Enhancing Large Language Models for Thai Legal Chatbots

Mr. Supachoke Hanwiboonwat

A Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of
Master of Engineering in Computer Engineering
Department of Computer Engineering
Faculty of Engineering
Chulalongkorn University
Academic Year 2024

การเพิ่มประสิทธิภาพโมเดลภาษาขนาดใหญ่สำหรับแชทบอทด้านกฎหมายภาษาไทย

นายศุภโชค หาญวิบูลย์วัฒน์

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิศวกรรมศาสตรมหาบัณฑิต
สาขาวิชาวิศวกรรมคอมพิวเตอร์ ภาควิชาวิศวกรรมคอมพิวเตอร์
คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2567

Thesis Title	Enhancing Large Language Models for Thai Legal Chatbots
By	Mr. Supachoke Hanwiboonwat
Field of Study	Computer Engineering
Thesis Advisor	Associate Professor Peerapon Vateekul, Ph.D.
Thesis Co Advisor	Apivadee Piyatumrong, Ph.D.

Accepted by the Faculty of Engineering, Chulalongkorn University in Partial Fulfillment of the Requirement for the Master of Engineering

..... Dean of the Faculty of Engineering
(Associate Professor Witaya Wannasuphoprasit, Ph.D.)

THESIS COMMITTEE

..... Chairman
(Professor Boonserm Kijsirikul, Ph.D.)

..... Thesis Advisor
(Associate Professor Peerapon Vateekul, Ph.D.)

..... Thesis Co-Advisor
(Apivadee Piyatumrong, Ph.D.)

..... Examiner
(Assistant Professor Pittipol Kantavat, Ph.D.)

..... External Examiner
(Prachya Boonkwan, Ph.D.)

ศุภโชค หาญวิบูลย์วัฒน์ : การเพิ่มประสิทธิภาพโมเดลภาษาขนาดใหญ่สำหรับแชทบอทด้านกฎหมายภาษาไทย. (Enhancing Large Language Models for Thai Legal Chatbots) อ.ที่ปรึกษาหลัก : รศ. ดร.พีรพล เวทีกูล, อ.ที่ปรึกษาร่วม : ดร.อภิวดี ปิยธรรมรงค์

ในปัจจุบัน การพัฒนาระบบตอบคำถามทางกฎหมายภาษาไทยสำหรับบุคคลทั่วไปเป็นงานที่มีความท้าทาย เนื่องจากภาษาที่ใช้ในประมวลกฎหมายมักซับซ้อนและเข้าใจยาก อีกทั้งเนื้อหายังยาวและซับซ้อนมาก งานวิจัยนี้นำเสนอระบบตอบคำถามทางกฎหมายภาษาไทยที่ออกแบบมาเพื่อบุคคลทั่วไป โดยมีเป้าหมายในการวางแนวทางที่ดีที่สุดสำหรับการพัฒนาระบบตอบคำถามทางกฎหมายอย่างมีประสิทธิภาพ เพื่อยกระดับประสิทธิภาพของระบบ เราได้สร้างชุดข้อมูลถาม-ตอบทางกฎหมายภาษาไทยขึ้นมาเอง และนำเข้าข้อมูลจากแหล่งอื่น ๆ มาร่วมด้วย จากนั้นได้ทำการทดลองเปรียบเทียบเพื่อค้นหาโมเดลภาษาที่เหมาะสมที่สุดสำหรับการใช้งานในบริบททางกฎหมายไทย รวมถึงการปรับแต่งโมเดลด้วยชุดข้อมูลหลากหลายรูปแบบเพื่อเพิ่มขีดความสามารถในงานถาม-ตอบและการทำข้อสอบทางกฎหมาย นอกจากนี้ ยังได้ทดลองใช้เทคนิคการค้นคืนข้อมูลด้วย Retrieval-Augmented Generation (RAG) โดยครอบคลุมทั้งการค้นหาด้วยคำสำคัญ การค้นหาตามบริบท และการจัดลำดับความเกี่ยวข้องของเนื้อหาในประมวลกฎหมาย เรายังได้เปรียบเทียบรูปแบบคำสั่ง (prompt) ที่แตกต่างกัน เพื่อประเมินว่ารูปแบบใดให้ผลลัพธ์ที่ดีที่สุดสำหรับการตอบคำถามทางกฎหมายแก่ประชาชนทั่วไป จากผลการทดลอง พบว่าระบบที่พัฒนาในงานวิจัยนี้มีประสิทธิภาพใกล้เคียงกับแบบจำลองขนาดใหญ่อย่าง GPT-4o ในการสอบความรู้ทางกฎหมาย และเหนือกว่าในแง่การตอบคำถามทางกฎหมายที่พบในสถานการณ์จริง โดยวัดจากค่า BERTScore และ ROUGE

สาขาวิชา วิศวกรรมคอมพิวเตอร์ ลายมือชื่อนิสิต .....

ปีการศึกษา 2567 ลายมือชื่อ อ.ที่ปรึกษาหลัก .....

ลายมือชื่อ อ.ที่ปรึกษาร่วม .....

6670246321 : MAJOR COMPUTER ENGINEERING

KEYWORD: Large Language Model, Thai Legal Documents, Question Answering

Supachoke Hanwiboonwat : Enhancing Large Language Models for Thai Legal Chatbots. Advisor: Assoc. Prof. Peerapon Vateekul, Ph.D. Co-advisor: Apivadee Piyatumrong, Ph.D.

Currently, developing a Thai legal question-answering system for the general public is highly challenging due to the complex, difficult-to-understand language and the extensive content of legal codes. This research proposes a Thai legal question-answering system designed for the public, aiming to establish best practices for developing effective legal QA systems. To improve performance, we created our own Thai legal QA dataset and incorporated data from various sources. We conducted comparative experiments to identify the most suitable language model for Thai legal contexts, and fine-tuned the models with diverse datasets for enhanced capabilities in legal QA and legal examinations. Additionally, we explored Retrieval-Augmented Generation (RAG) techniques, including keyword search, contextual search, and relevance ranking of legal documents. We also compared different prompt formats to determine which delivers the best results for answering legal questions for the general public. Our results show that the proposed system performs comparably to larger models like GPT-4o in legal knowledge exams and outperforms them in real-world legal QA tasks, as measured by BERTScore and ROUGE.

Field of Study: Computer Engineering

Academic Year: 2024

Student's Signature .....

Advisor's Signature .....

Co-advisor's Signature .....

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my advisor, Assoc. Prof. Dr. Peerapon Vateekul, for his invaluable guidance, support, and advice, all of which were crucial to the successful completion of this research. I am also deeply thankful to Ms. Apivadee Piyatumrong for her assistance in securing the necessary computing resources for this project, as well as for her advice and continuous support throughout the research process. My sincere thanks also go to Mr. Prachya Boonkwan for his guidance and ongoing assistance during this work. I would like to extend my appreciation to Mr. Chaichana Thavornthaveekul for his support in coding, reviewing experiments for appropriateness, and for his valuable input and help, which contributed significantly to the success of this research. I am also grateful to my friends in the Data Mind group, who provided help and encouragement in various aspects during my master's studies. Lastly, I would like to thank my family for their unwavering love and support, which has been the foundation that enabled me to complete this research.

Supachoke Hanwiboonwat

TABLE OF CONTENTS

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER I

INTRODUCTION

Artificial intelligence (AI) has rapidly emerged as a transformative force in the legal field [1-3], enabling innovative applications across a wide spectrum of legal processes. Notable AI-driven solutions include legal question answering (QA) systems [4-6], tools for legal judgment prediction [7], case prediction models [8], and contract review platforms [9]. The release of advanced conversational agents such as ChatGPT [10] has further accelerated progress in this domain, especially in the context of legal QA, where AI models are increasingly relied upon to address legal queries with speed and accessibility [11, 12]

Despite these advancements, significant challenges remain in tailoring legal QA systems for use by the non-experts, who often lack specialized legal knowledge. Major obstacles arise from the intricate nature of legal language, which frequently involves complex terminology and lengthy, nuanced idiomatic expressions [13]. Such linguistic challenges can impede a large language model' (LLM) ability to comprehend and accurately answer legal questions, particularly when posed by non-experts.

To address these concerns, previous research has explored the use of LLMs to generate simplified "legal stories" that translate complicated legalese into easily understandable narratives for the layperson. While these approaches have enhanced the readability of legal information, their primary focus has been on evaluating and improving the clarity and accessibility of legal content, rather than measuring the accuracy of answers to specific legal questions.

In contrast, our work is dedicated to the development of a comprehensive framework for legal question answering that is specifically designed for the non-experts. Our research aims to move beyond readability and instead centers on delivering precise, correct responses to legal queries. To this end, we have substantially enhanced the performance of LLMs, optimizing them to respond to legal questions with both efficiency and accuracy. Key to our approach is the

integration of a retrieval-augmented generation (RAG) system, meticulously adapted for the Thai legal context. This system empowers LLMs to identify and draw upon relevant legal provisions from extensive legal databases, significantly improving the relevance and correctness of their answers. Additionally, we have refined our prompt engineering strategies carefully structuring queries and instructions to guide the LLM in producing clear, user-friendly responses suitable for individuals without a legal background.

Our research is centered around the establishment of best practices for legal QA, guided by three core objectives. First, we investigate and identify the most effective LLM for answering Thai legal questions, taking into consideration the unique linguistic and structural characteristics of Thai law. Second, we focus on building a highly accurate RAG system that can efficiently extract pertinent legal codes and references from large databases. Third, we conduct an extensive comparison of various prompt formats, seeking to determine which structures and styles of questioning most effectively enhance the model's capacity to provide helpful and reliable answers for the non-experts. Moreover, recognizing the importance of diverse information sources, our study introduces new methods for improving LLM performance by leveraging datasets of varying formats, demonstrating that this multi-format strategy achieves superior results compared to single-format data utilization.

For evaluation, we rigorously compared our model against leading large language models such as GPT-4o. Our analysis considered two primary metrics: the appropriateness and correctness of the model's legal answers, and its effectiveness in solving official legal examination questions, such as those found in bar exams. Our findings reveal that our model not only provides more accurate and contextually appropriate answers than larger models like GPT-4o but also performs on par with them in formal legal exam settings. This indicates the robustness and practical applicability of our approach in real-world legal scenarios.

In summary, our research sets out to establish a best-practice framework for legal QA that is both efficient and precise, making legal information more readily accessible and understandable to Thai speakers, especially those without formal legal training. By combining advanced model selection, a tailored RAG system, and

innovative prompt engineering alongside diverse dataset utilization, we contribute a novel and effective solution to the legal AI landscape, with the goal of empowering the public to obtain reliable legal guidance in their own language.

1.1 Aims and Objectives

In this thesis, we concentrate on developing a legal question-answering framework called "Tamtanai", specifically tailored for specialized fields, with a focus on Thai law. Our objective is to create a framework that provides precise answers in complex domains, enhancing the capability to handle detailed, domain-specific interactions in Thai and establishing a new benchmark for reliability and accuracy in Thai legal question answering.

Our contributions to this endeavor are multi-faceted. First, we enhanced the RAG technique to better deliver contextually relevant answers, especially in Thai language and legal scenarios. Next, we developed a comprehensive dataset on Thai law to fine-tune LLM using the QLoRA adapter [14], to ensure the framework provides accurate legal responses. We also improved the language model, retrieval mechanism, and prompt design. Finally, we evaluated its performance against benchmarks like GPT-4o, setting new standards for reliability and accuracy in the Thai legal question-answering framework.

1.2 The Scope of Work

1. The dataset comprises Thai legal question-and-answer data created by converting Thai legal codes—such as land law, civil and criminal law, narcotics law, the revenue code, foreign labor law, commercial and consumer law, labor law, financial law, traffic law, and compensation and social security law—into a Q&A format using GPT-4. It also includes existing Thai legal Q&A from the internet and multiple-choice questions from lawyer qualification exams.

2. Develop a model to filter out queries unrelated to Thai law by utilizing the Thai legal Q&A data from this work, alongside general Thai Q&A data sourced from the internet.
3. Develop an LLM for answering legal questions in Thai, although it currently does not support drafting various types of contracts.
4. Compare and select the most effective methods and models in the areas of RAG and LLM selection.
5. Improve and assess the performance of both the RAG approach and the LLM to optimize their suitability for the Thai legal domain.
6. Compare various prompt formats used for answering legal questions in Thai to optimize the model's performance and ensure the highest efficiency in responding to Thai legal questions.

1.3 Expected Results

To evaluate the performance of our LLM against GPT-4o in legal question-answering and exam taking, we used the following metrics for contextual and semantic assessment:

1. BERTScore [15]: We aim for a higher BERTScore, indicating that our model's generated words closely match those in the ground truth.
2. ROUGE [16]: We seek a higher ROUGE score to show that our model's answers contain the correct set of words compared to the ground truth.
3. Exam scores: We expect our model to more accurately select the correct answers in multiple choice legal exams.

To evaluate the RAG system's effectiveness in retrieving the most relevant legal codes for a query, we employed the following metrics for contextual search, keyword search, and the reranking model:

1. Recall: Used to gauge the effectiveness of our contextual and keyword search methods in retrieving pertinent documents. In this thesis, precision is not reported for the entire RAG process because the ground truth comprises

only a single subsection of the legal code, which would result in a very low precision value, making it an unsuitable metric for overall assessment.

2. Top N Accuracy: Utilized to determine whether our reranking model can accurately prioritize the most relevant document to align with the ground truth.
3. MRR: To evaluate whether the reranking model can appropriately rank the documents that are relevant to the user's question.

1.4 Research Funding

We would like to extend our gratitude to the Chula Computer Engineering Graduate Scholarship for financially supporting the pursuit of a master's degree in the Department of Computer Engineering at Chulalongkorn University. Additionally, this research benefited from the support of the National Science and Technology Development Agency (NSTDA), which provided LANTA HPC computing resources.

1.5 Publication

This thesis was presented at the 30th International Conference on Natural Language & Information Systems (NLDB 2025), held in Japan from July 4th to 6th, 2025. The paper is entitled "A Comparative Study on the Development of a Thai Legal QA Framework Using Large Language Models and Mixed Legal Datasets."

CHAPTER II
BACKGROUND

This chapter will cover the key concepts related to this work, including TF-IDF, Transformer, BERT architecture, LLMs, QLoRA, RAG, and evaluation metrics.

2.1 TF-IDF

TF-IDF, or Term Frequency-Inverse Document Frequency, is a numerical statistic used to assess the significance of a word within a document relative to a collection or corpus of documents. It is widely utilized in information retrieval and text mining.

Term Frequency (TF): This measures the frequency of a term's occurrence in a document, calculated by dividing the term's occurrences by the total number of terms in that document. The premise is that a term that appears frequently is important.

Inverse Document Frequency (IDF): This evaluates a term's overall importance by factoring in the number of documents that contain the term. It is computed by dividing the total number of documents by the number of documents containing the term and then taking the logarithm of that quotient. A term that appears in many documents is considered less significant.

Together, TF-IDF assigns a score to each term in a document, facilitating the comparison of the term's importance across the entire document collection. It is extensively employed in tasks like text classification, clustering, and information retrieval to identify the most pertinent words in documents.

2.2 Transformer

Transformer architecture [17] is represents a major advancement in natural language processing technology. Central to this architecture is the robust encoder-

decoder framework, as illustrated in Figure 1, which enhances the model's capability to efficiently manage complex language tasks.

.

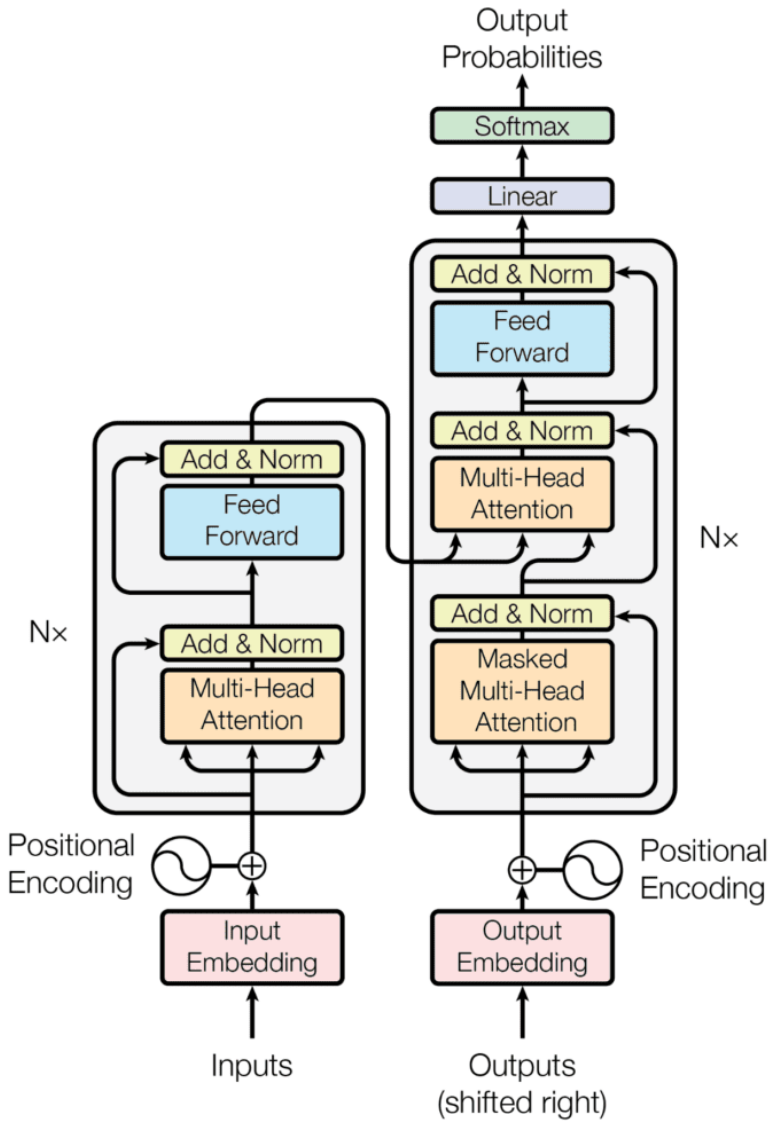

Figure 1 *Transformer architecture [17]*

2.2.1 Encoder

The encoder component of the Transformer architecture is crafted to process an input sequence and convert it into a set of continuous representations that encapsulate the entire input's contextual information. This process starts with input

embeddings, transforming each word or token into a vector representation. Positional encodings are added to indicate the order of words, helping the model understand word positioning. Central to the encoder is the multi-head self-attention mechanism, enabling it to assess the influence of all other words in the sequence for each word being processed. This captures complex dependencies and relationships within the input.

Each encoder layer combines the self-attention mechanism with a feedforward neural network for further data processing. Layer normalization and residual connections are included to stabilize and accelerate training, facilitating the effective training of deep networks. Encoders consist of stacked identical layers, each enhancing the input sequence's representation. The final output is a collection of encoded vectors that provide a comprehensive contextual summary of the input, essential for tasks like translation, where these representations aid the decoder in generating precise outputs.

2.2.2 Decoder

The decoder component of the Transformer architecture generates the output sequence from the encoded representations provided by the encoder. It processes the input to the decoder, which may include partially generated sequences or a start token, to produce coherent and contextually appropriate output. The decoder has a similar structure to the encoder but includes additional mechanisms to attend to the encoder's outputs. Each decoder layer begins with a self-attention mechanism, allowing it to focus on different parts of the generated output sequence thus far, ensuring coherence and fluency during generation.

Moreover, each decoder layer incorporates a cross-attention mechanism to focus on the encoder's outputs, enabling the model to integrate information from the entire input sequence while generating each token in the output. This cross-attention is vital for aligning language generation with the source sequence. The decoded output is refined by feedforward neural networks, like the encoder, and enhanced by layer normalization and residual connections to ensure stability during training. The final decoder layer produces a sequence of predictions or translations based on the input data, effectively converting the encoded input into a relevant

and comprehensible output sequence for tasks such as machine translation and text summarization.

2.3 BERT Architecture

BERT [18], which stands for Bidirectional Encoder Representations from Transformers, is a groundbreaking architecture in natural language processing. Developed by Google researchers in 2018, BERT is built on the Transformer model, focusing specifically on the encoder component. Its key innovation is the bidirectional training approach, allowing it to consider text input from both left-to-right and right-to-left during training. This bidirectionality enables BERT to grasp the full context of a word based on its surrounding words, resulting in representations that capture deeper linguistic nuances and meanings.

BERT's architecture includes multiple layers (typically 12 for BERT_BASE and 24 for BERT_LARGE) of transformer encoders stacked together. Each layer features self-attention mechanisms and feedforward neural networks, enabling the model to process and refine input for enhanced contextual understanding. During the pre-training phase, BERT learns from a large corpus using two primary tasks: masked language modeling (MLM) and next sentence prediction (NSP). MLM involves masking random words in a sentence and training the model to predict them, aiding in context comprehension. NSP requires predicting whether one sentence logically follows another, helping the model understand sentence relationships. BERT's architecture and training objectives allow it to excel in a variety of NLP tasks, such as question answering, sentiment analysis, and text classification, by fine-tuning the pre-trained model on specific datasets.

2.4 Large Language Models

Large language models (LLMs) have transformed artificial intelligence through advanced language processing and generation capabilities. These models leverage state-of-the-art architectures like transformers, enabling them to handle complex tasks across various domains. By using self-attention mechanisms, LLMs can assess the relevance of words in different contexts, enhancing their understanding of linguistic structures. Training involves feeding the models extensive datasets from diverse textual sources such as the internet, books, and academic papers, giving them a comprehensive grasp of syntax, semantics, and pragmatics across many languages and dialects. This diversity allows LLMs to learn and predict complex language patterns, making them adept at tasks like translation, summarization, and creative writing.

The true power of LLMs is in their massive scale, typically involving billions of parameters that encapsulate learned knowledge. During training, these parameters are refined as the model processes large volumes of text, improving the connections between words, phrases, and concepts. While larger parameter sizes enhance the model's text comprehension and generation abilities, they also require significant computational resources, posing technological and ethical challenges. The deployment of LLMs, capable of generating human-like text, must be managed carefully to avoid spreading misinformation or biases inherent in the training data. Despite these challenges, advancements in LLMs continue to promise transformative effects across various sectors, improving applications in natural language processing, automated content creation, and more, ultimately facilitating more sophisticated human-machine interactions.

2.5 Quantized Low-Rank Adaptation

Quantized Low-Rank Adaptation (QLoRA) [14] is an advanced technique that efficiently fine-tunes LLMs by integrating quantization and low-rank adaptation. This method is particularly beneficial as it allows researchers and developers to tailor pre-

trained models to specific tasks without heavy computational demands. QLoRA consists of two main components:

1. Quantization: This involves reducing the model's weight precision from 32-bit floating points to lower-bit representations (like 8-bit or 4-bit), which decreases memory usage and computational requirements. Quantization enhances resource efficiency, allowing large models to run on hardware with limited capacity.

2. Low-Rank Adaptation (LoRA): This is a fine-tuning strategy where low-rank matrices are added to the neural network's existing layers. These matrices enable the model to learn task-specific insights without significantly changing the original weights, thus retaining the knowledge from initial pre-training. Focused adjustments on these low-rank matrices minimize the number of parameters updated, further reducing computational costs.

By combining these techniques, QLoRA allows the adaptation of LLMs to specialized tasks with lower resource requirements, making it an appealing option for industries and researchers aiming to efficiently deploy powerful AI solutions.

2.6 Retrieval-Augmented Generation

Retrieval-Augmented Generation (RAG) is a method that boosts language model performance by combining document retrieval with natural language generation. It features two main components: a retriever, which searches large databases or corpora to find relevant information for a given query, and a generator, which uses this information to create informed and contextually accurate responses. By merging these processes, RAG enhances language models' ability to provide precise and relevant outputs, especially in situations requiring current or specialized knowledge. This approach is particularly useful in areas like customer service or technical support, where accessing up-to-date and comprehensive information is essential.

In this framework, document retrieval involves extracting keywords from the user query and identifying documents based on contextual similarity. The documents

obtained through these methods are then combined, and a reranking model selects the most relevant document.

2.7 Large Language Model Evaluation Metrics

2.7.1 BERTScore, 2019

BertScore [15] is a method for evaluating the quality of text summarization by assessing the similarity between the text summary and the original text.

Step 1: Contextual Embeddings - Reference and candidate sentences are represented using contextual embeddings, considering surrounding words. These embeddings are computed using models like BERT, RoBERTa, XLNet, and XLM.

Step 2: Cosine Similarity - The similarity between the contextual embeddings of the reference and candidate sentences is assessed using cosine similarity. where $\mathbf{x_i}$ indicate a word in the reference sentence and $\hat{\mathbf{x}}_\mathbf{j}$ indicate a word in the candidate sentence.

$$\text{similarity}(\mathbf{x_i}, \hat{\mathbf{x}}_\mathbf{j}) = \frac{\mathbf{x_i}^\top \hat{\mathbf{x}}_\mathbf{j}}{\|\mathbf{x_i}\| \|\hat{\mathbf{x}}_\mathbf{j}\|} \quad (1)$$

Step 3: Token Matching for Precision and Recall - Each token in the candidate sentence is aligned with the most similar token in the reference sentence, and vice versa, to compute Precision and Recall. These metrics are then combined to calculate the F1 score.

$$Precison_{\text{BERT}} = \frac{1}{|\hat{x}|} \sum_{\hat{\mathbf{x}}_\mathbf{j} \in \hat{\mathbf{x}}} \max_{\mathbf{x}_i \in x} \mathbf{x_i}^\top \hat{\mathbf{x}}_\mathbf{j} \quad (2)$$

$$Recall_{\text{BERT}} = \frac{1}{|x|}\sum_{\mathbf{x_i}\in x} \max_{\hat{\mathbf{x}}_\mathbf{j}\in \hat{x}} \mathbf{x_i}^\top \hat{\mathbf{x}}_\mathbf{j} \tag{3}$$

$$F1_{\text{BERT}} = 2 \times \frac{P_{\text{BERT}} \times R_{\text{BERT}}}{P_{\text{BERT}} + R_{\text{BERT}}} \tag{4}$$

Step 4: Importance Weighting - The significance of rare words is considered using Inverse Document Frequency (IDF), which can be integrated into BERTScore calculations. This step is optional and depends on the domain.

Step 5: Baseline Rescaling - BERTScore values are linearly adjusted to enhance human readability, ensuring they fit within a more intuitive range, based on Common Crawl monolingual datasets.

$$\hat{R}_{\text{BERT}} = \frac{R_{\text{BERT}} - b}{1 - b} \tag{5}$$

2.7.2 ROUGE Score, 2004

The ROUGE Score [16] consists of measures commonly used in text summarization tasks to automate the generation of concise summaries from longer texts. ROUGE was developed to evaluate the effectiveness of machine-generated summaries by comparing them to human-produced reference summaries. In this work, we utilize two types of ROUGE Scores: ROUGE-N and ROUGE-L.

ROUGE-N: This evaluates the similarity of n-grams, which are continuous sequences of n words, between the candidate and reference texts. It measures precision, recall, and F1-score based on the overlap of these n-grams. For example, ROUGE-1 considers individual word matches, ROUGE-2 evaluates pairs of words, and this pattern extends to larger sequences. ROUGE-N is often used to assess the grammatical accuracy and fluidity of the generated text.

$$\text{ROUGE-N} = \frac{\sum_{S \in \text{Reference Texts}} \sum_{\text{n-gram} \in S} \text{Match}(\text{n} - \text{gram})}{\sum_{S \in \text{Reference Texts}} \sum_{\text{n-gram} \in S} \text{Count}(\text{n} - \text{gram})} \quad (6)$$

ROUGE-L: This metric assesses the longest common subsequence (LCS) shared between the candidate and reference texts. It computes precision, recall, and F1-score based on the LCS length. ROUGE-L is frequently used to evaluate semantic similarity and content coverage in generated text, as it considers matching word sequences regardless of their order. ROUGE-L precision, recall, and F-measure can be calculated using Equations (7), (8), and (9). Where $LCS(X, Y)$ represents the length of the longest common subsequence between the reference text X and the candidate text Y, m is the length of the reference text and n is the length of the candidate text.

$$Precision_{lcs} = \frac{LCS(X, Y)}{n} \quad (7)$$

$$Recall_{lcs} = \frac{LCS(X, Y)}{m} \quad (8)$$

$$\beta = \frac{Precision_{lcs}}{Recall_{lcs}} \quad (9)$$

$$F_{lcs} = \frac{(1 + \beta^2) Recall_{lcs} Precision_{lcs}}{R_{lcs} + \beta^2 Precision_{lcs}} \quad (10)$$

2.8 Retrieval Augmentation Generation Evaluation Metrics

2.8.1 Recall

Recall is used to measure the ability to retrieve all relevant documents. Where Rel_{ret} denotes the number of relevant documents that were successfully

retrieved, and Rel denotes the total number of relevant documents in the collection.

$$\text{Recall} = \frac{Rel_{ret}}{Rel} \tag{11}$$

2.8.2 Top N Accuracy

Top N Accuracy is used to measure whether the correct document appears within the Top N results. Where $CorrectN$ represents the number of correct documents in the Top N results, and N represents the total number of documents in the Top N results.

$$\text{Top N Accuary} = \frac{CorrectN}{N} \tag{12}$$

2.8.3 Mean Reciprocal Rank

Mean Reciprocal Rank (MRR) is used to measure the rank position of the first relevant document, assigning higher weight to documents that appear earlier in the ranking. Where Q represents the total number of queries, and rank_i represents the rank position of the first relevant document for query i

$$\text{MRR} = \frac{1}{Q}\sum_{i=1}^{Q}\frac{1}{\text{rank}_i} \tag{13}$$

2.9 Query Classification Model Evaluation

To evaluate the query classification model, we use binary classification metrics: Precision, Recall, F1, and Accuracy. Where TP indicates both the prediction and the label are legal queries, TN signifies both the prediction and the label are

non-legal queries, FP means the model predicts a legal query, but the label is non-legal, and FN indicates the model predicts a non-legal query, but the label is legal.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{14}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{15}$$

$$\text{F1} = \frac{2 \times Precision \times \text{Recall}}{\text{Precison} + \text{Recall}} \tag{16}$$

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \tag{17}$$

CHAPTER III
RELATED WORKS

This section will discuss the various models mentioned in this work, including RAG, LLMs, query classification models and relevant work used in the experiments.

3.1 Relevant Text Embedding and Reranking Model

In this work, the RAG process includes document retrieval through keyword and contextual searches. All retrieved documents are then processed using a reranking model. The Multilingual E5 Text Embeddings model [19] is utilized for keyword searching, and the reranking model used is Bge-reranker-v2-m3 [20].

3.1.1 Multilingual E5 Text Embeddings, 2022

Released in December 2022, Multilingual E5 [19] is specifically designed for text embedding. Previously, the SentenceTransformers model called paraphrase-multilingual-mpnet-base-v2, introduced in 2019, was commonly used for multilingual embeddings in on-premises setups. However, Multilingual E5 provides a newer and more precise alternative.

The Multilingual E5 model is trained on a significantly larger and more diverse dataset known as CCPairs, which consists of cleaned text pairs from the internet. This dataset is compiled by integrating content from community Q&A platforms, CommonCrawl, and scientific papers, followed by a filtering process. Through this comprehensive and varied dataset, E5 generates more accurate and reliable embeddings.

3.1.2 Bge-reranker-v2-m3, 2024

The bge-v2-m3 [20] is a reranking model compatible with the Thai language. Unlike an embedding model, it directly outputs a similarity score by taking a question and document as input. By supplying a query and passage, you can get a relevance score, which can then be transformed into a float value between 0 and 1 using a sigmoid function.

3.2 Relevant Large Language Models

3.2.1 Openthaigpt1.5-7B-instruct, 2024

OpenThaiGPT 7b Version 1.5 [21] is an advanced Thai language chat model featuring 7 billion parameters. Based on Qwen v2.5 and released on September 30, 2024, it has been carefully fine-tuned with over 2,000,000 Thai instructional pairs, enabling it to effectively handle questions in Thai-focused domains.

3.2.2 SeaLLMs-v3-7B-Chat, 2024

Seallm3 [22] is the newest LLM in the SEALLM family, trained on linguistic data from Southeast Asia. It encompasses languages including English, Chinese, Indonesian, Vietnamese, Thai, Tagalog, Malay, Burmese, Khmer, Lao, Tamil, and Javanese. The base model for Seallm3 is QWEN2 [23].

3.2.3 Typhoon-2, 2024

Typhoon 2 [24] is a suite of LLMs developed by SCB10X. In this thesis, two variants are evaluated: typhoon2-qwen2.5-7b-instruct, which is based on Qwen-v2.5-7B [25] and contains 7 billion parameters, and llama3.1-typhoon2-8b-instruct, which builds upon llama-3.1-8B [26] with 8 billion parameters. Both models are continually pretrained using a mixture of English and Thai datasets. Furthermore, advanced post-training methods are employed to enhance their performance in Thai, while preserving the core capabilities of the original models.

3.2.4 OpenThaiLLM-Prebuilt-7B, 2024

OpenThaiLLM-Prebuilt-7B [27] is a 7-billion-parameter language model designed for both Thai and Chinese, developed on the Qwen2.5–7B base. Tailored for real-world applications, it is well-suited for a range of tasks, such as RAG, constrained text generation, and complex reasoning.

3.2.5 PathummaLLM-Text-V 1.0.0 Release, 2024

PathummaLLM-Text-V 1.0.0 [28] is a 7-billion parameter model that has been instruction-tuned from the OpenThaiLLM-Prebuilt foundation. Developed for practical deployment, it delivers performance on par with Openthaigpt1.5–7B-Instruct and is specially optimized for tasks like RAG, controlled text generation, and advanced

reasoning. The model is intended for both general users and enterprises seeking to improve multilingual support across Thai, Chinese, and English.

3.2.6 Sailor2, 2025

Sailor2 [29] is a community-driven project that delivers advanced multilingual language models for Southeast Asia, building upon Qwen 2.5 and continuously pretraining on 500 billion tokens to support 15 languages, including English, Chinese, Thai, Vietnamese, and Indonesian. Designed to meet regional needs, Sailor2 offers models in 1B, 8B, and 20B parameter sizes for both production and specialized research applications, all released under the Apache 2.0 license for broad accessibility. In this thesis, Sailor2-8B-Chat is utilized as one of the models for experimentation.

3.3 Relevant Query Classification Models

Fine-tuning an LLM can sometimes impair its ability to answer questions outside its specialized domain and may increase the risk of generating false information. Thus, having a model that can identify whether a user query pertains to the legal domain is advantageous. If the query falls outside the legal scope, employing the base model to respond can help minimize inaccuracies. In this study, the text classification models used for comparison are as follows:

3.3.1 XLM-RoBERTa, 2019

The XLM-RoBERTa [30] is an enhanced version of Facebook's RoBERTa model, introduced in 2019, designed specifically for multilingual applications. It is trained on 2.5 terabytes of filtered CommonCrawl data [31], allowing it to develop robust cross-lingual representations. This comprehensive training equips XLM-RoBERTa to perform effectively on a variety of natural language processing tasks across numerous languages, making it a highly adaptable model for multilingual contexts.

3.3.2 mpnet-base-v2, 2019

The all-mpnet-base-v2 model [32], developed by the sentence-transformers team, is designed to transform sentences and paragraphs into a 768-dimensional dense vector space. This makes it especially valuable for tasks like clustering and

semantic search. The model demonstrates strong performance across various language understanding tasks and integrates smoothly with the sentence-transformers library. As a variant of the MPNet model, it combines the strengths of BERT and XLNet, effectively capturing both bidirectional and autoregressive information.

3.3.3 WangchanBERTa, 2021

WangchanBERTa [33] was created in 2021 through a collaboration between VISTEC and PyThaiNLP. It is an encoder-only model trained with the MLM method using public data from diverse sources. This model is solely dedicated to supporting the Thai language.

3.3.4 PhayaThaiBERT, 2024

Phayathaibert [34], built on the Roberta architecture, involves training a new transformer-based model specifically for the Thai language. It has been optimized for improved understanding of code-switched language and unassimilated loanwords, and it has been trained on a larger dataset compared to WangchanBERTa.

3.4 Relevant Works

3.4.1 Legal Prompting, 2022

Legal Prompting [35] employs the IRAC (Issue, Rule, Application, Conclusion) framework to organize and structure legal responses, providing a systematic approach for analyzing and answering legal questions. This method contrasts with other prompt formats, such as Chain-of-Thought (CoT) prompting, by focusing specifically on the step-by-step reasoning process commonly used in legal analysis rather than encouraging free-form or general stepwise thinking.

3.4.2 Adapt-Retrieve-Revise, 2024

Adapt-Retrieve-Revise [6] is a pipeline designed to address legal questions through a structured three-step process. Initially, a Chinese pre-trained LLM is further trained on legal domain corpora to create a domain-adapted legal LLM. This specialized LLM generates a draft answer based on the query. In the second step, a sentence embedding model generates embedding for both the draft answer and

each paragraph in the relevant knowledge base. Evidence is retrieved by evaluating the similarity between these embeddings. In the final step, the query, draft answer, and retrieved evidence are combined into a prompt for GPT-4, which revises and constructs the final response. Figure 2 provides an overview of this methodology.

.

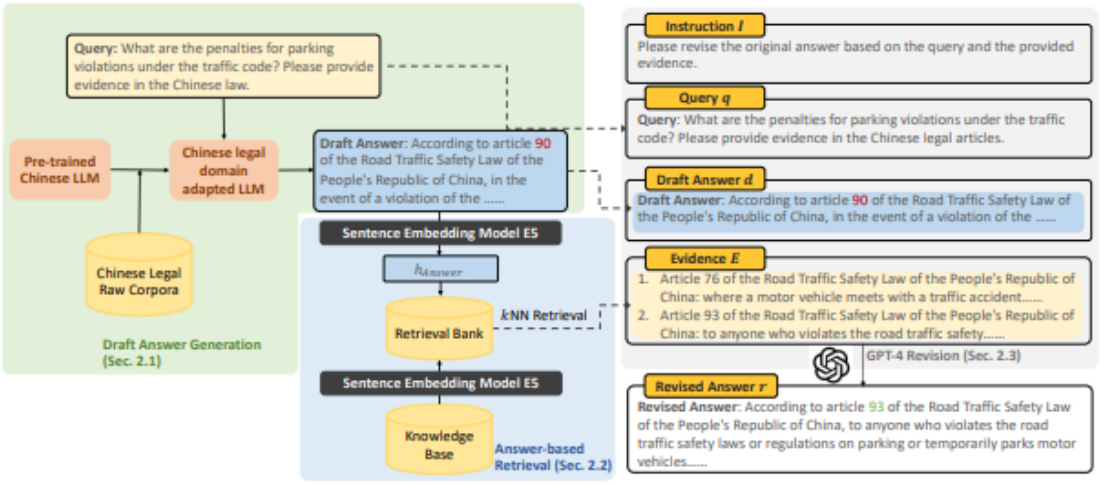

Figure 2 *Overview of Adapt-Retrieve-Revise methodology [6]*

3.4.3 The Cocktail Effect, 2024

The Cocktail Effect [36] is the phenomenon where fine-tuning LLM with a diverse combination of datasets—rather than just a single, task-specific dataset—leads to improved overall performance. By exposing the model to a wide range of data during fine-tuning, it can develop more generalizable knowledge and versatile skills, allowing it to perform better across a variety of downstream tasks. This approach leverages the strengths of rich and varied datasets to enhance the flexibility and robustness of the model.

3.4.4 SaulLM-7B, 2024

SaulLM-7B [5] is LLM specifically crafted for the legal domain. It features 7 billion parameters and is the first to be developed for interpreting and generating legal texts. Based on the Mistral 7B framework [37], SaulLM-7B has been trained on a corpus of over 30 billion English legal tokens, showcasing outstanding proficiency in understanding and managing legal documents.

3.4.5 KELLER, 2024

KELLER [8] is a reranking model that utilizes LLMs to improve the retrieval and understanding of legal cases. By incorporating expert insights from criminal law and legal statutes, KELLER allows LLMs to reorganize original case documents into clear and concise sub-facts, highlighting the essential elements of each case. This method ensures that the most pertinent legal details are emphasized for more accurate and interpretable retrieval.

CHAPTER IV
CONCEPT AND RESEARCH METHODLOGY

In this chapter, we provide a comprehensive overview of the research methodology, covering several key stages: data preparation, the overall framework, LLM experiments, RAG experiments, prompt engineering, and end-to-end evaluation.

4.1 Data Preparation

In this section, we will discuss all the datasets used in this work, focusing on the preparation of the datasets, the various features within the datasets, and the objectives for which these datasets are utilized.

4.1.1 Tamtanai Dataset

The dataset's primary source is an extensive collection of official legal documents obtained from a specialized Thai legal documentation website. We carefully selected 31 key documents frequently cited in legal consultations to ensure relevance and significance. This selection process was guided by consultations with legal experts who pinpointed the critical areas most pertinent to the Thai legal system. This targeted approach ensures that the dataset includes the essential elements necessary for effective legal assistance.

Once the core legal documents were collected, the data preparation involved several key steps to make the raw data suitable for training LLMs. Each document underwent thorough cleaning to remove irrelevant sections and standardize formatting, eliminating textual inconsistencies. We then divided the documents into smaller units based on specific subtopics, with each unit comprising a coherent idea or legal principle, which aids in more focused training and application.

To enhance the dataset further, we used the advanced language model GPT-4 to generate additional training and testing data. This synthetic data underwent a rigorous human verification process to assure its accuracy and relevance.

In summary, the Tamtanai dataset is a Thai legal question-and-answer resource featuring questions (legal-related queries), answers (responses to these questions), knowledges (the most pertinent legal articles for each query), references (the legal code containing the relevant article), and sources (the specific database file where the article can be found). This dataset is utilized for fine-tuning, evaluation, and selecting the winner in each module. The dataset contains a total of 4,534 entries, which are randomly split into 4,121 entries for training, 825 of which are further randomly selected for validation, and 413 entries for testing. Table 1 shows the distribution of queries across each law category within the Tamtanai dataset.

Table 1 *Distribution of queries across law categories in the Tamtanai dataset.*

Law Category	Training Samples	Testing Samples
ประมวลกฎหมายที่ดิน	115	14
ประมวลกฎหมายยาเสพติด	207	13
ประมวลกฎหมายวิธีพิจารณาความอาญา	259	30
ประมวลกฎหมายวิธีพิจารณาความแพ่ง	407	47
ประมวลกฎหมายอาญา	357	41
ประมวลกฎหมายแพ่งและพาณิชย์	347	31
ประมวลรัษฎากร	239	26
พระราชกำหนดการบริหารจัดการการทำงานของคนต่างด้าว	113	7
พระราชบัญญัติการขุดดินและถมดิน	45	8
พระราชบัญญัติการจัดตั้งสภาองค์กรของผู้บริโภค	14	0
พระราชบัญญัติการจัดสรรที่ดิน	64	7
พระราชบัญญัติการเช่าที่ดินเพื่อเกษตรกรรม	68	7
พระราชบัญญัติการเช่าอสังหาริมทรัพย์เพื่อพาณิชยกรรมและอุตสาหกรรม	10	0
พระราชบัญญัติการแข่งขันทางการค้า	77	6
พระราชบัญญัติขายตรงและตลาดแบบตรง	72	7
พระราชบัญญัติคุ้มครองผู้บริโภค	85	7

พระราชบัญญัติคุ้มครองแรงงาน	150	9
พระราชบัญญัติจราจรทางบก	165	20
พระราชบัญญัติประกันสังคม	121	10
พระราชบัญญัติภาษีที่ดินและสิ่งปลูกสร้าง	90	11
พระราชบัญญัติลิขสิทธิ์	105	11
พระราชบัญญัติล้มละลาย	294	30
พระราชบัญญัติว่าด้วยการกระทำความผิดเกี่ยวกับคอมพิวเตอร์	32	2
พระราชบัญญัติว่าด้วยการปรับเป็นพินัย	40	6
พระราชบัญญัติว่าด้วยข้อสัญญาที่ไม่เป็นธรรม	14	2
พระราชบัญญัติว่าด้วยราคาสินค้าและบริการ	45	5
พระราชบัญญัติศาลเยาวชนและครอบครัวและวิธีพิจารณาคดีเยาวชนและครอบครัว	157	17
พระราชบัญญัติสิทธิบัตร	115	12
พระราชบัญญัติเครื่องหมายการค้า	132	11
พระราชบัญญัติเงินทดแทน	72	7
พระราชบัญญัติแรงงานสัมพันธ์	110	9

4.1.2 WangchanX-Legal-ThaiCCL-RAG Dataset

The WangchanX-Legal-ThaiCCL-RAG dataset [38] is crafted to advance Thai legal question-answering systems using the RAG approach. It includes tailored training and test sets to enhance performance in the legal domain. "CCL" in the dataset's name stands for Corporate and Commercial Law, highlighting its focus on these essential areas of Thai legislation.

The training set features 35 legislative documents, covering a wide range of laws such as the Civil and Commercial Code, Securities and Exchange Act, and the Petroleum Income Tax Act. It particularly emphasizes finance-related laws like the Revenue Code and the Accounting Act (details in the Legislation section). Legal questions were generated from specific sections using Gemini 1.5 Pro, with relevant sections identified through the BGE-M3 model. Experts vetted these sections for

relevance and evaluated the generated questions. Answers were crafted based on these sections and reviewed using the Meta-Llama-3-70B model, with claude-3-sonnet providing Thai translations as needed. Final answers were reviewed by experts to ensure accuracy. The test set consists of expert-developed questions and answers based on 21 of the 35 major documents, with quality validated by an independent group of legal professionals for real-world applicability.

In summary, we utilized a portion of this dataset for multiple purposes. It was designed to enhance LLM performance through fine-tuning with additional legal Q&A data. Additionally, it was used for fine-tuning the reranking model and as a test dataset for end-to-end evaluation. The dataset includes features such as questions (legal-related queries), answers (corresponding responses), knowledges (relevant legal articles), and references (indicating the legal code). For LLM fine-tuning, we selected 409 entries, divided into 327 for training and 82 for validation. To tune the threshold of the reranking model, we used 4,121 entries of queries without corresponding legal codes, matching the Tamtanai training dataset. For end-to-end evaluation, 200 entries were chosen: 100 queries with legal code support and 100 without. Each dataset subset is distinct and does not overlap.

4.1.3 han-instruct-dataset-v2.0 Dataset

The Han Instruct Dataset [39], created by PyThaiNLP, this is a comprehensive Thai instruction dataset that consolidates all human- and model-developed Thai instructional datasets. It's designed for training instruction-following models, like ChatGPT, and similar applications. A large portion of the questions comes from the Reference Desk at Thai Wikipedia, adding to its diversity and usefulness for developing advanced language models.

In summary, this dataset serves as a general question-answering resource aimed at assessing the capability to determine which queries belong to the legal domain. The dataset includes features such as questions general inquiries encountered in daily life and answers providing the corresponding responses. It consists of a total of 3,200 entries, with 2,324 for training, 556 for validation, and 320 for testing.

4.1.4 Thai attorney Exam Dataset

We have collected examination questions from the past attorney exam books sold by the Lawyers Council of Thailand, covering a total of 27 years [40-42]. The collected data was thoroughly processed to eliminate duplicate questions and choices to ensure there was no overlap between the training and test datasets.

In summary, this dataset is specifically crafted as a legal exam resource to assess the ability to answer legal exam questions in Thailand. It features questions—each with four multiple choice options and answers, which indicate the correct choice for each question. The dataset contains a total of 489 entries, divided into 327 for training, 82 for validation, and 80 for testing.

4.1.5 Datasets Statistics and Summary

In this section, we will provide a detailed summary of the number and types of datasets that have been compiled for this research. Table 2 will showcase the types and names of each dataset, offering a comprehensive overview of their specific roles within the study. Additionally, Table 3 will present the distribution of all the datasets, providing a comprehensive overview of each category and its specific contributions to the study. Our discussion will cover various categories of data, each serving distinct purposes within the scope of our study. These categories include:

1. Thai Legal Question-and-Answer Data: This dataset contains structured question-and-answer pairs specifically focused on Thai legal matters. It serves as a valuable resource for understanding and interpreting Thai laws in a simplified format.
2. General Question-and-Answer Data: Alongside legal-specific data, this category encompasses a broader range of topics, providing general question-and-answer pairs that can aid in diverse applications of language models beyond the legal domain.
3. Exam Data: This dataset includes multiple-choice questions that are part of the bar examination process for obtaining attorney licensure in Thailand. It is meticulously curated to reflect the format and content of the actual licensing exams.

Table 2 *Details of datasets used in this work.*

Task	Dataset
Thai legal question answering	Tamtanai dataset, WangchanX-Legal-ThaiCCL-RAG dataset
General question answering	han-instruct-dataset-v2.0 dataset
Exam	Thai attorney exam dataset

Table 3 *Distribution of datasets used in this work.*

Dataset	Train	Validate	Test	Total
Tamtanai dataset	3,296	825	413	4,534
WangchanX-Legal-ThaiCCL-RAG dataset	327	82	0	409
han-instruct-dataset-v2.0 dataset	2,324	556	320	3,200
Thai attorney exam dataset	327	82	80	489

4.1.6 Legal Document Retrieval-Augmented Generation Database

After selecting 31 legal codes relevant to everyday life for inclusion in the Tamtanai dataset, we supplemented them with 5 additional codes commonly used in legal examinations. We then refined these 36 legal codes by removing outdated or repealed articles. Prior to integrating this material into the RAG database, we segmented the documents into subsections according to distinct subtopics, with each subsection forming a chunk that typically corresponds to a chapter in the original legal code. Each chapter, in turn, comprises multiple sections that address a single subject area. In total, this process yielded 617 chunks. Furthermore, we stored the processed data together with metadata specifying the legal sections present in each chunk and identifying which provisions remain currently active.

4.2 Overall Process of This Work

An overview of the framework is illustrated in Figure 3. The process begins by determining whether a user query falls within the legal domain. If the query is not related to legal topics, a base model generates the response. If the query is legal in

nature, it proceeds through the RAG pipeline. To optimize retrieval effectiveness, we employ a two-stage approach, utilizing both keyword and contextual search, followed by a reranking step. During keyword search, key terms are extracted from the query while low-IDF terms are excluded to minimize the retrieval of irrelevant documents. The refined keywords are then used to search for relevant documents. For contextual search, the system retrieves documents with a similarity score meeting or exceeding a predefined threshold. Results from both keyword and contextual searches are merged and passed through a reranking model to prioritize the most relevant documents.

Next, we evaluate whether the probabilities assigned by the reranking model to the retrieved documents exceed a predefined threshold. The threshold is determined by using queries from the WangchanX-Legal-ThaiCCL-RAG dataset that do not have any corresponding legal codes in our database. We conduct retrieval experiments with these queries and select the threshold that provides the highest accuracy in distinguishing between queries that do and do not have corresponding legal codes in the database. This step serves to determine whether relevant legal codes exist in the database for the user's query. If the probability does not meet the threshold, the system will respond that there are no relevant legal codes available in the database. However, if the probability surpasses the threshold, the reranked documents are incorporated into a prompt template, which is then used to guide the LLM in generating the final response.

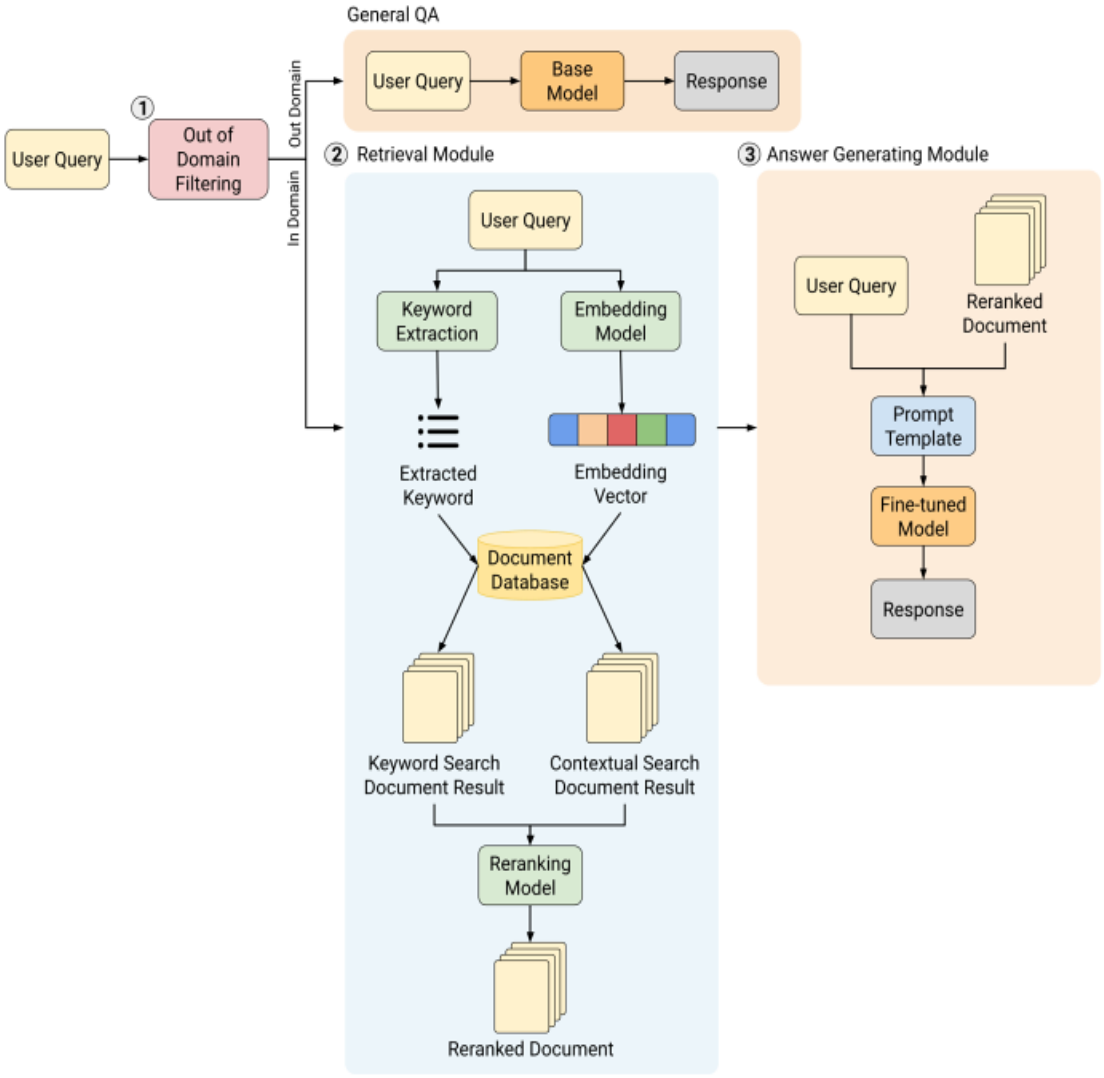

Figure 3 *The overall process diagram of this work*

4.3 Large Language Model and Query Classification Model Selection

To select the most appropriate LLM for our experiments, we focused on models that support the Thai language and have between 7B and 8B parameters. These models were fine-tuned and evaluated on the Tamtanai dataset, with the model achieving the highest BERTScore chosen as the best performer. For the query classification task, specifically Out of Domain Query Filtering, we fine-tuned a classification model to distinguish between legal and non-legal queries using two labels: 0 for non-legal domain queries and 1 for legal domain queries. The Tamtanai

dataset, which consists of legal domain question-answer pairs, was used for label 1. For non-legal queries, we utilized the Han Instruct v2.0 dataset, leveraging GPT-4o to assist in classifying the queries. Queries determined to be within the legal domain by GPT-4o were labeled as 1, and those outside the legal domain were labeled as 0. We also conducted a random manual review to ensure the accuracy and appropriateness of the labels generated by GPT-4o. The fine-tuning of Query Classification Model was conducted using 10 epochs, a learning rate of 2×10^{-5}, and a batch size of 8 per device for both training and evaluation phases.

4.4 Retrieval-Augmented Generation System

In the contextual search phase, we first identified the most suitable contextual search model based on the MRR metric. For the reranking process, we utilized the bge-reranker-v2-m3 model, which was fine-tuned using the Tamtanai dataset by designating each question as a query, the most relevant legal document as the positive example, and five similar yet non-identical documents retrieved via contextual search as negative examples. To determine the most effective retrieval approach, we compared the accuracy of several methods: (1) contextual search, (2) keyword search combined with reranking, (3) contextual search combined with reranking, and (4) a combination of keyword search, contextual search, and reranking. However, we did not report results for keyword search alone, as it does not produce a clear ranking documents that contain the same number of matching keywords are scored equally, regardless of actual relevance necessitating the use of a subsequent ranking step.

4.5 Enhancing Large Language Model

In this thesis, we fine-tuned LLM using two types of datasets: a legal question-and-answer dataset and a legal multiple-choice exam dataset. The primary dataset for fine-tuning was Tamtanai, with additional experiments conducted using multitask fine-tuning to evaluate whether combining datasets improves question-answering

performance compared to single-task fine-tuning on similar datasets. Specifically, we used the Thai attorney exam dataset for multitask learning, and we used the portion of the WangchanX-Legal-ThaiCCL-RAG dataset where the queries correspond to legal codes in the training dataset as a comparable single-task dataset. The QLoRA [14] framework was employed for model fine-tuning. Training was conducted over 3 epochs with a per-device training batch size of 3 and an evaluation batch size of 4. The optimizer selected was PagedAdamW32bit [43], set with a learning rate of 2.5e-4, and the temperature parameter was fixed at 0.1.

4.6 Prompt Engineering

In this study, our objective is to determine which prompt format is most effective for answering legal questions intended for the general public, as well as which is best suited for legal examination settings. To achieve this, we compare several prompt types: the standard instruction prompt, the CoT prompt, and the IRAC prompt [35] [44]. The CoT prompt guides the model to first examine the subsections and key elements of the relevant legal code before generating a response. In contrast, the IRAC prompt employs a structured legal reasoning framework, where I stands for Issue (identifying the core legal issue), R for Rule (reviewing and applying relevant principles or laws), A for Application/Analysis (analyzing and applying those rules to the specific case, considering relevant subsections), and C for Conclusion (summarizing the answer based on the applicable legal principles).

4.7 End to End Evaluation

In our end-to-end evaluation, after integrating all modules, we analyze how many legal documents selected by the reranking model should be included in the prompt to optimize the LLM's performance in answering legal questions. Additionally, we constructed four new test datasets to comprehensively evaluate our

system. The details of these new test datasets are provided in Table 4, and the process for creating each test dataset is described as follows.

1. New test dataset 1**:** This dataset is designed to evaluate end-to-end system performance and the RAG module. It is constructed from legal articles previously used in earlier test datasets, comprising a total of 413 rows. The number of relevant legal codes per query matches that of the original test datasets. Generation was carried out using GPT-4o. This test dataset will also be used to benchmark our system against GPT-4o, enabling an assessment of how our system performs on data that has never been tuned within this framework.
2. New test dataset 2**:** This dataset is designed for both end-to-end and RAG evaluation, consisting of legal codes from the database but specifically including articles that have not been used in previous training or test datasets. The dataset contains 100 rows for each label, totaling 200 rows, and was generated using GPT-4o. Label 0 consists of queries generated from articles that have never appeared in any prior datasets, while label 1 includes queries generated from articles that were used in previous datasets; these 100 label 1 rows are randomly sampled from new test dataset 1.
3. New test dataset 3**:** This dataset is designed to evaluate the RAG module's performance when processing queries related to legal codes that may or may not exist in the database. The dataset consists of 300 rows in total, with 100 queries labeled 0 and 200 queries labeled 1. Label 0 includes queries regarding legal codes that are not found in the database, with all 100 rows sourced from the WangchanX-Legal-ThaiCCL-RAG dataset and selected specifically to exclude any entries previously used in earlier experiments. For label 1, which includes queries about legal codes present in the database, 100 rows are also taken from the WangchanX-Legal-ThaiCCL-RAG dataset and an additional 100 rows are randomly selected from new test dataset 1.
4. New test dataset 4**:** This dataset is created to evaluate the performance of the query classification model. It consists of 400 queries in total, with 200 queries for each label. Label 0 comprises queries related to legal matters,

while label 1 includes queries unrelated to law. For each label, the dataset includes a mix of both simple and difficult queries to assess the model's ability to distinguish between legal and non-legal questions. All queries were generated using GPT-4o.

Table 4 *Details of all new test datasets.*

Dataset	Task	Samples
New test dataset 1	End to end and RAG evaluation	413
New test dataset 2	End to end and RAG evaluation	200
New test dataset 3	RAG evaluation	300
New test dataset 4	Query classification evaluation	400

CHAPTER V
RESULTS

This chapter presents the results from a comprehensive series of experiments conducted in this thesis, encompassing key components in the development and evaluation of a legal question answering system. These include the selection and comparison of large language models, evaluation of query classification and contextual search models, and fine-tuning the reranking model for improved answer relevance. The chapter also details the outcomes of retrieval-augmented generation techniques, enhancements to large language models, and the analysis of question answering performance across legal categories. Additionally, results from prompt optimization, end-to-end question answering evaluation, and assessment with a new test dataset are discussed, in that order, to provide thorough insight into each stage of system development.

5.1 Results of Large Language Model Selection

This experiment seeks to identify the most suitable LLM for the answer-generating module in Figure 3. As shown in Table 5, the typhoon2-qwen2.5-7b-instruct model consistently outperforms other LLMs that support Thai when trained and evaluated on the Tamtanai dataset. It not only achieves the highest BERTScore but is also based on Qwen2.5, whose Technical Report [25] demonstrates superior performance compared to Qwen-2 and Llama 3.1 across a range of tasks. Additionally, the Technical Report for typhoon2-qwen2.5-7b-instruct [24] highlights improvements gained through fine-tuning with long context prompts, which enables the model to analyze lengthy legal documents effectively. For these reasons, typhoon2-qwen2.5-7b-instruct was selected as the primary LLM for our experiments.

Table 5 *BERTScore F1 for each model when fine-tuned on the Tamtanai training dataset and tested on the Tamtanai test dataset. Bold values indicate the winner.*

Model	BERTScore F1 (↑)
typhoon2-qwen2.5-7b-instruct	**0.8979**
SeaLLMs-v3-7B-Chat	0.8965
Openthaigpt1.5-7B-instruct	0.8960
OpenThaiLLM-Prebuilt-7B	0.8912
Pathumma-llm-text-1.0.0	0.8900
Sailor2-8B-Chat	0.8898
llama3.1-typhoon2-8b-instruct	0.8888

5.2 Results of Query Classification Model Selection

This experiment aims to determine the most appropriate query classification model for the retrieval module shown in Figure 3. According to Table 6, paraphrase-multilingual-mpnet-base-v2 is selected as the main out of domain query filtering Model for this study, as it achieved the highest precision, recall, F1 score, and accuracy among the four models evaluated.

Table 6 *Precision, recall, F1 score, and accuracy for each model are reported after fine-tuning on the Tamtanai and han-instruct-dataset-v2.0 training sets and evaluated on the Tamtanai and han-instruct-dataset-v2.0 test sets. Bold values indicate the winner.*

Model	Precision (↑)	Recall (↑)	F1 (↑)	Accuracy (%) (↑)
paraphrase-multilingual-mpnet-base-v2	**0.9954**	0.9908	**0.9931**	**99.18**
wangchanberta-base-att-spm-uncased	0.9908	0.9908	0.9908	98.91
phayathaibert	0.9908	0.9908	0.9908	98.91
xlm-roberta-base	0.9863	**0.9954**	0.9908	98.91

5.3 Results of Contextual Search Model Selection

This experiment aims to identify the most effective contextual search model for use in the retrieval module illustrated in Figure 3. Table 7 displays the MRR for each model on the Tamtanai test set, offering a straightforward comparison of their retrieval performance. Among all evaluated models, multilingual-e5-large achieved the highest MRR, demonstrating a superior capability to accurately rank and retrieve relevant documents. This indicates that multilingual-e5-large excels at distinguishing relevant texts from irrelevant ones, especially in the context of diverse legal queries. Accordingly, we selected multilingual-e5-large as the primary contextual search model for our study. Additionally, to maximize retrieval accuracy, we fine-tuned its similarity threshold to 0.55 and set the top-k parameter to 50, ensuring that the most relevant documents are reliably retrieved.

Table 7 *MRR and execution time (seconds) for each model tested on the Tamtanai test dataset. Bold values indicate the winner.*

Model	MRR (↑)	Time (s) (↓)
multilingual-e5-large	**0.7079**	0.0152
multilingual-e5-large-instruct	0.6719	0.0157
multilingual-e5-base	0.6345	0.0096
multilingual-e5-small	0.6348	**0.0095**

5.4 Results of Fine Tuning Reranking Model

This experiment aims to further enhance the performance of the reranking model within the retrieval module shown in Figure 3. Table 8 demonstrates that, after fine-tuning on the Tamtanai test dataset, the Top-1 Accuracy improved by 12.49% over the pre-fine-tuning result. This notable increase underscores the effectiveness of fine-tuning in boosting the model's ability to accurately rank relevant legal documents and improve overall retrieval outcomes.

Subsequently, we configured a threshold for the reranking model to determine whether a given query is supported by any legal code in our system. To set this threshold, we utilized queries from the WangchanX-Legal-ThaiCCL-RAG dataset that do not correspond to any legal codes in our database and selected the threshold value that maximized accuracy in distinguishing between queries with and without matching legal codes in the database.

Table 8 *Top-1 accuracy for both base model and fine-tuned model tested on the Tamtanai test dataset. Bold values indicate the winner.*

Model	Top-1 Accuracy (↑)
Base Model	0.8198
Fine-Tuned Model	**0.9222**

5.5 Results of Retrieval-Augmented Generation Techniques

This experiment seeks to compare different RAG methods in the legal domain to determine the most effective approach for the retrieval module depicted in Figure 3. As presented in Table 9, combining keyword search, contextual search, and reranking yields the highest performance in retrieving documents related to the code of law. This is because each method brings distinct advantages: keyword search excels at locating exact terms, contextual search interprets the underlying meaning of queries, and reranking further improves relevance by refining the results.

By integrating these methods, the search process becomes more comprehensive—keyword search identifies a broad range of relevant candidates, contextual search assesses semantic relevance, and reranking selects the most appropriate documents. This layered approach is particularly beneficial for legal documents, which often contain complex language and domain-specific terminology. As a result, it ensures accurate term matching while also considering the legal context, effectively minimizing the retrieval of irrelevant materials.

Table 9 *Retrieval size, recall, top-1, 2, 5, 10 accuracy, MRR, and retrieval time (seconds) for each RAG methods tested on the Tamtanai test dataset. Bold values indicate the winner.*

Metric	Context	Keyword + Reranking	Context + Reranking	**Keyword + Context + Reranking**
Retrieval size (↓)	**44.68**	177.85	**44.68**	195.42
Recall (↑)	0.9782	0.9709	0.9782	**1.0000**
Top-1 accuracy (↑)	0.6053	0.9370	0.9540	**0.9637**
Top-2 accuracy (↑)	0.7094	0.9370	0.9685	**0.9879**
Top-5 accuracy (↑)	0.8402	0.9661	0.9758	**0.9952**
Top-10 accuracy (↑)	0.9056	0.9709	0.9782	**1.0000**
MRR (↑)	0.7064	0.9525	0.9633	**0.9784**
Retrieval time (s) (↓)	**0.0143**	0.3161	0.1246	0.2906

5.6 Results of Large Language Model Enhancement

This experiment systematically evaluates different fine-tuning strategies to improve LLM performance in the answer-generating module shown in Figure 3, with a particular focus on legal question-answering and legal exam tasks. In Table 10, “LLM” refers to the typhoon2-qwen2.5-7b-instruct model fine-tuned using the Tamtanai training dataset. “QA” indicates additional fine-tuning with the WangchanX-LegalThaiCCL-RAG dataset, while “EX” denotes fine-tuning with the Thai attorney exam dataset. The (H) label signifies the use of half of that respective dataset.

As detailed in Table 10, the findings reveal that fine-tuning the model with a combination of both QA and multiple-choice exam datasets consistently outperforms fine-tuning with the QA dataset alone. Incorporating legal exam questions during training leads to marked improvements in the model’s performance on both QA and exam-based tasks—an advantage that surpasses merely expanding the QA dataset. Even when the QA and exam datasets are combined in equal

proportions, the model still outperforms the QA-only setup, though its performance is slightly below scenarios where the exam data is further increased. This indicates that training with exam-specific questions contributes valuable knowledge for tackling a range of legal benchmarks.

Overall, our results confirm that the most effective approach is to fine-tune on both types of datasets: the resulting LLM not only surpasses GPT-4o on legal QA tasks but also matches its performance on legal exam questions, underscoring the importance of employing diverse, task-specific data for optimal domain adaptation.

Table 10 *BERTScore F1, ROUGE Scores, and multiple-choice exam scores for each model tested on the Tamtanai test dataset (2nd – 5th columns) and the Thai Attorney Exam test dataset (last column). Bold values indicate the winner.*

Model	BERTScore F1 (↑)	Rouge-1 (↑)	Rouge-2 (↑)	Rouge-L (↑)	Exam scores (↑)
GPT-4o	0.8591	0.6590	0.6037	0.6256	**47/80**
LLM	0.8979	0.7719	0.7231	0.7402	34/80
LLM+EX	0.9159	0.8201	0.7717	0.7852	38/80
LLM+QA	0.9133	0.8076	0.7693	0.7820	31/80
LLM+QA+EX	**0.9367**	**0.8589**	**0.8273**	**0.8376**	46/80
LLM+QA(H)+EX(H)	0.9137	0.8109	0.7719	0.7853	36/80

5.7 Results of Question Answering in Each Law Category

This experiment aims to assess the capability of the best-performing model from Experiment 5.6 in answering questions from various legal categories. As shown in Table 11, the Unfair Contract Terms Act category (พระราชบัญญัติว่าด้วยข้อสัญญาที่ไม่เป็นธรรม) achieves the highest average BERTScore F1. This is because the statute employs simpler language and the chunk sizes contain fewer words. For these reasons, the LLM can answer questions related to this law more effectively.

Table 11 *Statistical metrics of BERTScore F1 for each law category, measured on the Tamtanai test dataset. Bold values indicate the winner.*

Law Category	Min (↑)	Max (↑)	Mean (↑)	Median (↑)	Variance (↓)
ประมวลกฎหมายที่ดิน	0.8433	1	0.9524	1	0.0041
ประมวลกฎหมายยาเสพติด	0.7926	1	0.9433	0.9724	0.0044
ประมวลกฎหมายวิธีพิจารณาความอาญา	0.7617	1	0.9210	0.9424	0.0068
ประมวลกฎหมายวิธีพิจารณาความแพ่ง	0.7559	1	0.9326	0.9679	0.0056
ประมวลกฎหมายอาญา	0.7826	1	0.9322	0.9742	0.0055
ประมวลกฎหมายแพ่งและพาณิชย์	0.7838	1	0.9495	0.9746	0.0045
ประมวลรัษฎากร	0.6767	1	0.8874	0.9070	0.0106
พระราชกำหนดการบริหารจัดการการทำงานของคนต่างด้าว	0.7956	1	0.9315	0.9530	0.0059
พระราชบัญญัติการขุดดินและถมดิน	0.7297	1	0.8998	0.9144	0.0081
พระราชบัญญัติการจัดสรรที่ดิน	0.7671	1	0.9129	0.9512	0.0078
พระราชบัญญัติการเช่าที่ดินเพื่อเกษตรกรรม	0.8487	0.9996	0.9304	0.9247	0.0036
พระราชบัญญัติการแข่งขันทางการค้า	0.8979	1	0.9459	0.9322	0.0019
พระราชบัญญัติขายตรงและตลาดแบบตรง	0.8412	1	0.9473	0.9649	0.0036
พระราชบัญญัติคุ้มครองผู้บริโภค	0.8115	1	0.9430	0.9772	0.0051
พระราชบัญญัติคุ้มครองแรงงาน	0.7685	1	0.9382	0.9683	0.0062
พระราชบัญญัติจราจรทางบก	0.7530	1	0.9284	0.9700	0.0071
พระราชบัญญัติประกันสังคม	0.7804	1	0.9363	0.9591	0.0059
พระราชบัญญัติภาษีที่ดินและสิ่งปลูกสร้าง	0.8087	1	0.9442	0.9874	0.0055
พระราชบัญญัติลิขสิทธิ์	0.8749	1	0.9674	0.9967	0.0020

พระราชบัญญัติล้มละลาย	0.7649	1	0.9164	0.9278	0.0063
พระราชบัญญัติว่าด้วยการกระทำความผิดเกี่ยวกับคอมพิวเตอร์	0.7882	0.9627	0.8755	0.8755	0.0152
พระราชบัญญัติว่าด้วยการปรับเป็นพินัย	0.7530	1	0.9100	0.9328	0.0111
พระราชบัญญัติว่าด้วยข้อสัญญาที่ไม่เป็นธรรม	0.9551	1	**0.9776**	0.9776	0.0010
พระราชบัญญัติว่าด้วยราคาสินค้าและบริการ	0.8970	1	0.9518	0.9701	0.0024
พระราชบัญญัติศาลเยาวชนและครอบครัวและวิธีพิจารณาคดีเยาวชนและครอบครัว	0.7867	1	0.9399	0.9861	0.0059
พระราชบัญญัติสิทธิบัตร	0.8373	1	0.9635	0.9864	0.0029
พระราชบัญญัติเครื่องหมายการค้า	0.8294	1	0.9485	0.9893	0.0049
พระราชบัญญัติเงินทดแทน	0.8620	1	0.9568	1	0.0035
พระราชบัญญัติแรงงานสัมพันธ์	0.8225	1	0.9516	1	0.0043

5.8 Results of Prompt Optimization

The purpose of this experiment is to determine which prompt template works best for answering legal questions for the general public using the answer-generation module depicted in Figure 3. Table 12 shows that a standard instruction prompt delivers the most effective results. This can be attributed to the Tamtanai dataset, which is tailored to address general legal queries rather than following structured formats like IRAC or CoT. As a result, when the model is fine-tuned with the Tamtanai dataset, generating answers with a regular instruction prompt leads to better performance compared to prompts that focus on legal reasoning.

Moreover, the standard instruction prompt is also the most time-efficient, likely because it is less complex and requires fewer computational resources, resulting in quicker response times. Table 13 indicates that, while the CoT prompt

achieves the highest efficiency in simulated legal exam scenarios, the performance difference among the three prompt types is minor and statistically insignificant. This is probably due to the inherent randomness of LLM inferences, which can cause slight fluctuations in outcomes. Therefore, any of the prompt formats can be used for legal exams without a significant difference in effectiveness. However, in situations where processing speed is a priority, the standard instruction prompt stands out as the most practical choice thanks to its straightforwardness and rapid execution.

Table 12 *BERTScore F1, ROUGE scores, and response time (seconds) for testing with different prompt formats on the Tamtanai test dataset. Bold values indicate the winner.*

Prompt	BERTScore F1 (↑)	ROUGE-1 (↑)	ROUGE-2 (↑)	ROUGE-L (↑)	Time (s) (↓)
Normal	**0.9367**	**0.8589**	**0.8273**	**0.8376**	**37.41**
COT	0.9336	0.8541	0.8229	0.8345	37.95
IRAC	0.9346	0.8535	0.8233	0.8333	38.03

Table 13 *Multiple-choice exam scores and response time (seconds) for testing with different prompt formats on the Thai Attorney Exam test dataset. Bold values indicate the winner.*

Prompt	Exam Scores (↑)	Time (s) (↓)
Normal	46/80	**37.74**
COT	**47/80**	38.50
IRAC	45/80	38.93

5.9 End-to-End Question and Answering Evaluation

This experiment aims to identify the ideal number of legal documents, chosen by the reranking model, to supply to the LLM for optimal end-to-end legal question answering. After incorporating the most effective RAG strategy, top-

performing LLM, and best prompt template into our legal QA system, the results in Table 14 reveal that providing just a single legal document leads to better QA outcomes than supplying multiple documents. This improvement likely stems from the fact that giving the LLM too much legal information at once can create confusion about which specific provision to use, thereby reducing the system's accuracy and consistency in answering legal questions.

Table 14 *BERTScore F1, ROUGE scores, and response time (seconds) for end-to-end evaluation tested on the Tamtanai test dataset. Bold values indicate the winner.*

Number of documents	BERTScore F1 (↑)	ROUGE-1 (↑)	ROUGE-2 (↑)	ROUGE-L (↑)	Time (s) (↓)
1	**0.9052**	**0.7818**	**0.7398**	**0.7558**	**41.89**
2	0.8924	0.7481	0.7002	0.7155	42.16
5	0.8681	0.6867	0.6252	0.6453	48.15
10	0.8487	0.6221	0.5378	0.5669	70.36

5.10 Results from the Evaluation of the New Test Dataset

This experiment aims to evaluate the performance of various system components using test datasets that have not been included in previous experiments. As shown in Table 15, end-to-end evaluation with new test dataset 1 produces a lower BERTScore F1 compared to the Tamtanai test dataset. This is because the LLM was trained on data that closely resembles the Tamtanai dataset, so when it is tested with differently worded queries, as in new test dataset 1, its BERTScore F1 decreases due to unfamiliarity with the phrasing.

Table 16 compares our framework with GPT-4o, using new test dataset 1 as the benchmark. The results show that our framework achieves a higher BERTScore F1 than GPT-4o, demonstrating more effective question-answering performance. However, in terms of response time, our system is currently slower because time

optimization for our LLM has not yet been implemented, unlike GPT-4o which responds more quickly.

As illustrated in Table 17, the Top-1 accuracy measured with new test dataset 1 is lower than with the Tamtanai test dataset, since the reranking model is not familiar with the different wording in new test dataset 1 and thus performs less effective retrieval. Table 18 then evaluates RAG retrieval performance using new test dataset 2, revealing that the system retrieves queries referencing articles seen during training much more effectively than those from unseen articles, highlighting the importance of prior exposure in the training data. Similarly, Table 19 shows that end-to-end performance with new test dataset 2 yields a higher BERTScore F1 for queries related to articles included in the training set, as opposed to novel articles.

Finally, Table 20 assesses the RAG system's ability to distinguish between queries supported and not supported by legal codes in the database using new test dataset 3. The overall classification accuracy is 80.33%, with 89.50% accuracy for queries covered by legal codes and 62.00% for those not covered. Table 21, which evaluates legal-domain query classification using new test dataset 4, reports an overall accuracy of 85.50%, with 92.50% accuracy for legal-domain queries and 78.50% for non-legal queries.

Table 15 *BERTScore F1, ROUGE scores, and response time (seconds) for end-to-end evaluation tested on new test dataset 1 and the Tamtanai test dataset. Bold values indicate the winner.*

Dataset	BERTScore F1 (↑)	ROUGE-1 (↑)	ROUGE-2 (↑)	ROUGE-L (↑)	Time (s) (↓)
New test dataset 1	0.8669	0.6683	0.6025	0.6276	41.90
Tamtanai test dataset	**0.9052**	**0.7818**	**0.7398**	**0.7558**	**41.89**

Table 16 *BERTScore F1, ROUGE scores, and response time (seconds) for end-to-end evaluation tested on new test dataset 1, comparing GPT-4o and our framework. Bold values indicate the winner.*

Model	BERTScore F1 (↑)	ROUGE-1 (↑)	ROUGE-2 (↑)	ROUGE-L (↑)	Time (s) (↓)
Our framework	**0.8669**	**0.6683**	0.6025	0.6276	41.90
GPT-4o	0.8625	0.6600	**0.6176**	**0.6327**	**1.59**

Table 17 *Retrieval size, recall, top-1, 2, 5, 10 accuracy, MRR, and retrieval time (seconds) when tested on new test dataset 1 and the Tamtanai test dataset. Bold values indicate the winner.*

Metric	New Test Dataset 1	**Tamtanai Test Dataset**
Retrieval size (↓)	**169.09**	195.42
Recall (↑)	0.9927	**1**
Top-1 accuracy (↑)	0.8015	**0.9637**
Top-2 accuracy (↑)	0.8765	**0.9879**
Top-5 accuracy (↑)	0.9274	**0.9952**
Top-10 accuracy (↑)	0.9564	**1**
MRR (↑)	0.8594	**0.9784**
Retrieval time (s) (↓)	**0.2597**	0.2906

Table 18 *Retrieval size, recall, top-1, 2, 5, 10 accuracy, MRR, and retrieval time (seconds) when tested on new test dataset 2. Bold values indicate the winner.*

Metric	Unseen Article Queries	**Seen Article Queries**	Overall Average
Retrieval size (↓)	**144.78**	176.06	160.42
Recall (↑)	0.9400	**1**	0.9700
Top-1 accuracy (↑)	0.7000	**0.8200**	0.7600
Top-2 accuracy (↑)	0.7600	**0.8600**	0.8100
Top-5 accuracy (↑)	0.8700	**0.9200**	0.8950
Top-10 accuracy (↑)	0.9000	**0.9600**	0.9300
MRR (↑)	0.7668	**0.8632**	0.8150
Retrieval time (s) (↓)	**0.2372**	0.2656	0.2514

Table 19 *BERTScore F1, ROUGE scores, and response time (seconds) for end-to-end evaluation tested on the new test dataset 2. Bold values indicate the winner.*

Query Type	BERTScore F1 (↑)	ROUGE-1 (↑)	ROUGE-2 (↑)	ROUGE-L (↑)	Time (s) (↓)
Unseen article queries	0.8427	0.6090	0.5123	0.5437	42.34
Seen article queries	**0.8717**	**0.6889**	**0.6315**	**0.6573**	**42.14**
Overall average	0.8572	0.6490	0.5719	0.6005	42.24

Table 20 *Precision, recall, F1 score, and accuracy for legal document support detection performance of the RAG system tested on new test dataset 3. Bold values indicate the winner.*

Query Type	Precision (↑)	Recall (↑)	F1 (↑)	Accuracy (%) (↑)
Queries without legal documents	0.7470	0.6200	0.6776	62.00
Queries with legal documents	**0.8249**	**0.8950**	**0.8585**	**89.50**
Overall average	0.7989	0.8033	0.7982	80.33

Table 21 *Precision, recall, F1 score, and accuracy for out-of-domain query classification performance tested on new test dataset 4. Bold values indicate the winner.*

Query Type	Precision (↑)	Recall (↑)	F1 (↑)	Accuracy (%) (↑)
Out of legal domain queries	**0.9128**	0.7850	0.8441	78.50
In legal domain queries	0.8114	**0.9250**	**0.8645**	**92.50**
Overall average	0.8621	0.8550	0.8543	85.50

5.11 Discussion

In this thesis, the LLMs used in the experiments are not particularly large, which results in suboptimal performance when answering questions related to certain legal codes. According to the paper “Lost in the Middle: How Language Models Use Long Contexts”, [45] when longer contexts are included in the prompt, LLMs tend to remember only the information at the beginning and the end, while forgetting the content in the middle. This leads to poorer question-answering performance. These findings are consistent with Experiment 5.9 in this thesis, which shows that the more legal code documents are provided in the prompt, the lower the accuracy in answering questions becomes.

CHAPTER VI
CONCLUSION

This research presents a Thai legal question-answering (QA) framework for the general public. We evaluated several large language models (LLMs) and enhanced performance using a multitask dataset. For the retrieval-augmented generation (RAG) component, we optimized prompt templates, retrieval strategies, and document count. Experiments on the Thai Attorney Exam and our dataset showed that Typhoon2-qwen2.5-7b-instruct achieved the highest BERTScore. We used paraphrase-multilingual-mpnet-base-v2 to filter non-legal queries, which were then handled by the base model. The RAG pipeline combined multilingual-e5-large for retrieval and bge-reranker-v2-m3 for reranking, outperforming single-method approaches. Training with both QA and multiple-choice data improved results, and instruction-based prompts yielded the best performance when using only one retrieved document.
When evaluating the system with a new test dataset, we found that its performance remains limited when answering questions about legal statutes that were not present in the training data.

For future work, we plan to extend the framework to serve both general users and legal experts. Additionally, we aim to deploy the system in real-world settings to enhance accessibility and reproducibility, as well as continue to optimize inference time for better performance. We also plan to further improve the QA and RAG systems to enhance the accuracy of legal document retrieval, ensuring more effective answers to questions involving all statutes present in the system.

REFERENCES

1. Sourdin, T., *Judge v Robot?: Artificial intelligence and judicial decision-making.* University of New South Wales Law Journal, The, 2018. **41**(4): p. 1114-1133.
2. Rodrigues, R., *Legal and human rights issues of AI: Gaps, challenges and vulnerabilities.* Journal of Responsible Technology, 2020. **4**: p. 100005.
3. ARIAI, F. and G. DEMARTINI, *Natural Language Processing for the Legal Domain: A Survey of Tasks, Datasets, Models, and Challenges.* ACM Comput. Surv, 2024. **1**(1).
4. Yao, R., et al., *Elevating Legal LLM Responses: Harnessing Trainable Logical Structures and Semantic Knowledge with Legal Reasoning.* arXiv preprint arXiv:2502.07912, 2025.
5. Colombo, P., et al., *Saullm-7b: A pioneering large language model for law.* arXiv preprint arXiv:2403.03883, 2024.
6. Wan, Z., et al. *Reformulating Domain Adaptation of Large Language Models as Adapt-Retrieve-Revise: A Case Study on Chinese Legal Domain*. in *Findings of the Association for Computational Linguistics ACL 2024*. 2024.
7. Chlapanis, O., I. Androutsopoulos, and D. Galanis. *Archimedes-AUEB at SemEval-2024 Task 5: LLM explains Civil Procedure*. in *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*. 2024.
8. Deng, C., K. Mao, and Z. Dou. *Learning Interpretable Legal Case Retrieval via Knowledge-Guided Case Reformulation*. in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. 2024.
9. Mamooler, S., R. Lebret, S. Massonnet, and K. Aberer. *An Efficient Active Learning Pipeline for Legal Text Classification*. in *Proceedings of the Natural Legal Language Processing Workshop 2022*. 2022.
10. Ouyang, L., et al., *Training language models to follow instructions with human feedback.* Advances in neural information processing systems, 2022. **35**: p. 27730-27744.
11. Lai, J., et al., *Large language models in law: A survey.* AI Open, 2024.

12. Iu, K.Y. and V.M.-Y. Wong, *Chatgpt by openai: The end of litigation lawyers?* Available at SSRN 4339839, 2023.
13. Jiang, H., et al. *Leveraging Large Language Models for Learning Complex Legal Concepts through Storytelling*. in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2024.
14. Dettmers, T., A. Pagnoni, A. Holtzman, and L. Zettlemoyer, *Qlora: Efficient finetuning of quantized llms.* Advances in Neural Information Processing Systems, 2024. **36**.
15. Zhang, T., et al. *BERTScore: Evaluating Text Generation with BERT*. in *International Conference on Learning Representations*.
16. Lin, C.-Y. *Rouge: A package for automatic evaluation of summaries*. in *Text summarization branches out*. 2004.
17. Vaswani, A., *Attention is all you need.* Advances in Neural Information Processing Systems, 2017.
18. Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova. *Bert: Pre-training of deep bidirectional transformers for language understanding*. in *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*. 2019.
19. Wang, L., et al., *Multilingual e5 text embeddings: A technical report.* arXiv preprint arXiv:2402.05672, 2024.
20. Chen, J., et al. *M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation*. in *Findings of the Association for Computational Linguistics ACL 2024*. 2024.
21. Yuenyong, S., K. Viriyayudhakorn, A. Piyatumrong, and J. Jaroenkantasima, *OpenThaiGPT 1.5: A Thai-Centric Open Source Large Language Model.* arXiv preprint arXiv:2411.07238, 2024.
22. Zhang, W., et al., *Seallms 3: Open foundation and chat multilingual large language models for southeast asian languages.* arXiv preprint arXiv:2407.19672, 2024.
23. Yang, A., et al., *Qwen2 Technical Report.* arXiv preprint arXiv:2407.10671, 2024.

24. Pipatanakul, K., et al., *Typhoon 2: A Family of Open Text and Multimodal Thai Large Language Models.* arXiv preprint arXiv:2412.13702, 2024.
25. Yang, A., et al., *Qwen2. 5 Technical Report.* arXiv preprint arXiv:2412.15115, 2024.
26. Dubey, A., et al., *The llama 3 herd of models.* arXiv preprint arXiv:2407.21783, 2024.
27. Knowledge Sharing from NECTEC. *OpenThaiLLM-Prebuilt Release*. 2024; Available from: https://medium.com/nectec/openthaillm-prebuilt-release-f1b0e22be6a5.
28. Knowledge Sharing from NECTEC. *PathummaLLM-Text-V 1.0.0 Release*. 2024; Available from: https://medium.com/nectec/pathummallm-text-v-1-0-0-release-1fd41344b061.
29. Dou, L., et al., *Sailor2: Sailing in South-East Asia with Inclusive Multilingual LLMs.* arXiv preprint arXiv:2502.12982, 2025.
30. Ruder, S., A. Søgaard, and I. Vulić. *Unsupervised cross-lingual representation learning*. in *Proceedings of the 57th annual meeting of the association for computational linguistics: Tutorial abstracts*. 2019.
31. Wenzek, G., et al., *CCNet: Extracting High Quality Monolingual Datasets from Web Crawl Data.*
32. Song, K., et al., *Mpnet: Masked and permuted pre-training for language understanding.* Advances in neural information processing systems, 2020. **33**: p. 16857-16867.
33. Lowphansirikul, L., C. Polpanumas, N. Jantrakulchai, and S. Nutanong, *Wangchanberta: Pretraining transformer-based thai language models.* arXiv preprint arXiv:2101.09635, 2021.
34. Sriwirote, P., J. Thapiang, V. Timtong, and A.T. Rutherford, *Phayathaibert: Enhancing a pretrained thai language model with unassimilated loanwords.* arXiv preprint arXiv:2311.12475, 2023.
35. Yu, F., L. Quartey, and F. Schilder, *Legal prompting: Teaching a language model to think like a lawyer.* arXiv preprint arXiv:2212.01326, 2022.
36. Brief, M., et al., *Mixing It Up: The Cocktail Effect of Multi-Task Fine-Tuning on*

LLM Performance--A Case Study in Finance. arXiv preprint arXiv:2410.01109, 2024.

37. Jiang, A.Q., et al., *Mistral 7B.* arXiv preprint arXiv:2310.06825, 2023.
38. VISTEC-depa AI Research Institute of Thailand, *WangchanX-Legal-ThaiCCL-RAG*. 2024: Hugging Face.
39. PyThaiNLP, *Han Instruct Dataset v2.0*. 2024: Hugging face.
40. Lawyers Council Under The Royal Patronage. รวมข้อสอบพร้อมแนวคำตอบการฝึกอบรมวิชาว่าความ ภาคทฤษฎี ตั้งแต่รุ่นที่ 60-รุ่นที่ 61. Nonthaburi: Dharmasarn Printing.
41. Lawyers Council Under The Royal Patronage. รวมข้อสอบพร้อมแนวคำตอบการฝึกอบรมวิชาว่าความ ภาคทฤษฎี ตั้งแต่รุ่นที่ 55-รุ่นที่ 59. Nonthaburi: Dharmasarn Printing.
42. Lawyers Council Under The Royal Patronage. รวมข้อสอบพร้อมแนวคำตอบการฝึกอบรมวิชาว่าความ ภาคทฤษฎี ตั้งแต่รุ่นที่ 16-รุ่นที่ 54. Nonthaburi: Dharmasarn Printing.
43. Loshchilov, I. and F. Hutter. *Decoupled Weight Decay Regularization*. in *International Conference on Learning Representations*.
44. Yu, F., L. Quartey, and F. Schilder. *Exploring the effectiveness of prompt engineering for legal reasoning tasks*. in *Findings of the Association for Computational Linguistics: ACL 2023*. 2023.
45. Liu, N.F., et al., *Lost in the Middle: How Language Models Use Long Contexts.* Transactions of the Association for Computational Linguistics, 2024. **11**: p. 157-173.

VITA

NAME	Supachoke Hanwiboonwat
DATE OF BIRTH	30 March 2000
PLACE OF BIRTH	Bangkok
INSTITUTIONS ATTENDED	Department of Nuclear Enginerring, Faculty of Engineering, Chulalongkorn University Department of Computer Enginerring, Faculty of Engineering, Chulalongkorn University
HOME ADDRESS	599/42 Ladprao 1 Road Jompol Subdistrict, Chatuchak District, Bangkok 10900 Thailand