pahune, S., & Chandrasekharan, M. (2023). Several categories of Large Language Models (LLMs): A Short Survey. *International Journal for Research in Applied Science and Engineering Technology*, *11*(7), 615–633. https://doi.org/10.22214/ijraset.2023.54677

Pan, Y., & Zhang, L. (2021). Roles of artificial intelligence in construction engineering and management: A critical review and future trends. *Automation in Construction*, *122*, 103517. https://doi.org/10.1016/j.autcon.2020.103517

Papajohn, D., Cui, Q., & Bayraktar, M. E. (2011). Public-Private Partnerships in U.S. Transportation: Research Overview and a Path Forward. *Journal of Management in Engineering*, *27*(3), 126–135. https://doi.org/10.1061/(ASCE)ME.1943-5479.0000050

Park, J. Y., Mistur, E., Kim, D., Mo, Y., & Hoefer, R. (2022). Toward human-centric urban infrastructure: Text mining for social media data to identify the public perception of COVID-19 policy in transportation hubs. *Sustainable Cities and Society*, *76*, 103524. https://doi.org/10.1016/j.scs.2021.103524

Pellegrino, R., Vajdic, N., & Carbonara, N. (2013). Real option theory for risk mitigation in transport PPPs. *Built Environment Project and Asset Management*, *3*. https://doi.org/10.1108/BEPAM-05-2012-0027

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543. https://aclanthology.org/D14-1162.pdf

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). *Deep contextualized word representations* (arXiv:1802.05365). arXiv. https://doi.org/10.48550/arXiv.1802.05365

Pham, H. T. T. L., & Han, S. (2023). Natural Language Processing with Multitask Classification for Semantic Prediction of Risk-Handling Actions in Construction Contracts. *Journal of Computing in Civil Engineering*, *37*(6), 04023027. https://doi.org/10.1061/JCCEE5.CPENG-5218

Qi, X., Chen, Y., Lai, J., & Meng, F. (2024). Multifunctional Analysis of Construction Contracts Using a Machine Learning Approach. *Journal of Management in Engineering*, *40*(2), 04024002. https://doi.org/10.1061/JMENEA.MEENG-5604

Qiao, S., Ou, Y., Zhang, N., Chen, X., Yao, Y., Deng, S., Tan, C., Huang, F., & Chen, H. (2023). *Reasoning with Language Model Prompting: A Survey* (arXiv:2212.09597). arXiv. https://doi.org/10.48550/arXiv.2212.09597

Qiu, X., Sun, T., Xu, Y., Shao, Y., Dai, N., & Huang, X. (2020). Pre-trained models for natural language processing: A survey. *Science China Technological Sciences*, *63*(10), 1872–1897. https://doi.org/10.1007/s11431-020-1647-3

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). *Improving Language Understanding by Generative Pre-Training*. https://www.mikecaptain.com/resources/pdf/GPT-1.pdf

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). *Language Models are Unsupervised Multitask Learners*.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, *21*(1), 140:5485-140:5551.

Sainz, O., Qiu, H., Lopez de Lacalle, O., Agirre, E., & Min, B. (2022). ZS4IE: A toolkit for Zero-Shot Information Extraction with simple Verbalizations. In H. Hajishirzi, Q. Ning, & A. Sil (Eds.), *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: System Demonstrations* (pp. 27–38). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.naacl-demo.4

Santu, S. K. K., & Feng, D. (2023, May 19). *TELeR: A General Taxonomy of LLM Prompts for Benchmarking Complex Tasks*. arXiv.Org. https://arxiv.org/abs/2305.11430v1

Scao, T. L., Fan, A., Akiki, C., Pavlick, E., Ilić, S., Hesslow, D., Castagné, R., Luccioni, A. S., Yvon, F., Gallé, M., Tow, J., Rush, A. M., Biderman, S., Webson, A., Ammanamanchi, P. S., Wang, T., Sagot, B., Muennighoff, N., del Moral, A. V., … Wolf, T. (2022, November 9). *BLOOM: A 176B-Parameter Open-Access Multilingual Language Model*. arXiv.Org. https://arxiv.org/abs/2211.05100v4

Shen, T., Long, G., Geng, X., Tao, C., Zhou, T., & Jiang, D. (2023). *Large Language Models are Strong Zero-Shot Retriever* (arXiv:2304.14233). arXiv. https://doi.org/10.48550/arXiv.2304.14233

Shi, X., Liu, Y.-S., Gao, G., Gu, M., & Li, H. (2018). IFCdiff: A content-based automatic comparison approach for IFC files. *Automation in Construction*, *86*, 53–68. https://doi.org/10.1016/j.autcon.2017.10.013

Smith, N., Peña, P. de la, Kussy, E., Sethi, S., Wheeler, P., Gifford, J., & Ybarra, S. (2019). *Public-Private Partnership (P3) Procurement: A Guide for Public Owners* (FHWA-HIN-18-004). https://www.fhwa.dot.gov/ipd/p3/toolkit/publications/other_guides/p3_procurement_guide_0319/ch_1.aspx

Sun, Z. (2023). *A Short Survey of Viewing Large Language Models in Legal Aspect* (arXiv:2303.09136). arXiv. https://doi.org/10.48550/arXiv.2303.09136

Tenney, I., Das, D., & Pavlick, E. (2019, May 15). *BERT Rediscovers the Classical NLP Pipeline*. arXiv.Org. https://arxiv.org/abs/1905.05950v2

Thomas, A. V., Kalidindi, S. N., & Ananthanarayanan, K. (2003). Risk perception analysis of BOT road project participants in India. *Construction Management and Economics*, *21*(4), 393–407. https://doi.org/10.1080/0144619032000064127

Tixier, A. J.-P., Hallowell, M. R., Rajagopalan, B., & Bowman, D. (2016). Automated content analysis for construction safety: A natural language processing system to extract precursors and outcomes from unstructured injury reports. *Automation in Construction*, *62*, 45–56. https://doi.org/10.1016/j.autcon.2015.11.001

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., & Lample, G. (2023, February 27). *LLaMA: Open and Efficient Foundation Language Models*. arXiv.Org. https://arxiv.org/abs/2302.13971v1

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Ferrer, C. C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., … Scialom, T. (2023). *Llama 2: Open Foundation and Fine-Tuned Chat Models* (arXiv:2307.09288). arXiv. https://doi.org/10.48550/arXiv.2307.09288

Tsoumakas, G., & Katakis, I. (2007). Multi-Label Classification: An Overview. *International Journal of Data Warehousing and Mining (IJDWM)*, *3*(3), 1–13. https://doi.org/10.4018/jdwm.2007070101

Turing, A. M. (1950). I.—COMPUTING MACHINERY AND INTELLIGENCE. *Mind*, *LIX*(236), 433–460. https://doi.org/10.1093/mind/LIX.236.433

U.S. Department of Transportation. (2004). *United States Department of Transportation report to Congress on public-private partnerships, December 2004.* https://rosap.ntl.bts.gov/view/dot/28035

van den Hurk, M., & Verhoest, K. (2017). On the fast track? Using standard contracts in public–private partnerships for sports facilities: A case study. *Sport Management Review*, *20*(2), 226–239. https://doi.org/10.1016/j.smr.2016.07.004

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is All you Need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems* (Vol. 30). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

Wang, M.-T., & Chou, H.-Y. (2003). Risk Allocation and Risk Handling of Highway Projects in Taiwan. *Journal of Management in Engineering*, *19*(2), 60–68. https://doi.org/10.1061/(ASCE)0742-597X(2003)19:2(60)

Wang, N., Issa, R. R. A., & Anumba, C. J. (2022). NLP-Based Query-Answering System for Information Extraction from Building Information Models. *Journal of Computing in Civil Engineering*, *36*(3), 04022004. https://doi.org/10.1061/(ASCE)CP.1943-5487.0001019

Wang, S., Sun, Y., Xiang, Y., Wu, Z., Ding, S., Gong, W., Feng, S., Shang, J., Zhao, Y., Pang, C., Liu, J., Chen, X., Lu, Y., Liu, W., Wang, X., Bai, Y., Chen, Q., Zhao, L., Li, S., … Wang, H. (2021). *ERNIE 3.0 Titan: Exploring Larger-scale Knowledge Enhanced Pre-training for Language Understanding and Generation* (arXiv:2112.12731). arXiv. https://doi.org/10.48550/arXiv.2112.12731

Wei, J., Bosma, M., Zhao, V. Y., Guu, K., Yu, A. W., Lester, B., Du, N., Dai, A. M., & Le, Q. V. (2022). *Finetuned Language Models Are Zero-Shot Learners* (arXiv:2109.01652). arXiv. https://doi.org/10.48550/arXiv.2109.01652

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., Chi, E. H., Hashimoto, T., Vinyals, O., Liang, P., Dean, J., & Fedus, W. (2022, June 15). *Emergent Abilities of Large Language Models*. arXiv.Org. https://arxiv.org/abs/2206.07682v2

Wei, X., Cui, X., Cheng, N., Wang, X., Zhang, X., Huang, S., Xie, P., Xu, J., Chen, Y., Zhang, M., Jiang, Y., & Han, W. (2023). *Zero-Shot Information Extraction via Chatting with ChatGPT* (arXiv:2302.10205). arXiv. https://doi.org/10.48550/arXiv.2302.10205

White, J., Fu, Q., Hays, S., Sandborn, M., Olea, C., Gilbert, H., Elnashar, A., Spencer-Smith, J., & Schmidt, D. C. (2023, February 21). *A Prompt Pattern Catalog to Enhance Prompt Engineering with ChatGPT*. arXiv.Org. https://arxiv.org/abs/2302.11382v1

William, I. C., & Ashley, D. B. (1987). Impact of Various Construction Contract Clauses. *Journal of Construction Engineering and Management*, *113*(3), 501–521. https://doi.org/10.1061/(ASCE)0733-9364(1987)113:3(501)

Williamson, O. E. (1985). *The Economic Institutions of Capitalism: Firms, Markets, Relational Contracting*. New York: The Free Press.

Wimalasuriya, D. C., & Dou, D. (2010). Ontology-based information extraction: An introduction and a survey of current approaches. *Journal of Information Science*, *36*(3), 306–323. https://doi.org/10.1177/0165551509360123

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J.,

Xu, C., Le Scao, T., Gugger, S., … Rush, A. (2020). Transformers: State-of-the-Art Natural Language Processing. In Q. Liu & D. Schlangen (Eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* (pp. 38–45). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-demos.6

World Bank. (2008). *Matrix of risks distribution – roads*. PPP in Infrastructure Resource Center for Contracts, Laws and Regulations (PPPIRC) The World Bank Group, Washington, DC,. https://ppp.worldbank.org/public-private-partnership/sites/ppp.worldbank.org/files/documents/roadriskmatrix_1.pdf

Wu, C., Li, X., Guo, Y., Wang, J., Ren, Z., Wang, M., & Yang, Z. (2022). Natural language processing for smart construction: Current status and future directions. *Automation in Construction*, *134*, 104059. https://doi.org/10.1016/j.autcon.2021.104059

Xu, Y., Qiu, X., Zhou, L., & Huang, X. (2020). *Improving BERT Fine-Tuning via Self-Ensemble and Self-Distillation* (arXiv:2002.10345). arXiv. https://doi.org/10.48550/arXiv.2002.10345

Yan, H., Ma, M., Wu, Y., Fan, H., & Dong, C. (2022). Overview and analysis of the text mining applications in the construction industry. *Heliyon*, *8*(12), e12088. https://doi.org/10.1016/j.heliyon.2022.e12088

Yan, H., Yang, N., Peng, Y., & Ren, Y. (2020). Data mining in the construction industry: Present status, opportunities, and future trends. *Automation in Construction*, *119*, 103331. https://doi.org/10.1016/j.autcon.2020.103331

Yang, J., Chen, Y., Yao, H., & Zhang, B. (2022). Machine Learning–Driven Model to Analyze Particular Conditions of Contracts: A Multifunctional and Risk Perspective. *Journal of*

Management in Engineering, *38*(5), 04022036. https://doi.org/10.1061/(ASCE)ME.1943-5479.0001068

Yescombe, E. R., & Farquharson, E. (2018). *Public-Private Partnerships for infrastructure: Principles of policy and finance*. Butterworth-Heinemann.

Zhang, F. (2022). A hybrid structured deep neural network with Word2Vec for construction accident causes classification. *International Journal of Construction Management*, *22*(6), 1120–1140. https://doi.org/10.1080/15623599.2019.1683692

Zhang, J., & El-Gohary, N. M. (2015). Automated Information Transformation for Automated Regulatory Compliance Checking in Construction. *Journal of Computing in Civil Engineering*, *29*(4), B4015001. https://doi.org/10.1061/(ASCE)CP.1943-5487.0000427

Zhang, J., & El-Gohary, N. M. (2016). Semantic NLP-Based Information Extraction from Construction Regulatory Documents for Automated Compliance Checking. *Journal of Computing in Civil Engineering*, *30*(2), 04015014. https://doi.org/10.1061/(ASCE)CP.1943-5487.0000346

Zhang, L., Yao, H., Fu, Y., & Chen, Y. (2023). Comparing Subjective and Objective Measurements of Contract Complexity in Influencing Construction Project Performance: Survey versus Machine Learning. *Journal of Management in Engineering*, *39*(4), 04023017. https://doi.org/10.1061/JMENEA.MEENG-5331

Zhang, R., & El-Gohary, N. (2021). A deep neural network-based method for deep information extraction using transfer learning strategies to support automated compliance checking. *Automation in Construction*, *132*, 103834. https://doi.org/10.1016/j.autcon.2021.103834

Zhang, S., Zhang, S., Gao, Y., & Ding, X. (2016). Contractual Governance: Effects of Risk Allocation on Contractors' Cooperative Behavior in Construction Projects. *Journal of*

Construction Engineering and Management, *142*(6), 04016005. https://doi.org/10.1061/(ASCE)CO.1943-7862.0001111

Zhao, H., Chen, H., Yang, F., Liu, N., Deng, H., Cai, H., Wang, S., Yin, D., & Du, M. (2024). Explainability for Large Language Models: A Survey. *ACM Transactions on Intelligent Systems and Technology*, *15*(2), 20:1-20:38. https://doi.org/10.1145/3639372

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., Du, Y., Yang, C., Chen, Y., Chen, Z., Jiang, J., Ren, R., Li, Y., Tang, X., Liu, Z., … Wen, J.-R. (2023). *A Survey of Large Language Models* (arXiv:2303.18223). arXiv. https://doi.org/10.48550/arXiv.2303.18223

Zhong, B., Pan, X., Love, P. E. D., Sun, J., & Tao, C. (2020). Hazard analysis: A deep learning and text mining framework for accident prevention. *Advanced Engineering Informatics*, *46*, 101152. https://doi.org/10.1016/j.aei.2020.101152

Zhou, P., & El-Gohary, N. (2017). Ontology-based automated information extraction from building energy conservation codes. *Automation in Construction*, *74*, 103–117. https://doi.org/10.1016/j.autcon.2016.09.004

Zou, Y., Kiviniemi, A., & Jones, S. W. (2017). Retrieving similar cases for construction project risk management using Natural Language Processing techniques. *Automation in Construction*, *80*, 66–76. https://doi.org/10.1016/j.autcon.2017.04.003