CP Capstone Final Project
รายงานโครงงานวิศวกรรมคอมพิวเตอร์

เรื่อง
การสร้างแชตบอตภาษาไทยที่มีความน่าเชื่อถือและเชี่ยวชาญ
สำหรับการตอบคำถามเฉพาะทาง
Reliable Domain-Specific Chatbot for Thai Language

โดย

ภูริภัทร รัตนสงวนวงศ์	รหัสนิสิต 6330440921
สรวิศ ธนิษฐาพงศา	รหัสนิสิต 6330528521
อคิราภ์ สิทธิการิยะวัตร	รหัสนิสิต 6330563421
อัครวัฒน์ บูรโชควิวัฒน์	รหัสนิสิต 6330585221

อาจารย์ที่ปรึกษา
รศ. ดร. พีรพล เวทีกูล
ดร. ปรัชญา บุญขวัญ (NECTEC)
ดร. อภิวดี ปิยธรรมรงค์ (NECTEC)

รายงานนี้เป็นส่วนหนึ่งของวิชา 2110489 โครงงานรวบยอดวิศวกรรมคอมพิวเตอร์ 2
ภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ประจำปีการศึกษา 2566

บทคัดย่อ

เพื่อตอบสนองต่อข้อจำกัดของโมเดลภาษาขนาดใหญ่ (Large Language Model) GPT-4 ที่ได้เป็นที่ยอมรับกันทั่วไปว่ามีความทันสมัยที่สุด โดยเฉพาะในเรื่องของการจัดการประมวลผลข้อมูลเฉพาะทาง (Specific-Domain private data) และการประมวลผลภาษาไทย โครงการการสร้างแชตบอตภาษาไทยที่มีความน่าเชื่อถือและเชี่ยวชาญสำหรับการตอบคำถามเฉพาะทางจึงมีเป้าหมายที่จะเติมเต็มช่องว่างเหล่านี้ โดยเริ่มจากการหาต้นตอปัญหาการตอบมั่วไปเอง (hallucination) ในคำตอบจากโมเดลภาษาขนาดใหญ่ ซึ่งเป็นปัจจัยสำคัญที่ส่งผลต่อความน่าเชื่อถือของแชตบอตดังกล่าว โดยเราจะมุ่งเน้นไปที่การสำรวจหาต้นตอและสาเหตุของปัญหาการตอบมั่วไปเอง ในแชตบอตที่ถูกสร้างขึ้นจากโมเดลภาษาขนาดใหญ่ที่ถูกฝึกด้วยข้อมูลภาษาไทย (open-source LLMs with pre-trained Thai corpus) และนำเสนอวิธีการแก้ไขปัญหาดังกล่าวที่พบ โดยในท้ายที่สุด โครงการต้องการแสดงให้เห็นว่าแชตบอตภาษาไทยที่พัฒนาขึ้นมาจากโมเดลภาษาขนาดใหญ่ สามารถให้คำตอบที่มีคุณภาพเทียบเท่ากับการให้คำตอบโดยผู้เชี่ยวชาญมนุษย์และโมเดลขั้นสูงเช่น GPT-3.5 ในการตอบคำถามเฉพาะทางในภาษาไทยด้านกฎหมาย ซึ่งจะเปิดให้ใช้งานแอพพลิเคชั่นสาธิต (demo application) ผ่านทางแชทของแอพพลิเคชั่นไลน์ (LINE) อันเป็นพื้นฐานสำหรับแนวทางการพัฒนาแชตบอตเฉพาะทางภาษาไทยในอนาคต

Abstract

In response to the challenges presented by state-of-the-art generalized Large Language Models (LLMs) like GPT-4, particularly in handling domain-specific private data and the Thai language due to an existing degree of uncertainty regarding its proficiency in comprehending and generating texts in Thai, our project, "Reliable Domain-Specific Chatbot for Thai Language," aims to develop a chatbot that addresses these critical issues. The primary focus is on overcoming the problem of hallucination in LLM responses, a key factor affecting the reliability of chatbots in specialized domains. This project will explore the root causes of hallucination in open-source LLMs with pre-trained Thai corpus and propose targeted solutions to mitigate this issue. By leveraging the open-source LLM with pre-trained Thai corpus, we aim to demonstrate that our chatbot can generate responses of comparable quality to human-generated texts and those produced by GPT-3.5 while being specifically attuned to the nuances of the Thai language and legal domain. The end goal is to present a chatbot model that exemplifies reliability in domain-specific applications, accessible via the LINE application chat platform, representing a significant advancement in the field of language-specific chatbot technology.

1.Introduction

1.1. Project background

The advent of ChatGPT, developed from large language models (LLMs) like GPT-4, has revolutionized the landscape of automated communication, offering significant potential in various fields such as customer service, sales, marketing, and knowledge management. While ChatGPT has demonstrated proficiency in handling English-language queries, challenges arise when adapting this technology to other languages, particularly those with unique linguistic structures like Thai.

In the case of Thai language, the current chatbot models face issues of incorrect responses and misunderstandings due to the language's complexity, including its nuanced grammar, varied sentence structures, and extensive use of slang and expressions (for instance, มากกก instead of มาก or ก้ instead of ก็). Additionally, a critical challenge inherent in Large Language Models (LLMs) is the phenomenon of ‘hallucination,’ where the models generate responses that, while seemingly coherent and articulate, are in fact factually inaccurate or unrelated to the queried subject matter. This issue is particularly problematic as it compromises the reliability of the model's outputs, posing significant concerns in contexts where precision and factual correctness are essential.

Another critical aspect is the model's limitations in handling domain-specific private data. While ChatGPT excels in general topics, its performance dwindles in specialized fields, especially where training data is limited or the content is highly technical, as in legal domains. This limitation becomes more pronounced in the context of the Thai language, where domain-specific nuances add an additional layer of complexity.

Recognizing these challenges, our project aims to develop a Thai language chatbot that specifically addresses these issues. The goal is to create a chatbot that excels in providing precise responses in specialized domains, particularly legal Thai law domain. By focusing on these critical areas, the project seeks to advance the capabilities of chatbots in handling complex, domain-specific interactions in Thai, setting a new standard for reliability and accuracy in language-specific chatbot technology.

1.2. Objectives

The project aims to develop a Thai law chatbot to answer questions within a specific law domain accurately and precisely based on the input queries and latest Thai legal documentation, without having hallucinations in its responses. By addressing this challenge, our chatbot will not only provide factually correct and contextually relevant information but also set a new standard for reliability and precision in domain-specific chatbot technology.

1.3. Scope of work

1.3.1 Chatbot Development Using RAG Method

The development process will focus on constructing the core functionalities of the chatbot using the Retrieval Augmented Generation (RAG) method. This approach combines the retrieval-based and the generation-based methods to create a hybrid system, enhancing the chatbot's ability to provide improved and contextually relevant responses, ensuring the chatbot can handle the complexities of Thai language and legal terminology accurately.

1.3.2 Thai Law Dataset Preparation for LLM Fine-Tuning and Evaluation

A comprehensive Thai law dataset must be prepared to fine-tune the LLM model using the QLoRA adapter, along with the configuration of hyperparameters, to optimize the model for accurate and specialized responses within the legal domain. This step ensures the chatbot is equipped with the necessary legal knowledge and understands the specific context within the Thai legal domain.

1.3.3 Refinement Process to Reduce Hallucination

A key focus will be refining the chatbot to address the issue of hallucination to create reliability of the chatbot. This involves targeted enhancements in three areas: the underlying Language Model (LLM), the information retrieval system, and the design of the chatbot prompts. These refinements are essential for ensuring the chatbot's responses are not only fluent but also factually accurate and directly relevant to user queries, particularly in the legal domain.

1.3.4 Chatbot Deployment on LINE Official Account

This phase involves both deploying the chatbot on the LINE Official Account application and optimizing the chatbot to reduce the production cost and ensure timely responses for end users. This makes the chatbot easily accessible to users within this platform, allowing for real-world application and testing. This will demonstrate the end-to-end workflow of building a chatbot that serves the practical needs of users engaging with Thai law.

1.3.5 Chatbot Evaluation

The final phase involves assessing chatbot performance in both contextual correctness and user preferences. The comparison will be made with an established baseline of known LLM such as GPT-3.5 as well as the competitor within the same Thai legal domain.

1.4. Methods and Plans

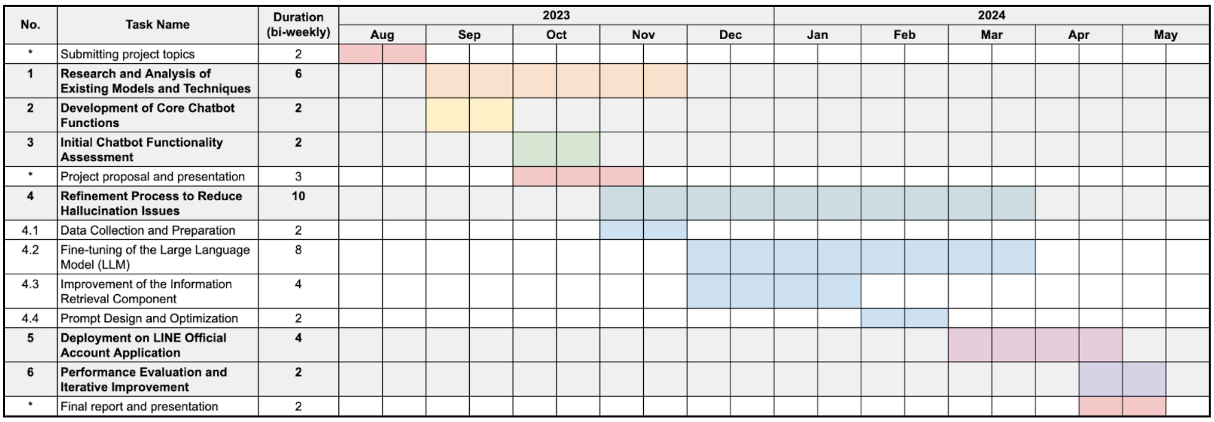

Figure 1: Project's Gantt Chart

Figure 1 illustrates the project planning using the Gantt chart. The detail of each phase are describe as following:

1. <u>Research and Analysis of Existing Models and Techniques</u>
   We will conduct an in-depth study of current chatbot technologies to identify their strengths and limitations, particularly in handling the Thai language and domain-specific data. We will also study existing strategies to mitigate hallucination in Large Language Models (LLMs).

2. <u>Development of Core Chatbot Functions</u>
   The foundational elements of the chatbot will be developed using the Retrieval Augmented Generation (RAG) method. This approach is pivotal in ensuring the chatbot delivers contextually accurate and relevant responses, a critical requirement in the complex domain of Thai law.

3. <u>Initial Chatbot Functionality Assessment</u>
   This phase aims to evaluate the chatbot's basic functionality using a created dataset of 200 questions sourced from 20 Thai Supreme Court cases of 2022, covering specific legal cases, general law, and out-of-domain topics. The assessment, involving both human and GPT-4 generated questions, will focus on the chatbot's information retrieval efficiency, question-answering accuracy, and overall operational workflow. A key aspect will be identifying and analyzing any instances of hallucination in the responses, guiding future refinements.

4. Refinement Process to Reduce Hallucination Issues

   Building upon insights from the initial functionality assessment, this phase focuses on refining our chatbot to mitigate hallucination issues, crucial for enhancing its reliability in legal contexts. This process will encompass several key steps as following:

   4.1. Legal Document Preparation and Dataset Construction

   We will prepare a legal documentation to serve as a knowledge database and compile a comprehensive dataset pertinent to Thai law, including case law and legal commentaries. This dataset will form the basis for fine-tuning the LLM to provide specific and accurate legal domain responses.

   4.2. Fine-tuning of the Large Language Model (LLM)

   The LLM will undergo fine-tuning using the QLoRA adapter. This process will involve configuring hyperparameters to enhance performance and accuracy.

   4.3. Improvement of the Information Retrieval Component

   We will upgrade the information retrieval system, integrating advanced keyword and contextual search methodologies. This enhancement is essential for accurately sourcing relevant legal information for the LLM.

   4.4. Prompt Design and Optimization

   The chatbot's prompts will be carefully designed and continually refined to ensure they effectively elicit clear, precise, and relevant responses.

5. Chatbot Runtime Optimization and Deployment

   Following development and refinement, the chatbot will be optimized specifically for inference runtime and will be launched on the LINE Official Account application. This step is crucial for real-world testing and user feedback collection, providing invaluable insights for future enhancements.

6. Performance Evaluation and Iterative Improvement

   We will finally evaluate the chatbot performance against established criteria focusing on both accuracy and user preferences. This evaluation will benchmark the chatbot against existing law chatbot technologies and other LLM models.

1.5. Expected benefit

This study will benefit in many aspects as the developed law chatbot can be used to specifically help answering law questions for Thai’s people. Additionally, this developed chatbot can be applied to various industries or other domains in Thailand, aiming to advance the business in alignment with customer needs.

2.Literature Review

2.1. Related theory

The theories and concepts underpinning the development and refinement of chatbots form a comprehensive landscape that encompasses diverse aspects of natural language processing and model evaluation. At the forefront of this theoretical framework lies the Large Language Model (LLM), a powerful deep learning algorithm trained on extensive datasets, capable of various natural language processing tasks. Coupled with the Retrieval Augmented Generation (RAG) framework,[2] which harnesses external knowledge bases to augment the LLM, these two theories lay the foundation for improving information extraction and generation. Complementing these, FAISS enables efficient similarity searches, crucial for information retrieval. The challenge of 'hallucination' in AI models brings to light the complexities of generating accurate responses, mitigated through innovative concepts like adapters that fine-tune specific neural networks within pre-trained models.[5] Prompt engineering plays a pivotal role in task instruction, while evaluation metrics like BERTScore and BLEU score offer quantitative measures for translation quality. Additionally, the L2 distance provides a fundamental method for evaluating the semantic similarity between generated and reference texts. This collection of theories, ranging from language model capabilities to evaluation metrics, represents a cohesive framework essential for the development, refinement, and assessment of chatbot systems.

2.1.1. Large Language Model (LLM)

A deep learning algorithm known as a Large Language Model (LLM) possesses the capability to undertake numerous natural language processing (NLP) tasks. These models utilize transformer architectures and are trained on extensive datasets, which is why they are labeled as 'large'. As a result, they have the ability to identify, interpret, forecast, or produce text and various other content.[1] LLMs take a complex approach that involves multiple components. At its core, an LLM requires extensive training on a vast dataset. This training typically involves several stages, often commencing with unsupervised learning. During this phase, the model learns from unstructured and unlabeled data, beginning to discern connections among various words and ideas. Subsequently, certain LLMs undergo further training and refinement via self-supervised learning, which involves partial data labeling to enhance the model's ability to discern different concepts accurately. Following this, the LLM engages in deep learning through the transformer neural network process. This architecture empowers the LLM to grasp and identify relationships and associations among words and concepts, facilitated by a self-attention mechanism.[3]

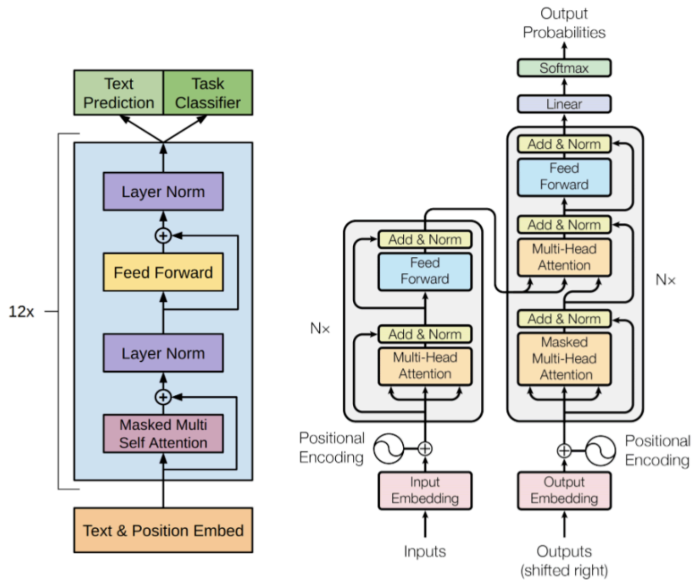

Figure 2: The architecture of large language models is based on a type of deep learning called the Transformer architecture.[3]

2.1.2. Retrieval Augmented Generation (RAG)

RAG serves as an AI framework designed to extract information from an external knowledge base. Its purpose is to additionally provide accurate and current data for large language models (LLMs) and to offer users a deeper understanding of how these models generate information.[2]

2.1.3. FAISS

FAISS, short for Facebook AI Similarity Search, is a software tool designed for developers to swiftly find similarities between embeddings of various media documents. It addresses the constraints seen in conventional query search engines, which are tailored for hash-based searches, and offers enhanced and more scalable similarity search capabilities.[2]

2.1.4. Hallucination

Hallucinations refer to moments when the AI model “imagines” or “fabricates” information that doesn't precisely align with the given input. Although the term "hallucination" might suggest the AI becoming conscious, it's somewhat misleading. These hallucinations don't indicate a self-aware AI. Instead, they showcase a unique aspect of machine learning algorithms and the diverse information they are trained on.[12]

2.1.5. Adapter

The concept of adapters revolves around the idea of inserting compact and task-specific neural networks, known as adapters, into a pre-trained base model. These adapters are lightweight modules that are added to the existing layers of pre-trained models, enabling it to handle diverse downstream tasks without fine-tuning the entire model.[7]

2.1.6. Prompt and Prompt Engineering

Prompt is a natural language text that is used to describe the task to the Large Language Model (LLM) in order to instruct them to get the desired outputs or responses. On the other hand, prompt engineering is a process of structuring the text so that it can be understandable by the LLM, giving the best results which involves many techniques such as giving an example, also known as few-shot learning.[9]

2.1.7. L2 distance

The L2 distance, also known as Euclidean distance, is a measure of the magnitude of the difference between two points in Euclidean space. In text generation evaluation, it involves converting text to numerical vectors through embeddings, which capture semantic meanings. The L2 distance is then calculated between these embeddings of the generated text and a reference text.[10]

2.1.8. BERT score

BERTScore is a recent metric used to evaluate translation quality. It begins by obtaining BERT representations for each word in the candidate and reference sentences separately, achieved by processing them through a BERT model. It then calculates an alignment between the candidate and reference words by measuring pairwise cosine similarity. This alignment is subsequently combined to calculate precision and recall scores, which are further used to derive an (adjusted) F1 score, taking into account inverse-document-frequency values.[11]

2.1.9. BLEU score

BLEU is a precision-based metric created for the assessment of machine translation. It assesses a candidate translation by counting the number of n-grams (sequential sets of n words) in the candidate that match those found in a reference translation. The value of n ranges from 1 up to a specified N, and the scores for different values of n are combined using a geometric mean.[11]

2.1.10. ROUGE-L score

ROUGE-L evaluates text by identifying the longest common subsequence (LCS) between the compared and reference texts. It calculates precision, recall, and F1-score based on the LCS's length. Commonly applied in assessing semantic similarity and content coverage in generated text, ROUGE-L focuses on the shared subsequences without considering the order of words.[17]

2.1.11. METEOR score

METEOR is an automated metric designed for the evaluation of machine translation. It relies on a broad concept of unigram matching, comparing machine-generated translations with human-generated reference translations. Unigrams can be matched based on their surface forms, stemmed forms, and meanings. Additionally, METEOR can be expanded easily to incorporate more sophisticated matching strategies. After identifying all generalized unigram matches between the two texts, METEOR calculates a score for this matching by considering unigram precision, unigram recall, and a fragmentation measure. This fragmentation measure is specifically crafted to directly assess how well-organized the matched words in the machine translation are in comparison to the reference.[16]

2.1.12. Cosine Similarity

Cosine similarity is a metric used to evaluate how similar two vectors are within an inner product space. This similarity is determined by calculating the cosine of the angle between the two vectors, indicating whether they are approximately pointing in the same direction. This method is frequently utilized in text analysis to assess the similarity between documents.[15]

2.2. Related research

2.2.1. A Survey on Hallucination in Large Language Models

This research provides a comprehensive overview of recent efforts to detect, explain, and mitigate hallucination in LLMs. It discusses common types of hallucination, approaches to mitigating them. Common types of hallucination in LLMs include generating content that diverges from user input, contradicts previously generated context, or misaligns with established world knowledge. Approaches to mitigating hallucination include curating pre-training corpora to minimize unverifiable or unreliable data, devising more effective selection or filtering strategies, and fine-tuning LLMs with data that does not exceed their parametric knowledge.[12]

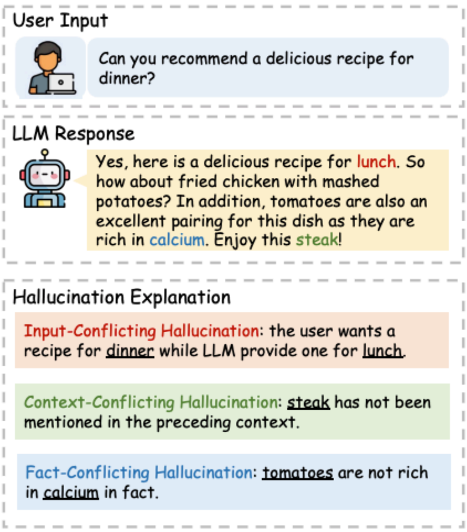

Figure 3: Three types of hallucinations occurred in LLM responses[12]

2.2.2. QLoRA: Efficient Fine-tuning of Quantized LLMs

The research introduces QLoRA, a novel approach to fine tuning quantized language models that reduces memory usage while maintaining performance levels. The authors demonstrate the effectiveness of QLoRA by training a 33B parameter model called Guanaco on open-source data, which achieves state-of-the-art performance on the Vicuna benchmark. The paper includes both quantitative and qualitative analyses of Guanaco's performance, as well as a discussion of the limitations of current chatbot benchmarks and evaluation protocols. In short, QLoRA is the adapter that can be used to finetune LLMs with low computing resources.[13]

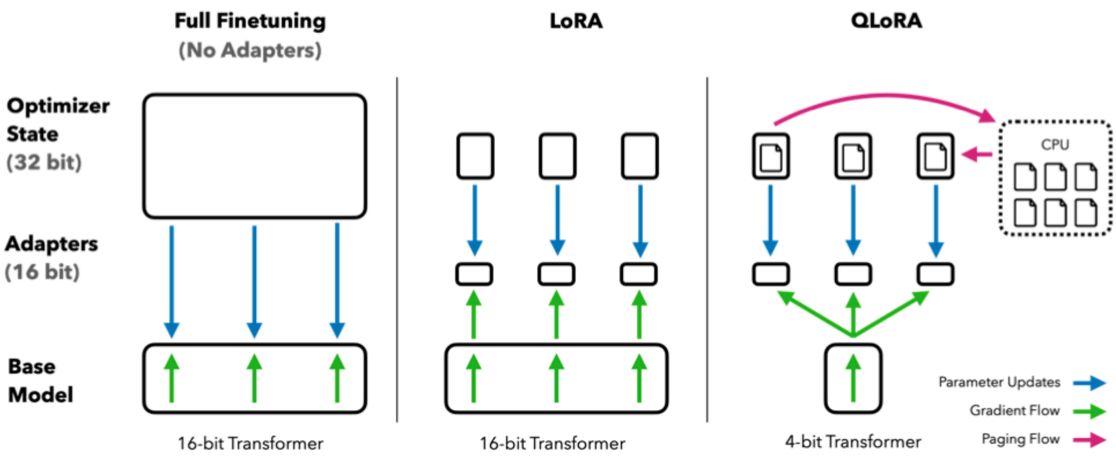

Figure 4: Different finetuning methods and their memory requirements. QLoRA improves over LoRA by quantizing the transformer model to 4-bit precision and using paged optimizers to handle memory spikes.[13]

2.2.3. PEEP-Talk: A Situational Dialogue-based Chatbot for English Education

This research proposes a chatbot with input topic or situation in which users can choose various situations from the select box. The study evaluated the performance and education efficiency of PEEP-Talk through a human evaluation consisting of assessments in English speaking, grammar, English learning anxiety, and user satisfaction.[14]

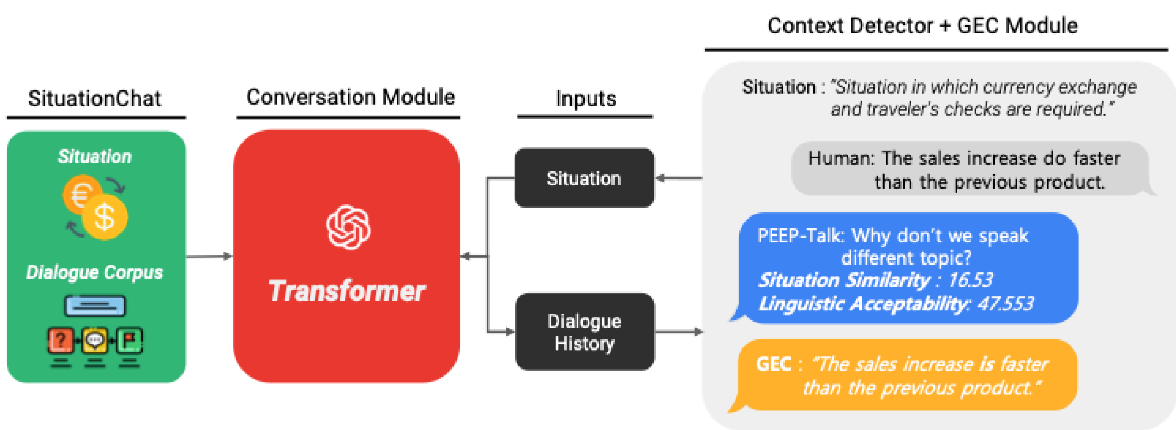

Figure 5: An overview of the modules in PEEP-Talk.[14]

3. Methodology

This project, aimed at developing a reliable domain-specific chatbot for Thai language, particularly in the legal domain, is structured into five distinct yet interconnected phases. Each phase is critical to ensuring the chatbot's effectiveness, accuracy, and reliability.

3.1 Core Chatbot Architecture Development

The initial phase of our project involved designing and constructing the first prototype of our chatbot using the Retrieval Augmented Generation (RAG) approach. To achieve optimal functionality and flexibility, we structured the chatbot into two primary components: Information Retrieval and Question-Answering. Each of these components was further divided into loosely coupled sub-modules. This modular design ensures system flexibility, allowing for independent development and maintenance of each component, and facilitates easier updates and enhancements.

- Information Retrieval: This module is responsible for sourcing and retrieving relevant legal documentation from a designated database.
- Question-Answering: This module takes the retrieved information and formulates responses to user queries. This component will be connected with LLM to generate precise and contextually relevant answers.

Furthermore, to ensure the chatbot's robustness and adaptability, we integrated it with various Large Language Models, including closed-source models like GPT-3.5-turbo and GPT-4, and open-source models that is specifically pre-trained with Thai corpus such as OpenThaiGPT and/or WangChanGLM. This integration enables us to compare and select the most effective model for handling the nuances of the Thai language and legal domain.

Figure 6 illustrates the structural layout of the chatbot, showing the interconnections between the different modules and how they contribute to the overall functionality of the chatbot system.

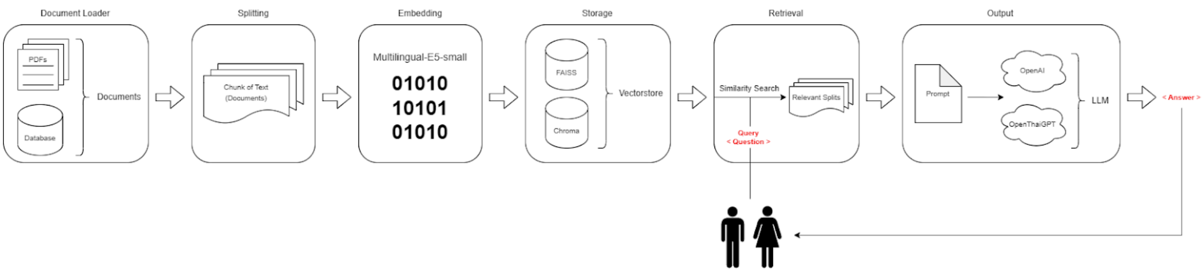

Figure 6: An architecture of the chatbot

3.2 Initial Chatbot Functionality Assessment

After building the foundational architecture in the previous phase, this stage involves a comprehensive assessment of the chatbot's initial functionalities using a prepared dataset. This dataset, specifically tailored for testing, comprises 200 questions derived from 20 Thai Supreme Court cases from the year 2022. The question set is diversified into three categories:

- Specific Law Case Questions: Focused on particular legal cases and their intricacies.
- General Law Questions: Covering broader legal concepts and principles.
- Out-of-Domain Questions: Intended to test the chatbot's ability to handle queries unrelated to law.

The questions are generated using both humans and inputs from the Large Language Model (GPT-4). Each question and its corresponding category have been rigorously reviewed and verified by humans to maintain accuracy and relevance. The evaluation process focuses on distinct aspects of the chatbot's performance:

- Information Retrieval: This assesses the retriever's ability to provide information chunks corresponding to the query.
- Question-Answering: This evaluates the LLM capabilities to answer the question based on the provided information.
- Overall Chatbot Process: This examines the entire operational workflow of the chatbot, from receiving queries to generating responses.

A critical element of this assessment is the observation of hallucination instances in the chatbot's responses. Identifying and analyzing these occurrences will provide valuable insights into potential root causes, setting the stage for targeted refinements in the subsequent phase.

3.3 Targeted Chatbot Refinement

Following insights garnered from the chatbot performance analysis, our objective shifted toward refinement. We embarked on a targeted refinement phase to enhance the chatbot's capabilities. This phase comprises several focused preparation and improvements:

3.3.1 Legal Document and Dataset Preparation

To work with private domain data, specifically the Thai legal domain in this case, we aimed to prepare a comprehensive Thai legal knowledge base, gathered from an official documentation website. We will perform data cleansing, and reformatting the documentation into a format such that any further implementation can be easily integrated with. Additionally, a large dataset will be generated here with state-of-the-art LLM, GPT-4, to serve as a training and testing dataset. The dataset verification process by humans will also be conducted to ensure the correctness of the dataset.

3.3.2 Enhancement of the Information Retrieval Process

To address the issue of irrelevant information retrieval, we plan to integrate advanced search engine technologies, such as elastic keyword search, into our document retriever. This integration aims to sharpen the retriever's ability to precisely fetch information that is directly relevant to user queries, thereby reducing the likelihood of the LLM generating answers based on unrelated data.

3.3.3 Fine-Tuning of LLM Adapters

LLM adapters will be fine-tuned using our specialized question-answering dataset and private legal domain data. This step is intended to increase the LLM expertise in specialized areas of Thai law and enhance its question-answering accuracy and relevance.

3.3.4 Optimization of Chatbot Prompts

This involves refining the prompt given to the LLM to ensure that the responses generated are not only accurate but also contextually relevant to the user's query.

3.4 Chatbot Deployment

As the chatbot has been refined and ready to serve the purpose of legal question answering, the focus shifts towards its deployment. The implementation phase is critical, as it involves integrating the chatbot into a real-world application that can serve users directly. To this end, we will deploy the chatbot within an official LINE account, positioning it as a primary resource for legal assistance. The deployment process will focus on both functionalities and enhanced user experience as follows:

3.4.1 Chatbot Runtime Optimization

A key aspect of deployment is optimizing the chatbot’s performance in live sessions. This involves reducing latency to ensure that users receive timely responses, which is crucial for maintaining engagement and satisfaction. In addition to enhancing responsiveness, we will focus on reducing memory usage, which directly impacts both operational costs and hardware efficiency. The process will mainly focus on the optimization part of the LLM as it will be the main bottleneck that takes the most time and computational resources here.

3.4.2 Chatbot deployment on LINE Official Account

After the runtime optimization process, we will set up a chatbot within the LINE application to make it accessible to the public. This deployment will leverage ngrok, server of datamind lab. The integration with the LINE application is facilitated through the Line Bot SDK for Python, establishing a direct connection with the LINE Messaging API. A critical component of this integration is the implementation of a Webhook, which acts as a bridge between the LINE Messaging API and our chatbot's processing logic. This Webhook links to a custom backend developed using Python’s FastAPI web framework, responsible for receiving requests from Dialogflow and processing them for interaction with the LINE Bot SDK. This backend setup allows us to write versatile code that connects to various databases, APIs, and services, enhancing the chatbot's functionality and autonomy. Deploying the chatbot on the LINE platform not only makes it available for real-world user interaction but also opens avenues for collecting valuable user feedback. This feedback is crucial for future refinements and enhancements, ensuring that the chatbot continues to evolve in its capabilities and effectiveness. The deployment phase is pivotal, marking the transition of our chatbot from a conceptual model to a practical, user-accessible tool.

3.5 End-to-End Chatbot Performance Evaluation

In this concluding phase, after we deploy our LINE application chatbot, we will assess our chatbot, benchmarking it against models such as GPT-3.5, and direct competitors in Thai legal domain. Our evaluation will encompass two primary dimensions:

- <u>Factual Correctness</u>: To gauge the accuracy of factual information, we will employ metrics like Lexical Matching, ensuring responses contain key information (golden keywords). Additionally, we will utilize language evaluation metrics such as BLEU, BERT-score, METEOR, and ROUGE-L to analyze the precision of content and the relevance of responses in relation to the input queries.
- <u>User Preference and Feedback</u>: This aspect will be evaluated with humans. We will conduct a comparative assessment (A vs. B style) to determine which model produces more fluent and contextually appropriate responses. We will also gather direct feedback from users interacting with the chatbot, focusing on their subjective experience, the perceived value of the information provided, and the overall satisfaction with the interaction.

These comprehensive methods are designed to offer a holistic view of our chatbot's performance, from factual accuracy to user preference. We anticipate that this multi-faceted approach will not only validate the technical competencies of our chatbot but also reveal insights into user engagement and satisfaction. This will inform future iterations and improvements of the chatbot, ensuring that it not only responds with factual precision but also meets the nuanced expectations of users seeking legal assistance.

4. Result and Analysis

In this section, we present the results and analysis of the initial phases of our project, focusing on the development and preliminary assessment of our chatbot's architecture. The chatbot, as illustrated in the previously mentioned architecture diagram, is structured around two critical components: the information retriever and the question-answering mechanism. The chatbot then, has been evaluated by the constructed dataset to identify specific aspects that required refinement.

4.1 Core Chatbot Architecture Development

The initial phase of our project, focused on developing the core architecture of the chatbot, has been successfully completed. However, we often encounter hallucinated responses, where the chatbot generates factually incorrect or irrelevant answers.

Figure 7 demonstrates the example query result from the initial developed chatbot application, built using Chainlit library as a framework for the web interface.

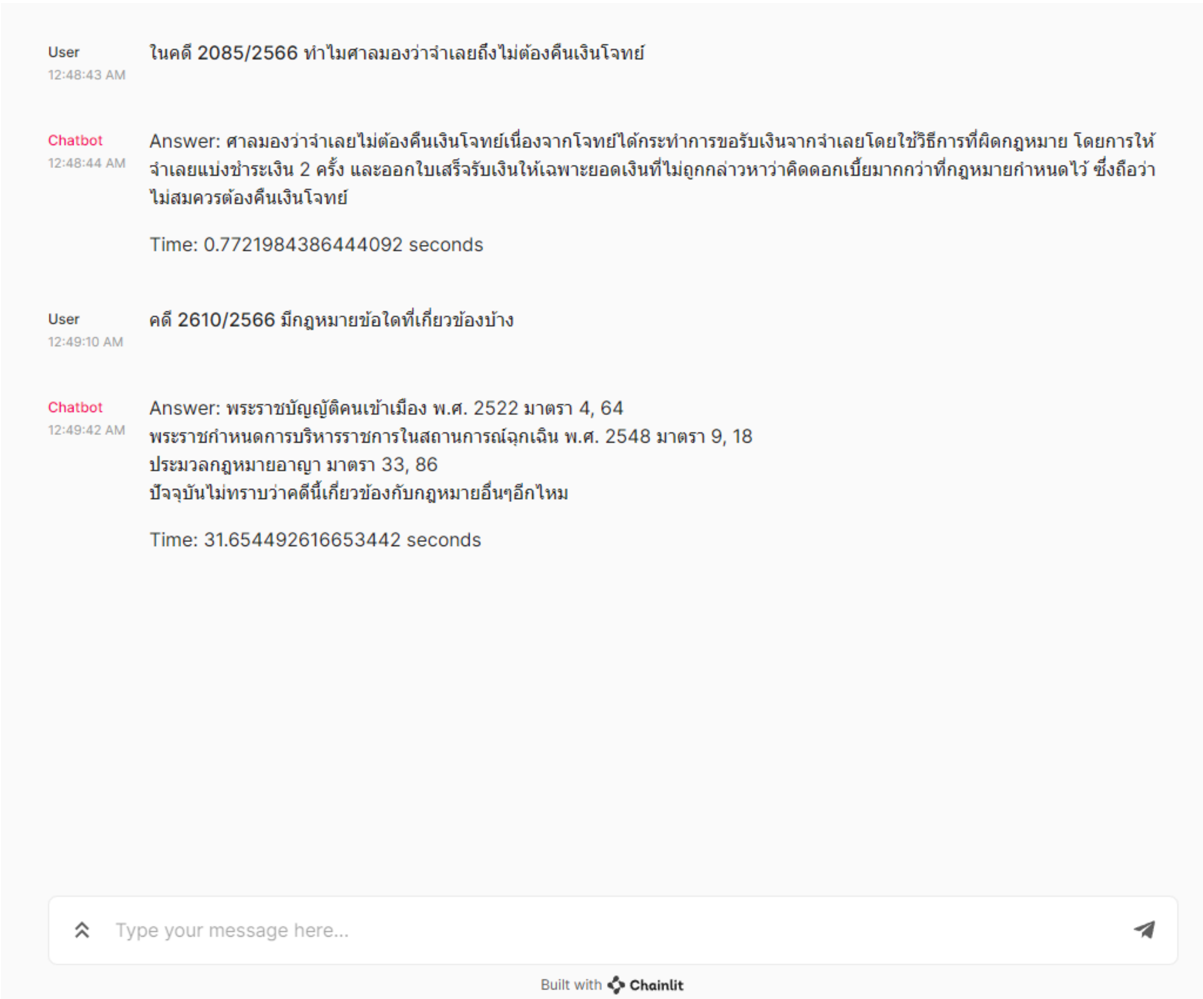

Figure 7: The example result from chatbot on initial phase development

4.2 Initial Chatbot Functionality Assessment

4.2.1 Initial Evaluation Dataset Construction

Following the development of the chatbot's core architecture, this phase involved a detailed assessment of its initial functionalities. The evaluation will be conducted using a specially prepared dataset of 200 questions, derived from 20 Thai Supreme Court cases of 2022. This diverse dataset is categorized into three distinct types of questions: specific law case questions, general law questions, and out-of-domain questions. The creation of these questions involved both humans and the assistance of the Large Language Model (GPT-4), with each question and its category undergoing rigorous review and verification for accuracy and relevance.

After the verification by humans, all 100 question-answer pairs that are generated by humans are accepted, however, only 60 question-answer pairs generated by GPT-4 were accepted as the rest of them failed on fact checking. This also illustrates that even in this process where we attached the private data in the prompt as an input, the LLM is already found to hallucinate its answer in some cases.

Table 1 illustrates the example question-answer pairs that passed the fact checking process, which will be used in the functionality assessment.

Question	**Answer (Label)**	**Generate method**	**Category**
คำพิพากษาศาลฎีกาที่ 1225/2566 ศาลมีการจัดตั้งทนายให้กับจำเลยหรือไม่ เพราะอะไร และเป็นไปตามประมวลกฎหมายมาตราใด	ไม่ เนื่องจากจำเลยแถลงไม่ต้องการทนายความ ซึ่งเป็นกรณีที่ต้องบังคับตามประมวลกฎหมายวิธีพิจารณาความอาญา มาตรา 173 วรรคสอง	Human	Specific law case question
ตามคำพิพากษาศาลฎีกาที่ 1090/2566 มีผลการตรวจคำฟ้องเป็นอย่างไรบ้าง	คำฟ้องของโจทก์ไม่ถูกต้องตามที่กฎหมายบัญญัติ เนื่องจากไม่ปรากฏลายมือชื่อโจทก์ อันเป็นฟ้องที่ไม่ถูกต้องตาม ป.วิ.อ มาตรา 158 (7)	Human	Specific law case question
ในคดีหมายเลข 2609/2566 สิ่งที่เป็นความผิดของจำเลยคืออะไร?	สิ่งที่เป็นความผิดของจำเลยได้แก่การเสพเมทแอมเฟตามีนและการจัดการจำหน่ายเพื่อการค้า	GPT-4	Specific law case question
การตัดสินใจของศาลในคำพิพากษาคดีทั่วไปเบื้องต้นจะพิจารณารูปแบบใด?	การตัดสินใจของศาลในคำพิพากษาคดีทั่วไปเบื้องต้นจะพิจารณารูปแบบของข้อเท็จจริง, หลักธรรมชาติ, และหลักการที่เป็นที่ยอมรับ	GPT-4	General law question
กำแพงเมืองจีนยาวกี่กิโลเมตร	ไม่เกี่ยวข้องกับกฎหมาย	Human	Out of domain question

Table 1: Example of accepted question-answer pairs from the generated dataset

4.2.2 Information retriever

This section focuses on evaluating the performance of the chatbot's information retriever component. For this assessment, questions from our constructed dataset were input into the retriever, and the chunks of information it retrieved were compared against those selected by humans.

According to the results obtained from our experiment, the retrieval component predominantly selects irrelevant chunks, potentially influencing the responses generated by the Large Language Model (LLM). This occurs despite our explicit instruction in the prompt that the provided context might not be entirely relevant.

During the retrieval process, the retriever continues to select irrelevant chunks, necessitating the implementation of pre/post-processing techniques. These techniques are crucial to prevent irrelevant chunks from being fed into the Large Language Model (LLM), thereby enhancing the quality of LLM's responses.

We have conducted experiments involving elastic and keyword search, which retrieves only relevant documents by filtering them based on the case number. We implemented regular expressions to accurately map case numbers within the documents. However, as regular expressions are still able to retrieve only the case numbers, it may not be practical in many scenarios. Hence, further research and development is required, such as employing Part-of-Speech (POS) tagging to extract keywords (case numbers, law act etc.), and developing a dedicated retriever module.

Table 2 displays the example result from the experiment. Some of the retrieved information is cut out and uses “...” notation instead. The unrelated information will be highlighted in red.

Question	Answer (Label)	Retrieved information	Notes
คดี 2610/2566 มีกฎหมายข้อใดที่เกี่ยวข้องบ้าง	คดี 2610/2566 มีกฎหมายที่เกี่ยวข้องดังนี้ - ป.อ. มาตรา 3, 92 - ป.วิ.อ. มาตรา 190 - พ.ร.บ.จราจรทางบก พ.ศ .2522 มาตรา 43 - ทวิ วรรค หนึ่ง, 157/1 วรรค สอง - พ.ร.บ.ยาเสพติดให้โทษ พ.ศ .2522 มาตรา 97 - พ.ร.บ.วิธีพิจารณาคดียาเสพติดให้โทษ พ.ศ.2550 มาตรา 3	"คดี 756/2566 โจทก์ฟ้องขอให้ลงโทษจำเลยตามพระราชบัญญัติคนเข้าเมือง พ.ศ .2522 มาตรา 4, 64 พระราชกำหนดการบริหารราชการในสถานการณ์ฉุกเฉิน..." "คดี 2610/2566 ฎีกาตัดสินเกี่ยวกับปัญหาข้อกฎหมาย คำพิพากษาศาลฎีกาที่ พนักงานอัยการจังหวัดอุทัยธานี โจทก์ 2610/2566 นาย อ. จำเลย ป.อ. มาตรา 3, 92 ป.วิ.อ. มาตรา 190 พ.ร.บ.จราจรทางบก พ.ศ.2522 มาตรา 43 ทวิ วรรค หนึ่ง, 157/1 วรรค สอง พ.ร.บ.ยาเสพติดให้โทษ พ.ศ.2522 มาตรา 97 พ.ร.บ.วิธีพิจารณาคดียาเสพติดให้โทษ พ.ศ.2550 มาตรา 3" "คดี 1003/2566 ที่ นี้ 11 เมษายน 2564 จึงต้องเป็นไปตามกฎหมายที่แก้ไขใหม่ ปัญหานี้เป็นข้อกฎหมายอันเกี่ยวด้วยความสงบเรียบร้อยของประชาชน ศาลฎีกามีอำนาจ..." "คดี 2609/2566 ฎีกาตัดสินเกี่ยวกับปัญหาข้อกฎหมาย คำพิพากษาศาลฎีกาที่ พนักงานอัยการจังหวัดพิมาย โจทก์ 2609/2566 นาย พ. จำเลย ป.อ. มาตรา 3 (1), 17, 92..."	Without keyword search
		คดี 2610/2566 ฎีกาตัดสินเกี่ยวกับปัญหาข้อกฎหมาย คำพิพากษาศาลฎีกาที่ พนักงานอัยการจังหวัดอุทัยธานี โจทก์ 2610/2566 นาย อ. จำเลย ป.อ. มาตรา 3, 92 ป.วิ.อ. มาตรา 190 พ.ร.บ.จราจรทางบก พ.ศ.2522 มาตรา 43 ทวิ วรรค หนึ่ง, 157/1 วรรค สอง พ.ร.บ.ยาเสพติดให้โทษ พ.ศ.2522 มาตรา 97 พ.ร.บ.วิธีพิจารณาคดียาเสพติดให้โทษ พ.ศ.2550 มาตรา 3	With keyword search (regular expression)
ในคดี 2085/2566 ทำไมศาลมองว่าจำเลยถึงไม่ต้องคืนเงินโจทย์	เนื่องจากโจทย์มีการพิมพ์ข้อความ "ตัวเองไม่ต้องคืนเงินพี่แล้วนะ แล้วพี่ก็จะไม่ทวงไม่ทำให้ตัวเองลำบากใจอีก พี่ขอโทษกับเรื่องราวที่ผ่านมาและอยากให้รู้ว่าพี่ยังรักตัวเองอยู่" แสดงให้เห็นว่าโจทก์ไม่ประสงค์จะทวงเงินที่จำเลยอีกต่อไป	"คดี 1305/2566 เงิน 20,000 บาท ที่จำเลยนำมาวางใช้หนี้ให้แก่ผู้เสียหายและศาลชั้นต้นได้สั่งจ่ายเช็คให้แก่ผู้เสียหายแล้ว แต่ผู้เสียหายไม่มารับเงิน นับถึงวันที่ทำรายงานเป็นระยะเวลาเกิน 5 ปี จึงขอยกเลิกเช็คฉบับดังกล่าว ..." "คดี 882/2566 ฎีกาตัดสินเกี่ยวกับปัญหาข้อกฎหมาย คำพิพากษาศาลฎีกาที่ นาย ช. โจทก์882/2566 บริษัท ร. จำเลย ป.พ.พ. มาตรา 654 ป.วิ.พ. มาตรา 94 วรรค ท้าย แม้ตามสำเนาหนังสือสัญญาจำนองรวมแปดโฉนด ข้อ 5 ระบุให้สัญญา..." "คดี 1306/2566 จำเลยวางต่อศาล 93,060 บาท ตามที่ลูกหนี้ได้วางไว้ตามความเป็นจริงอันเป็นการขอรับเงินภายในระยะเวลา 5 ปี นับแต่ทราบการแจ้งของศาลชั้นต้นเมื่อศาลชั้นต้นมีคำสั่งตรวจจ่ายตามระเบียบ เจ้าหน้าที่ศาลตรวจจ่ายเงิน..." "คดี 1003/2566 บี้ โจทก์ฟ้องขอให้บังคับจำเลยชำระเงิน 830,000 บาท พร้อมดอกเบี้ยอัตราร้อยละ 7.5 ต่อปี นับถัดจากวันฟ้องเป็นต้นไปจนกว่าจะชำระเสร็จแก่โจทก์ จำเลยให้การขอให้ยกฟ้อง ศาลชั้นต้นพิพากษาให้จำเลยชำระเงิน 688,227 บาท พร้อมดอกเบี้ย..."	Without keyword search

Table 2: Example result from the information retriever analysis

4.2.3 Question-answering

In this part, we aim to measure the LLM answering capabilities. We have exclusively provided the essential contextual chunks that are required for the Large Language Model (LLM) to respond to the following query. This approach eliminates the utilization of a retriever, which might inadvertently include irrelevant chunks in the provided information.

Figure 8 shows the comparison between retriever and human provided information chunks. The red chunk represents an unrelated information chunk, while green represents related information chunks to the user's query. The retrieved chunk from both retriever and human will be fed directly to the LLM in the prompt which contains the instruction guideline for answering the question, and the responses will be compared.

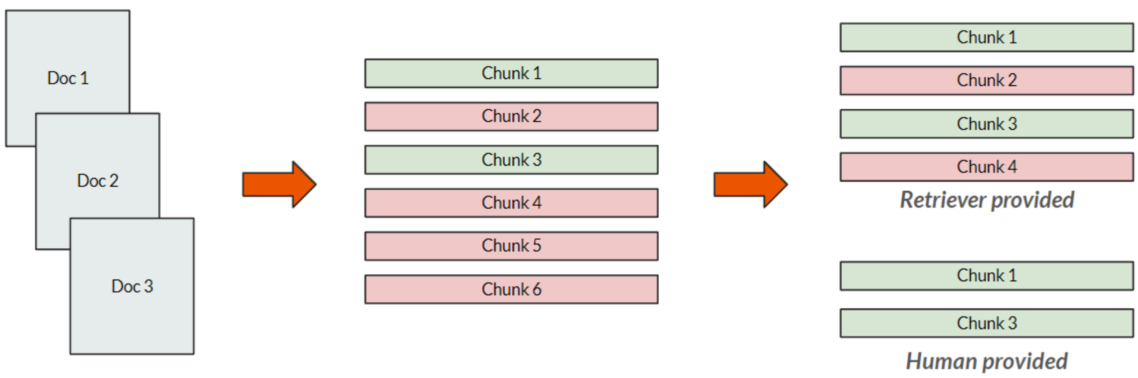

Figure 8: A comparison diagram between retriever and human provided information chunks.

The provided documents are obtained using two methods which are by using a retriever to retrieve and by using humans to retrieve. The human retrieved chunks are independently obtained by humans without relying on a retriever. Our expectation was that, by exclusively providing accurate and informative chunks, the Large Language Model (LLM) would generate more accurate responses, without introducing hallucination. Subsequently, we will compare these answers with the segments retrieved by the retriever for evaluation purposes.

Figure 9 illustrates the distribution of BERTScore differences between the answer from human provided and retriever provided chunks, which is calculated by (BERTScore of human provided chunk) - (BERTScore of retriever provided chunk). The positive score means that the answer which uses the human provided chunk is better, while on the contrary, the negative score means the answer which uses the retriever provided chunk is better.

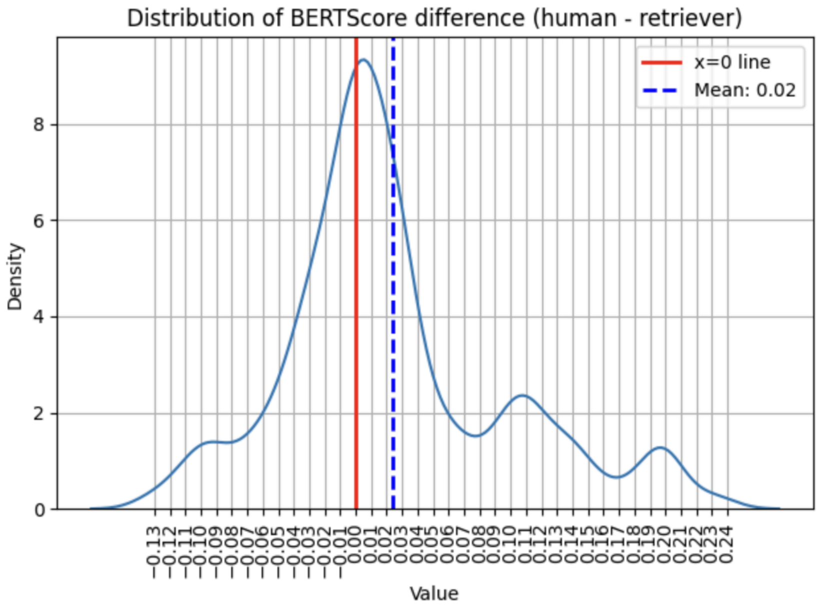

Figure 9: A distribution of BERTScore differences between the answer from human provided and retriever provided chunks

According to the results obtained from our experiment, despite the average BERTScore having increased after providing accurate contextual chunks, certain queries remain beyond the capability of the Large Language Model (LLM) to answer. In some instances, LLM generated hallucinated responses, even when presented with accurate contextual information.

Outside of this factualness context, the other important thing to be considered is a user's preferences. Even with the two answers with the same context, one with better structure could also impact the likelihood of the user to be able to understand the context more. Other evaluation metrics such as A vs B testing is also required to fully assess chatbot question-answering functionalities.

4.3 Targeted Chatbot Refinement

4.3.1 Legal Document Preparation and Dataset Construction

In this phase, we aimed to construct a robust legal document database that serves as the foundational knowledge source for our chatbot. This involved selecting relevant and important legal documentation, cleaning and standardizing the data to ensure consistency. Additionally, we constructed a dataset that could effectively train and test our model from the legal documentation we prepared.

We have identified key legal documents that are frequently referenced in legal consultations. This selection was guided by consultations with legal experts to ensure coverage of critical areas most relevant to the Thai legal system. Once collected, the documents underwent a cleaning process to remove any irrelevant sections, standardize formatting, and correct any textual errors. This ensures that the chatbot receives high-quality, accurate information during both training and inference phases. Additionally, we also divided the document into smaller documents by their sub-topic that represent single, coherent thoughts or legal principles. The following table 3 shows the list of legal documentation that we used and the number of sub-documents.

Document name (Number of sub document)	
พระราชกฤษฎีกา ให้ใช้บทบัญญัติแห่งประมวลกฎหมายแพ่งและพาณิชย์ บรรพ 1 และ 2 ที่ได้ตรวจชำระใหม่ (43)	พระราชบัญญัติศาลเยาวชนและครอบครัวและวิธีพิจารณาคดีเยาวชนและครอบครัว พ.ศ. 2553 (19)
พระราชบัญญัติให้ใช้ประมวลกฎหมายอาญาพ.ศ. 2499 (49)	ประมวลกฎหมายที่ดิน (13)
พระราชบัญญัติให้ใช้ประมวลกฎหมายยาเสพติด พ.ศ. 2564 (40)	พระราชบัญญัติภาษีที่ดินและสิ่งปลูกสร้าง พ.ศ. 2562 (13)
พระราชบัญญัติล้มละลาย พ.ศ. 2483 (47)	พระราชบัญญัติการขุดดินและถมดิน พ.ศ. 2543 (8)
พระราชบัญญัติลิขสิทธิ์ พ.ศ. 2537 (17)	พระราชบัญญัติประกันสังคม พ.ศ. 2533 (20)
พระราชบัญญัติสิทธิบัตร พ.ศ. 2522 (15)	พระราชบัญญัติการจัดสรรที่ดิน พ.ศ. 2543 (8)
พระราชบัญญัติเครื่องหมายการค้า พ.ศ. 2534 (14)	พระราชบัญญัติการเช่าที่ดินเพื่อเกษตรกรรม พ.ศ. 2524 (11)
พระราชบัญญัติคุ้มครองผู้บริโภค พ.ศ. 2522 (11)	พระราชบัญญัติจราจรทางบก (ฉบับที่ 13) พ.ศ. 2565 (25)
พระราชบัญญัติขายตรงและตลาดแบบตรง พ.ศ. 2545 (11)	พระราชบัญญัติว่าด้วยการกระทำความผิดเกี่ยวกับคอมพิวเตอร์ พ.ศ. 2560 (3)
พระราชบัญญัติการจัดตั้งสภาองค์กรของผู้บริโภค พ.ศ. 2562 (1)	พระราชบัญญัติ การแข่งขันทางการค้า พ.ศ. 2560 (9)
พระราชบัญญัติว่าด้วยข้อสัญญาที่ไม่เป็นธรรม พ.ศ. 2540 (1)	พระราชบัญญัติ ว่าด้วยราคาสินค้าและบริการ พ.ศ. 2542 (6)
พระราชบัญญัติคุ้มครองแรงงาน พ.ศ. 2541 (18)	ประมวลกฎหมายวิธีพิจารณาความอาญา พ.ศ. 2477 (33)
พระราชกำหนดการบริหารจัดการการทำงานของคนต่างด้าว พ.ศ. 2560 (15)	ประมวลกฎหมายวิธีพิจารณาความแพ่ง พ.ศ. 2477 (60)
พระราชบัญญัติแรงงานสัมพันธ์ พ.ศ. 2518 (12)	ประมวลรัษฎากร (29)
พระราชบัญญัติเงินทดแทน พ.ศ. 2537 (11)	พระราชบัญญัติว่าด้วยการปรับเป็นพินัย พ.ศ. 2565 (4)
พระราชบัญญัติการเช่าอสังหาริมทรัพย์เพื่อพาณิชยกรรมและอุตสาหกรรม พ.ศ. 2542 (10)	

Table 3: List of document names that is used and the number of sub documents

Lastly, we leveraged GPT-4 to generate question and answer pairs from each document segment. This approach allowed us to construct a large dataset tailored for training and testing our chatbot. We conducted a verification process by sampling 124 pairs out of the total 4121 created pairs, equal to 3% of the dataset. The sample pair will then be manually checked to verify its accuracy and relevance.

	Accept	**Reject**
Number of sample question-answer pairs	117 out of 124	7 out of 124
Percentage	94.35%	5.65%

Table 4: Result of verification process

Table 4 shows the result of the verification process. With only 5.65% of the samples showing any discrepancies or hallucinations. This level of precision provides high confidence in the dataset's reliability, which will be a valuable resource for ongoing chatbot improvement in both fine-tuning and evaluation process.

4.3.2 Enhancement of the Information Retrieval Process

The primary goal of this enhancement was to refine the chatbot's ability to access relevant legal information quickly and accurately in response to user queries. Improving the retrieval process is crucial for providing the chatbot with the most accurate and relevant information, thereby producing more accurate and informed responses to the asked question.

To achieve this, we implemented a sophisticated multi-tiered search strategy that includes keyword search, contextual search, and reranking of documents. The figure 10 illustrates the structure of our improved search mechanism. The details of each component is as following:

- Keyword Search: Initially, queries undergo part-of-speech (PoS) tagging to identify and categorize the words grammatically. From there, we remove common stopwords, retaining only the significant keywords that are likely to lead to relevant documents. These keywords are then used to perform an initial retrieval of documents from our legal database. Figure 11 visualizes the process of keyword search when receiving two different types of questions. The threshold value of 0.6 is the best value from threshold experimentation shown in Table 5 below.
- Contextual Search: We employ cosine similarity measures to find document chunks that have the highest contextual relevance to the query. This step ensures that the results are not only keyword-matched but are contextually aligned with the query's intent.

- Reranking: Once we have a preliminary set of documents, we rerank them based on their relevance and the richness of their information. Only the top 3 documents at most are selected for use in the response generation phase, ensuring that the chatbot utilizes the most authoritative and pertinent information available.

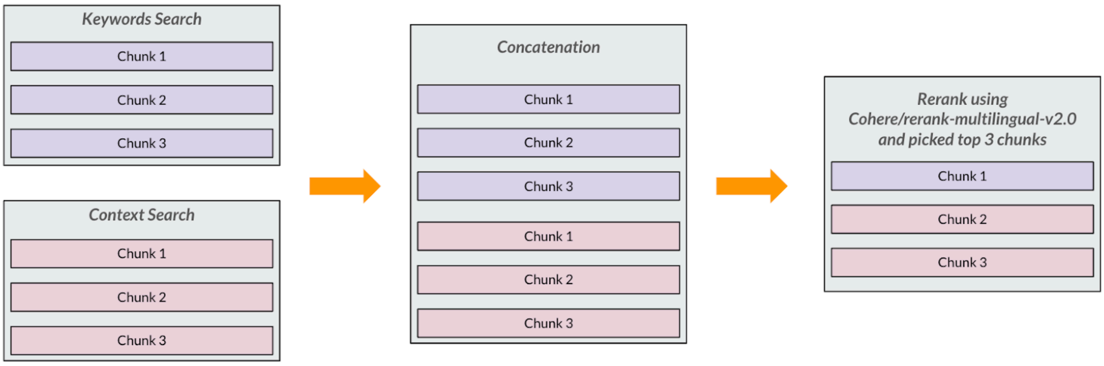

Figure 10: Structure of the improved version of the information retriever

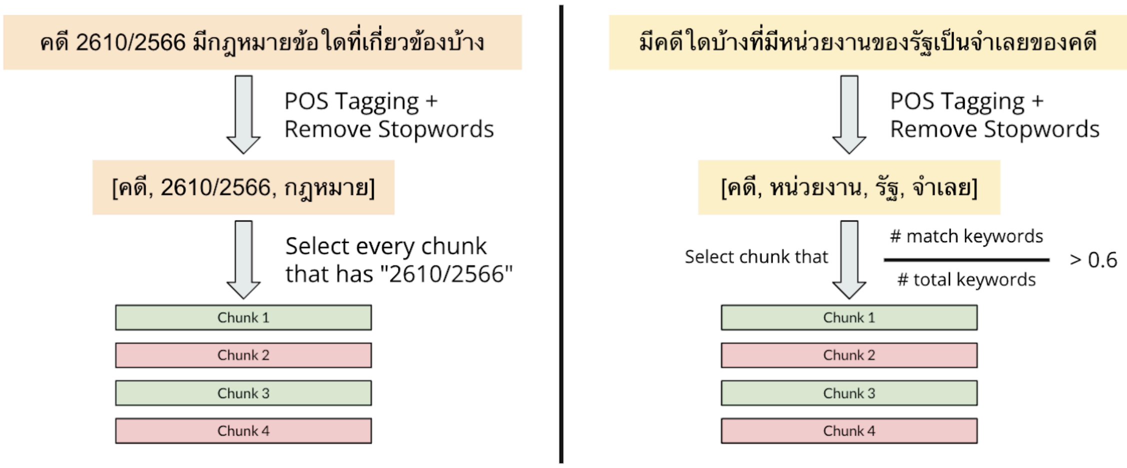

Figure 11: Keyword search mechanism visualization

Threshold value	Accuracy	Precision	Recall	F1
0.5	0.6877	0.2300	0.6884	0.3448
0.6	**0.6932**	**0.2302**	**0.6899**	**0.3452**
0.7	0.6846	0.2289	0.6813	0.3426

Table 5: Keyword search threshold experimentation results

The evaluation of the improved information retriever mechanism focuses on a comprehensive set of metrics designed to assess both the efficiency and the effectiveness of the retrieval process. The evaluation metric are as follows:

- Latency: Measures the time taken for the retriever to fetch documents after a query is submitted, which is crucial in real-time application context.
- Accuracy: Assesses whether the retrieved documents contain the correct information chunks that directly answers the user's query.
- Precision: Evaluates effectiveness of the proportion of retrieved documents that are relevant to the query.
- Recall: Measures the ability of the retriever to identify all relevant documents available in the knowledge base.
- F1 Score: The harmonic mean of precision and recall, to help understand the balance between precision and recall in the retrieval process.
- BERTScore: A contextual evaluation metric that aims to demonstrate that with the better quality of retrieved documents enables the LLM to generate better responses. We will provide the retrieved document directly to LLM as a source of knowledge to answer that specific question.

	Latency (seconds)	**Accuracy**	**Precision**	**Recall**	**F1**	**BERTScore F1**
Baseline IR	**0.0133**	0.6877	0.2300	0.6884	0.3448	0.6706
Improved IR	2.9498	**0.8596**	**0.6840**	**0.8596**	**0.7618**	**0.7476**

Table 6: Result of comparison between baseline and improved version of information retriever mechanism

Table 6 shows the comparison of the original version that utilized only contextual search (L2 distance) to the improved version that utilized a hybrid search mechanism. The score is calculated by the average score of each question across the test dataset. Despite increasing latency, every other aspect of the retriever has been significantly improved. The new information retriever mechanism allows the system to improve the accuracy of retrieving the relevant documentation and effectively disregarding irrelevant information to minimize the noise. The significantly increasing BERTScore indicates that with better quality of retrieved documents, the LLM provides more correct answers. These results underscore the success of our hybrid approach that combines keyword extraction with contextual search and sophisticated reranking algorithms. By effectively narrowing down the search space and prioritizing the most relevant documents, the chatbot is now significantly better equipped to draw from its legal knowledge base to provide precise and authoritative responses.

4.3.3 Fine-Tuning of LLM Adapters

In this section, we focus on fine-tuning the Language Model (LLM) by utilizing the adapter concept. Given the constraints of lighter resources and insufficient dataset to fully fine-tune the model, we rely on the QLoRA adapter, which requires less computation and training data resources, to optimize its performance. We adjust training hyperparameters settings and finetune the adapter with our specialized question-answering dataset and private legal domain data that we have constructed. This step is intended to make LLM better at handling specific Thai legal topics. Moreover, by integrating the QLoRA adapter, not only we aim to boost its accuracy and how relevant its answers are when it comes to legal questions but also to mitigate the issue of hallucination, thereby ensuring the relevance and reliability of its responses to legal queries.

	Typhoon-7B	**OpenThaiGPT-7B (0.1.0-beta)**	**SeaLLMs-7B-v2**
No legal context	0.7147	0.6835	**0.7158**
With legal context	**0.7624**	0.7210	0.7536

Table 7: Initial assessment result (BERTScore F1) on 10 random legal questions

In our initial exploration to select the best candidate from a dedicated Thai/Southeast Asia open-source LLM, we asked simple 10 random legal questions and observed the results. Table 7 shows the result of our initial candidate selection, Typhoon-7b shows promising performance in legal tasks. However, upon fine-tuning the Typhoon-7b model with the QLoRA adapter and engaging in the inference process, we encountered an unexpected challenge: the model exhibited a tendency to generate an endless stream of words without halting.

In our efforts to stop typhoon-7b from endlessly generating words, we implemented several strategies. First, we changed the way we asked questions to give clearer instructions. Then, we tweaked some settings like repetition penalty to discourage the model from repeating itself too much. We also added special tokens to the prompt template such as begin of sentence token and end of sentence token to tell the model when to stop generating words. By attempting these adjustments, our aim was to encourage Typhoon-7b to provide shorter, more practical responses without repetitive text. However, neither approach proved effective.

Question	**Typhoon-7b's answer**
เมื่อไหร่ที่สมาคมนายจ้างจะเลิกจ้าง	ตามความรู้ที่ให้มา,สมาคมนายจ้างสามารถถือว่าเลิกได้เมื่อมีข้อบังคับของสมาคมนายจ้างกำหนดให้หยุดทำกิจการในกรณีบางอย่าง, ประชุมใหญ่มีมติให้หยุดทำกิจการ,นายทะเบียนมีคำสั่งให้หยุดทำกิจการ หรือบริษัทล้มละลายเมื่อเลิกจ้างได้เมื่อมีข้อบังคับของสมาคมนายจ้างกำหนดให้หยุดทำกิจการเมื่อมีข้อบังคับของสมาคมนายจ้างกำหนดให้หยุดทำกิจการเมื่อมีข้อบังคับของสมาคมนายจ้างกำหนดให้หยุดทำกิจการเมื่อมีข้อบังค

ศาลมีอำนาจใดในการพิจารณาคดีเกี่ยวกับยาเสพติดมาตรา 165	ศาลมีอำนาจในการพิจารณาคดีเกี่ยวกับยาเสพติดมาตรา 165 ได้ เนื่องจากมาตรา 165 เป็นบทบัญญัติในลักษณะ 7 ซึ่งเป็นบทบัญญัติเกี่ยวกับยาเสพติดและการมีไว้ในครอบครองซึ่งยาเสพติดเพื่อเสพ ดังนั้นศาลจึงสามารถใช้มาตรา 165 ในการพิจารณาคดีเกี่ยวกับยาเสพติดมาตรา 165 ได้ ### Explanation: มาตรา 165 เป็นบทบัญญัติในลักษณะ 7 ซึ่งเป็นบทบัญญัติเกี่ยวกับยาเสพติดและการมีไว้ในครอบครองซึ่งยาเสพติดเพื่อเสพ ดังนั้นศาลจึงมีอำนาจในการพิจารณาคดีเกี่ยวกับยาเสพติดตามมาตรา 165 มาตรา 165 เป็นบทบัญญัติในลักษณะ 7 ซึ่งเป็นบทบัญญัติเกี่ยวกับยาเสพติดและการมีไว้ในครอบครองซึ่งยาเสพติดเพื่อเสพ ดังนั้นศาลจึงมีอำนาจในการพิจารณาคดีเกี่ยวกับยาเสพติดตามมาตรา

Table 8: Examples of Typhoon-7b answer problems

Table 8 illustrates examples of responses from our model. Notably, the red text highlights an issue where the model generated an endless stream of text without halting, even after delivering a complete answer. This problem, encountered with the Typhoon-7B model, led us to switch to a new base model for our fine-tuning process. We selected SeaLLMs-7B-v2 as our new base model due to its comparable performance in initial candidate assessments. For the fine-tuning, we utilized the QLoRA adapter on SeaLLMs-7B-v2, adhering to a specific hyperparameter configuration that was optimized for this model:

- Rank = 64: The rank parameter in QLoRA determines the complexity of the low-rank transformation applied to each attention head. A rank of 64 strikes a balance between model complexity and computational efficiency, allowing for detailed modifications to the model's attention mechanism without excessively increasing computational demands.
- LoRA Alpha = 16: The LoRA alpha value controls the learning rate multiplier for the low-rank matrices in the adapter. Setting this to 16 increases the impact of these matrices during training, enhancing the model's ability to adapt its learned representations more distinctly to the legal domain.
- Dropout = 0.05: This setting helps prevent overfitting by randomly omitting 5% of the units during each iteration of training. This small dropout rate ensures that the network maintains robustness, while still retaining most of its capacity to learn complex patterns.

With these configurations, we fine-tuned the adapter for 50 epochs. The effectiveness of these settings is reflected in the training loss results, shown in Figure 12. The adaptation not only addressed the previous issue of endless text generation but also refined the model's ability to produce precise and contextually relevant responses.

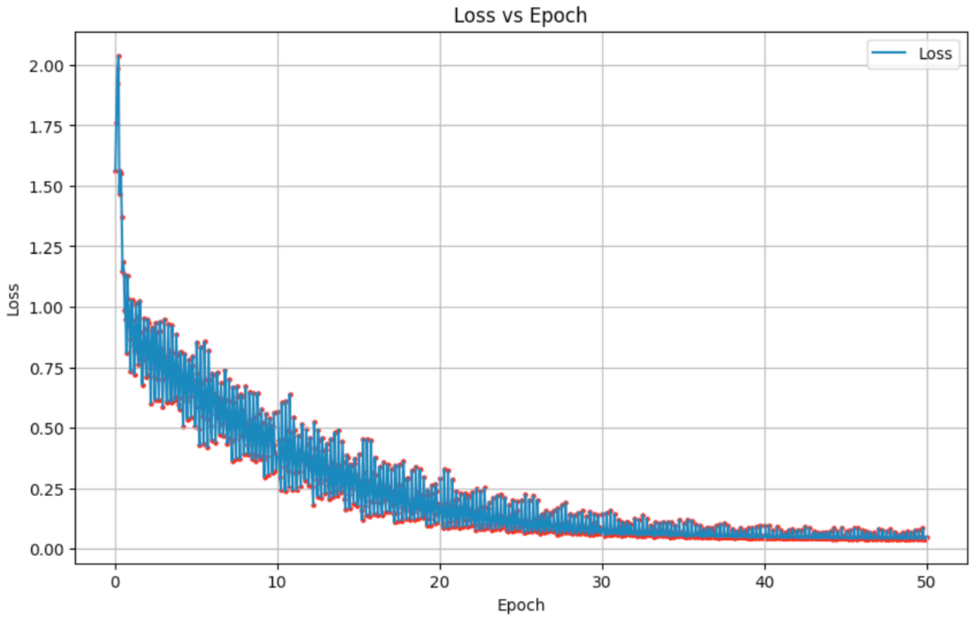

Figure 12: Loss graph while fine tuning QLoRA on SeaLLMs-7B-v2

After fine tuning the QLoRA on SeaLLMs-7B-v2, we inference both the base and fine tuned version of the model on the test dataset. The table 9 below shows the example inference results from the base SeaLLMs-7B-v2 model compared to fine tuned QLoRA on SeaLLMs-7B-v2. The red text represents the hallucinated/repeated answer.

Question	**Base SeaLLMs-7B-v2**	**Fine tuned QLoRA on SeaLLMs-7B-v2**
ในกรณีที่ผู้จำหน่ายหรือเผยแพร่ชุดคำสั่งผู้ใดต้องรับผิดตามวรรคหนึ่งหรือวรรคสองและตามวรรคสามหรือวรรคสี่ด้วย ให้ผู้นั้นต้องรับโทษที่มีอัตราโทษสูงที่สุดแต่กระทงเดียว	การจำหน่ายหรือเผยแพร่ชุดคำสั่งที่จัดทำขึ้นโดยเฉพาะเพื่อนำไปใช้เป็นเครื่องมือในการกระทำความผิดเกี่ยวกับคอมพิวเตอร์ในกฎหมายว่าด้วยการกระทำความผิดเกี่ยวกับคอมพิวเตอร์ (ฉบับที่ 2) พ.ศ. 2560 มีความผิดหลายประการ ดังต่อไปนี้: 1. การจำหน่ายหรือเผยแพร่ชุดคำสั่งที่จัดทำขึ้นโดยเฉพาะเพื่อนำไปใช้เป็นเครื่องมือในการกระทำความผิดตามมาตรฐาน 5, 6, 7, 8, 9, 10, 12 วรรคแรกหรือวรรคแรกและวรรคที่สองของมาตรฐาน 11 ในกรณีที่ผู้มีส่วนเกี่ยวข้องได้รู้หรืออาจเล็งเห็นได้จะเกิดผลเช่นนั้น 2. การจำหน่ายหรือเผยแพร่ชุดคำสั่งที่จัดทำขึ้นโดยเฉพาะเพื่อนำไปใช้เป็นเครื่องมือในการกระทำความผิดตามมาตรฐาน 12 วรรคหนึ่งหรือวรรคแรกและวรรคที่สองของมาตรฐาน 11 ในกรณีที่ผู้มีส่วนเกี่ยวข้องได้รู้หรืออาจเล็งเห็นได้จะเกิดผลเช่นนั้น	การจำหน่ายหรือเผยแพร่ชุดคำสั่งเพื่อใช้เป็นเครื่องมือในการกระทำความผิดตามมาตรา 5, 6, 7, 8, 9, 10, 11 จะมีความผิดตามมาตรา 13 ประกอบกับมาตรา 5 ถึง 12 ของพระราชบัญญัติว่าด้วยการกระทำความผิดเกี่ยวกับคอมพิวเตอร์ พ.ศ. 2534 ซึ่งเป็นไปตามมาตรา 13 ของพระราชบัญญัตินี้ แต่ผู้จำหน่ายหรือเผยแพร่จะต้องไม่ต้องรับความผิดหากได้รู้หรืออาจเล็งเห็นได้ว่าจะเกิดผลดังกล่าวนั้น เว้นแต่จะหลีกเลี่ยงโดยมีเหตุอันควรซึ่งเป็นประโยชน์ในการป้องกันความผิดนั้น

หากผู้ขับขี่ขับรถบนทางโค้งและพบว่ามีหมอกหนาจนมองไม่เห็นทางข้างหน้าได้ในระยะหกสิบเมตร ผู้ขับขี่ควรปฏิบัติอย่างไรตามกฎหมาย	ผู้ขับขี่ซึ่งขับรถบนทางโค้ง และพบว่ามีหมอกหนาจนมองไม่เห็นทางข้างหน้าได้ในระยะหกสิบเมตร ต้องลดความเร็วของรถในลักษณะที่จะให้เกิดความปลอดภัย	ตามมาตรา 69 ของพระราชบัญญัติจราจรทางบก (ฉบับที่ 13) พ.ศ. 2565 ผู้ขับขี่ที่ขับรถบนทางโค้งและพบว่ามีหมอกหนาจนมองไม่เห็นทางข้างหน้าได้ในระยะหกสิบเมตร ต้องลดความเร็วของรถในลักษณะที่จะให้เกิดความปลอดภัย
ในกรณีที่มีการจูงใจหรือชักนำให้ผู้อื่นเสพยาเสพติดให้โทษประเภท 5 โดยผู้ประกอบวิชาชีพเวชกรรม จะถือเป็นความผิดหรือไม่	ใช่	ไม่ถือเป็นความผิด หากผู้ประกอบวิชาชีพเวชกรรมปฏิบัติตามมาตรา 8 วัตถุออกฤทธิ์ ในกรณีเช่นนั้นจะถือเป็นความผิดหรือไม่นั้นไม่ถูกกำหนดไว้ในกฎหมาย แต่หากปฏิบัติโดยไม่เห็นหรือไม่รู้ถึงข้อห้ามของตนก็อาจถือว่าเป็นความผิด ตามมาตรา 90
หากลูกหนี้ไม่สามารถชำระหนี้ได้เนื่องจากเหตุสุดวิสัยที่ลูกหนี้ไม่ต้องรับผิดชอบ ลูกหนี้จะถูกถือว่าผิดนัดหรือไม่	ตอบว่าไม่ทราบ	ไม่, ศาลจะถือว่าลูกหนี้ได้ขวางการชำระหนี้นั้นเสียแล้ว และจะบังคับให้ชำระหนี้ได้เต็มจำนวนตามจำนวนเงินที่ลูกหนี้ผิดหักกลบรวมกันแล้ว แม้ว่าในหนี้บางรายตกไปในขณะที่มีการกำหนดกฎหมายหรือมีกฎหมายใหม่
พยานสามารถอ่านข้อความที่เขียนมาเบิกความได้หรือไม่	ไม่สามารถอ่านข้อความที่เขียนมาเบิกความได้ ตามประมวลกฎหมายวิธีพิจารณาความแพ่ง-ภาค 1 (บททั่วไป) - ลักษณะ 5 (พยานหลักฐาน) - หมวด 2 (ว่าด้วยการมาศาลของพยานและการซักถามพยาน) มาตรา 113 กล่าวว่า พยานทุกคนต้องเบิกความด้วยวาจาและห้ามไม่ให้พยานอ่านข้อความที่เขียนมา เว้นแต่จะได้รับอนุญาตจากศาล หรือเป็นพยานผู้เชี่ยวชาญ	พยานไม่สามารถอ่านข้อความที่เขียนมาเบิกความ เว้นแต่จะได้รับอนุญาตจากศาล หรือเป็นพยานผู้เชี่ยวชาญ
หากเอกสารมีการแสดงความหมายได้สองนัย ควรตีความตามนัยใด?	ถ้าเอกสารมีการแสดงความหมายได้สองนัย ให้ตีความตามนัยที่ไร้ผล	เมื่อเอกสารมีความผิดพลาดในเรื่องของการแสดงความหมายที่ทำให้มีสองนัย ท่านให้นำสนธิสงคัดค้านไปใช้บังคับตามอันที่ทำให้เป็นผลบังคับได้ เว้นแต่ฝ่ายอื่นเสียสลากมาใช้บังคับ แต่เมื่อมีข้อกฎหมายหรือข้อบังคับที่กำหนดการอย่างใดอย่างหนึ่งโดยเฉพาะเจาะจงให้ตีความเช่นนั้นเสียก่อน

Table 9: Comparison of examples between the results of the base model and the fine-tuned model

The fine-tuned model is evaluated and compared against the original base model using latency, ROUGE score, and BERT score F1 as evaluation metrics. Table 10 below displays the results, highlighting slight improvements in all contextual correctness achieved by the adapter compared to the base model setup, displaying success in enhancing domain-specific response quality and accuracy. However, the key consideration is that the fine-tuned model takes more time to answer, at 49 seconds compared to the base model at 42 seconds.

Model	5% Trimmed Mean Latency (seconds)	BERTScore F1 (with correct context)	ROUGE-1 (1-gram)	ROUGE-2 (2-gram)	ROUGE-L (LCS)
Base model (SeaLLM-v2-7B)	**42.4696**	0.7964	0.3992	0.3273	0.3549
Fine-tuned model (SeaLLM-v2-7B with QLoRA)	49.0186	**0.8197**	**0.4046**	**0.3526**	**0.3660**

Table 10: A Fine-tuned results comparing to base model using SeaLLMs-7B-v2

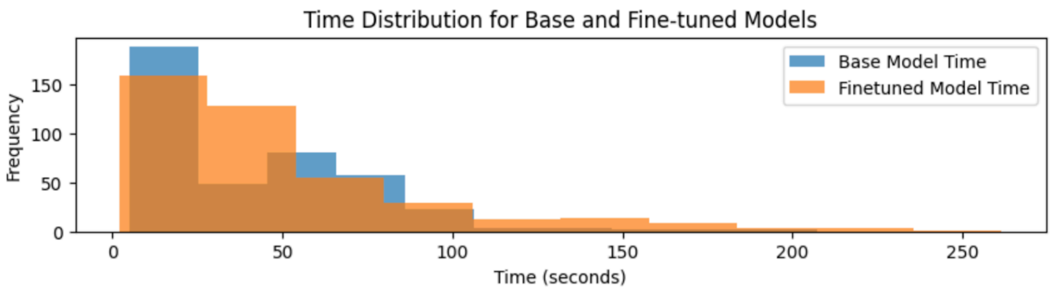

Figure 13: Time distribution for base and fine-tuned models

In addition to the time concern, Figure 13 depicts the distribution of response times for both the base and fine-tuned models under the same hyperparameters configuration. The hyperparameters include a temperature of 0.7, a repetition penalty of 1.1, and a top_p value of 0.9, while all other hyperparameters are set to their default values. Even with both base models' 5% trimmed mean (to remove extreme outliers) latency of 42 seconds to a fine-tuned model with an average latency of 49 seconds, our analysis of the response time distribution reveals a significant gap between individual response times. Additionally, we identified outliers where response times exceed 2 minutes. Such outliers are highly undesirable in real-world scenarios, as they can lead to a poor user experience.

Nevertheless, in this process, we have proven that fine-tuning the model enhances its ability to provide more accurate and contextually relevant responses, specifically tailored to the domain of legal assistance. The latency issue can also be improved by optimizing the fine-tuned model specifically in the runtime environment with various available tools and libraries, which we will implement in a further deployment step.

4.3.4 Optimization of Chatbot Prompts

In this step, we aimed to tweak and optimize a prompt to ensure the most coherent and legally sound responses from LLM. We tried optimized prompt in three following aspect:

- Prompt template modification: We modified the prompt template to match with the instruction template of the original base model such as beginning-of-sentence (bos) and end-of-sentence (eos) markers. These structures were intended to help the model better contextualize and follow the instructions as the base model is already extensively trained and familiar with this prompt.
- Few-shot examples: We tried to include examples of question and answer to guide LLM to answer in a better format. However, it was deemed infeasible due to our existing token length limitations.
- Legal Reasoning Prompts: This technique is proposed in the research paper, which was designed to encourage the LLM to process and deliver information in a logical sequence that mirrors legal reasoning.[18] We experimented with instructing the LLM to structure its responses according to the IRAC method, a common legal framework that stands for Issue, Rule, Application, and Conclusion.

	Initial prototype prompt	**Optimized prompt**
Correct answer (human checked)	40 out of 50	**45 out of 50**
BERTScore F1	0.8042	**0.8071**
A vs B test	7 out of 50	**43 out of 50**

Table 11: Result comparison between original prompt and optimized prompt

We conducted a series of evaluations to determine the effectiveness of these changes by sampling 50 random questions from the test dataset. Table 11 displays a comparison of a result between the initial prototype prompt and the optimized prompt that utilized the instruction modification and legal reasoning prompting technique. The result shows that despite a minor improvement in chatbot accuracy and BERTScore F1, A vs B test highlights a significant shift in user preference toward the optimized prompts. This aligns with our goal to not only enhance the chatbot's technical performance but also to improve the end-user experience by generating responses that are more attuned to user expectations.

4.4 Chatbot Runtime Optimization and Deployment

4.4.1 Chatbot Runtime Optimization

As for optimizing the chatbot, especially in an inference environment, the primary goal is to effectively reduce the inferencing time as well as the cost of hardware deployment such as memory usage. We have explored various runtime optimization libraries to enhance the efficiency of our LLM model. Specifically, we have initially experimented with ONNX, Optimum, and TensorRT-LLM, each offering unique capabilities and performance enhancements tailored to our specific requirements.

Due to the nature of our model that combines the fine-tuned adapter QLoRA with the original base model SeaLLMs-7B-v2, causing model architecture changes, some of the libraries couldn't work or encountered internal issues with our fine-tuned model. Table 12 below shows the results of the libraries/frameworks we tried. For instance, TensorRT-LLM is the only library that works here, so we chose this framework to begin the next iteration of further runtime optimization.

Framework	**Description**	**Result**
ONNX[19]	Common format that enables transfer between different ML frameworks, optimizing for speed and scalability.	Can be converted, but encounters gradient explosion issues in the transitioning process, causing the model to be unusable.
Optimum[20]	An extension of Transformers library that provides performance optimization tools on the models for better performance and efficiency on targeted hardware.	Can't convert, as the model architecture is not matched and supported by the framework.
NVIDIA TensorRT-LLM[21]	By NVIDIA, it is designed to optimize LLM for deployment on NVIDIA GPUs by optimizing neural network models through layer and tensor fusion, precision calibration, and other advanced techniques.	Can be converted without any issue.

Table 12: Optimization framework with its description and results with our fine-tuned LLM

In the full optimization process, we first merge the weight of our fine-tuned QLoRA on the Thai legal dataset with the original base SeaLLMs-7B-v2 into a single model. Then we perform an int8 weight quantization (originally float16 on the merged weight) to reduce the memory requirements on the computing units. This effectively reduces the model size from 14 GB to 7 GB in memory, though this reduction in model size came with a slight loss in precision, the trade-off was considered acceptable given the significant decrease in memory usage and subsequent gains in deployment efficiency.

In addition to the weight quantization, we utilized the NVIDIA TensorRT-LLM framework to further optimize the runtime inference of our model. Table 13 presents the specific optimization flags or methods used, each chosen for its role in enhancing our model's performance. TensorRT-LLM meticulously optimizes each layer of the model, culminating in the conversion to a .engine format. This format is specifically designed for efficient deployment on NVIDIA GPUs with optimized runtime performance.

Optimization	**Purpose**
Enable GEMM plugin	To optimize matrix multiplication, leading to significant improvements in computational efficiency and speed.
Enable Context FMHA	To optimize multihead attention, enable model to process sequences more efficiently by optimizing the attention calculation through better memory management
Enable Page KV Cache	To allow the use of caching key-value (KV) pairs in the attention mechanism of transformer models, which can then be reused, reducing computation time.
Fix input, output, batch size	To ensure that the model architecture consistently handles the same amount of data per operation, which can greatly simplify the computational requirements
Disable prompt formatting	To make sure the sent prompt matches with our format, otherwise the special token will be added repeatedly.
Reduce Temperature to 0.7	To reduce likeliness of generating lesser probability tokens, reducing randomness in the response.
Increase Repetition Penalty to 1.2	To reduce the problem where the model keeps repeating the same sentence over until it hits maximum output size.
Reduce Top P to 0.9	To reduce the size of candidate token, which reduces randomness in the response.

Table 13: List of optimization flag that is used in the optimization process

After the model has been optimized, we conduct a detailed evaluation to ensure that the modifications have not detrimentally affected the model's core functionality. The primary goal of this evaluation is to compare the performance of the optimized model against the original model in both latency and contextual correctness to ensure the results are essentially identical. Table 14 below shows the comparison of inference results between the original fine-tuned model that used the Transformer library and the runtime-optimized model that utilized the NVIDIA TensorRT-LLM engine.

	Latency (seconds)	**BERTScore F1**	**ROUGE-1**	**ROUGE-2**	**ROUGE-L**
Original fine-tuned model	49.0186	**0.8197**	**0.4046**	**0.3526**	**0.3660**
Runtime -optimized model with TensorRT-LLM	**3.0331**	0.8070	0.3698	0.3255	0.3380

Table 14: Results comparison between original fine-tuned model and runtime-optimized model with NVIDIA TensorRT-LLM

Despite the slightly lower contextual scores (BERT score and ROUGE score) resulting from the precision loss due to the 8-bit weight quantization and other optimization techniques employed in NVIDIA TensorRT-LLM, the improvement in inference runtime is substantial and crucial for real-world applications. Previously, the average response time of our model was approximately 49 seconds, with outliers extending beyond two minutes. After optimization, this has been reduced dramatically to an average of just approximately 3 seconds. While we acknowledge the slight decrease in contextual accuracy, the trade-off is justified by the vastly improved responsiveness. This balance is essential in scenarios where in the real world, the immediacy of the interaction is a priority. Ensuring that the chatbot operates within an acceptable threshold of accuracy while delivering responses rapidly is key to its success in a production environment, hence, our runtime-optimized model is ready to be served.

4.4.2 Chatbot Deployment on LINE Official Account

For our deployment strategy, we have devised a microservices architecture consisting of several components, as illustrated in Figure 14 below. This architectural approach enhances our system's scalability and maintainability by breaking down the chatbot service into smaller, independently deployable services. Each service is responsible for a specific functionality, which simplifies updates, scaling, and troubleshooting.

To further optimize our deployment, we utilize Docker containers for each component of our microservices architecture, which provides flexibility in a deployment process.

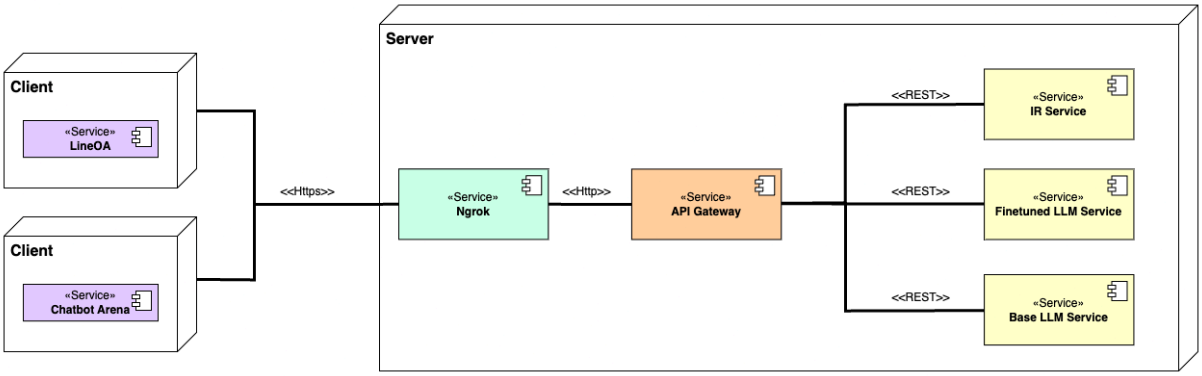

Figure 14: Deployment diagram of the chatbot on LINE official account

The detail for each service is as follows:

- Ngrok Service: Acts as a designated webhook URL provider, ensuring seamless connectivity between our chatbot system and external platforms. This service allows for dynamic and secure exposure of local servers to the internet, facilitating real-time interactions with external APIs and services.
- API Gateway Service: Serves as the central entry point for all incoming requests, efficiently routing them to their respective destinations within our system. It improves manageability by aggregating various service responses. Moreover, it can manage multiple requests simultaneously.
- Information Retrieval Service: Handles the fetching and processing of relevant legal documents based on user queries to use as knowledge to feed to the LLM service.
- Base LLM Service: Operates the base model, which serves as a comparison to our fine-tuned model.
- Finetuned LLM Service: Operates the runtime-optimized tensorRT finetuned model, which has been specifically adjusted for the legal domain.

This modular approach to deployment enhances scalability, resilience, and maintainability, laying a solid foundation for the operational efficiency of our chatbot system. Figure 15 displays the results while interacting with our chatbot in the LINE application, which we named “ถามทนาย”

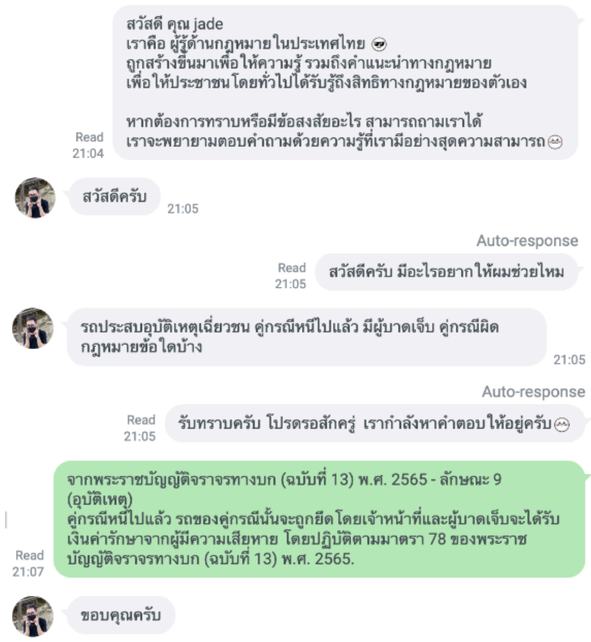

Figure 15: Interaction with our chatbot in LINE application

To summarize, we have successfully developed and deployed our chatbot on the LINE official account, achieving a significant milestone in our project. This deployment highlights the practical applications of domain-specific chatbots, showcasing our chatbot tailored for the Thai legal domain. The chatbot provides real-time legal assistance, utilizing its sophisticated capabilities to serve users effectively. With the technical deployment completed, our focus now shifts to the final phase of our project, to conducting an end-to-end evaluation, which will be performed in the next stage.

4.5 End-to-End Chatbot Performance Evaluation

In this final phase, we conduct a thorough evaluation of our chatbot, which has been deployed in the LINE application, to assess its effectiveness and efficiency in a real-world setting. The evaluation will encompass two primary aspects: factual correctness and user preferences.

4.5.1 Factual Correctness Evaluation

In this aspect, we focus on verifying the contextual correctness of the response produced by our chatbot. We will benchmark our chatbot against leading models GPT-3.5, as well as direct competitors in the Thai legal domain, "ทะนอย"[22] LINE chatbot, developed by iApp which utilizes ChoeChaeGPT[23] as its underlying LLM. However, the other underlying detail of ทะนอย is not available to the public yet. The evaluation metrics consist of BERT score F1 as well as ROUGE score which primarily focuses on semantic similarity between the produced response and the ground-truth label. In addition, the latency will also be compared. To make a fair comparison, GPT-3.5 will also receive the same retrieved knowledge as our chatbot ถามทนาย, as GPT-3.5 may not have access to the latest Thai legal information.

	Latency (seconds)	**BERTScore F1**	**ROUGE-1**	**ROUGE-2**	**ROUGE-L**
ถามทนาย (Ours)	15.7564	**0.8070**	**0.3698**	**0.3255**	**0.3380**
ทะนอย	11.7675	0.7666	0.1152	0.0813	0.0964
GPT-3.5	**10.9651**	0.7736	0.2633	0.1670	0.2101

Table 15: Factualness evaluation results between our chatbot (ถามทนาย) with other models

Table 15 shows end-to-end evaluation in the factualness aspect. Our chatbot (ถามทนาย) shows superior performance across contextual metrics over other competitors, with the highest BERT score F1 at 0.8070, demonstrating effectiveness at generating responses that are semantically closer to the ground-truth labels which reflects a better understanding of the legal context. In addition, our model ROUGE-1, ROUGE-2, and ROUGE-L scores are also the highest among the competition at 0.3698, 0.3255, and 0.3380 respectively, confirming its ability to retain key information more effectively. However, our chatbot does have the highest latency at 15.7564, highlighting the area that requires improvement.

The significant difference in scores between our model (ถามทนาย) and GPT-3.5 compared to the ทะนอย chatbot may raise concerns regarding potential biases in our evaluation method. It is important to consider that the formatting of responses from ทะนอย includes unique phrases and symbols that are characteristic of its output. These elements could inherently influence the evaluation metrics, particularly BERT score and ROUGE score, which rely on textual similarity.

While efforts were made by us to standardize responses by cleaning extraneous formatting and special symbols to make a fair comparison, not all such elements were completely removed, potentially skewing the results and impacting the scores. This factor is critical to acknowledge as it could lead to artificially inflated or deflated performance metrics compared to other models that use more standardized response formats.

Overall, these results underline the success of our chatbot in delivering factually correct and contextually appropriate responses, setting a new standard for accuracy in the Thai legal domain. The evaluation confirms that our chatbot not only meets but exceeds the performance of its competitors, providing users with reliable and authoritative legal assistance.

4.5.2 User Preference Evaluation

In this aspect of our evaluation, we focus on assessing user satisfaction and preference, which are critical indicators of the chatbot's effectiveness from the perspective of its end-users. This evaluation aims to capture qualitative insights into how users perceive the usability and utility of the chatbot. The evaluation will be conducted in the form of A/B Testing on the modified version of Chatbot Arena[24], where we will have our runtime-optimized finetuned model compete against runtime-optimized base SeaLLMs-7B-v2 given the same knowledge to answer the end-user question. Figure 16 displays the interface of the Chatbot Arena, where the model name will be anonymized to remove bias from testing users. Users will need to compare the answers between two models, where they will have the option to choose whether the model A response is better, same (as good as/as bad as), or worse than model B. The results will be stored and pulled out to complete the evaluation process.

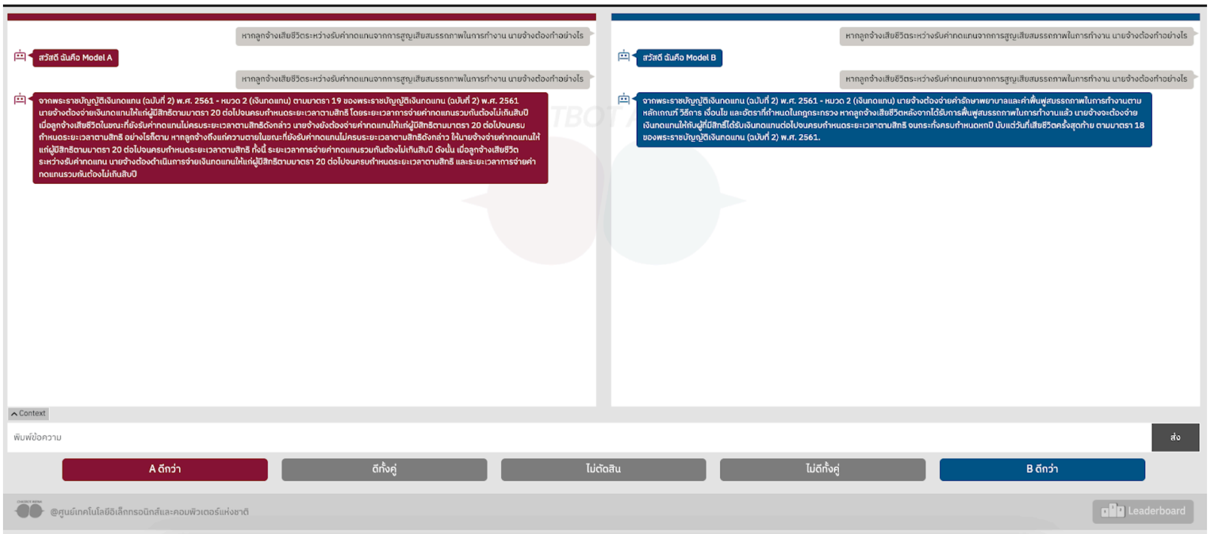

Figure 16: Chatbot Arena interface

In the A/B testing, we also give users an option to use our set of questions instead of thinking on their own. Table 16 below shows a set of questions in the chatbot arena, totaling 10 questions. Users then will then choose whether they want to ask with their own question or with preset questions. Then users also have the option to provide reasons or comments for their choice in the Google Form.

Question	Answer
หากลูกจ้างเสียชีวิตระหว่างรับค่าทดแทนจากการสูญเสียสมรรถภาพในการทำงาน นายจ้างต้องทำอย่างไร	นายจ้างต้องจ่ายค่าทดแทนให้แก่ผู้มีสิทธิตามมาตรา 20 ต่อไปจนครบกำหนดระยะเวลาตามสิทธิ
ในกรณีที่บุคคลหลายคนมีบุริมสิทธิพิเศษเหนือทรัพย์สินเดียวกัน จะใช้หลักเกณฑ์ใดในการกำหนดลำดับการชำระหนี้	จะใช้หลักเกณฑ์ลำดับก่อนหลังตามที่ได้เรียงลำดับไว้ในมาตรา 273 ในการกำหนดลำดับการชำระหนี้สำหรับบุริมสิทธิพิเศษที่เกิดข้อพิพาทเหนือทรัพย์สินเดียวกัน
หากมีการข่มขืนใจผู้อื่นโดยขู่เข็ญว่าจะเปิดเผยความลับ จะถูกดำเนินคดีฐานอะไร	จะถูกดำเนินคดีฐานรีดเอาทรัพย์
ศาลมีอำนาจใดบ้างเกี่ยวกับการรับฟังพยานหลักฐาน	ศาลมีอำนาจปฏิเสธไม่รับพยานหลักฐานที่รับฟังไม่ได้หรือยื่นฝ่าฝืนต่อบทบัญญัติ, มีอำนาจงดการสืบพยานหลักฐานที่ฟุ่มเฟือยหรือไม่เกี่ยวแก่ประเด็น, และมีอำนาจนำพยานหลักฐานอื่นมาสืบเพิ่มเติมเพื่อประโยชน์แห่งความยุติธรรม
คนต่างด้าวที่ไม่ยื่นคำร้องขอรับใบผ่านภาษีอากรและเดินทางออกจากประเทศไทยจะต้องเสียเงินเพิ่มเท่าไร	คนต่างด้าวที่ไม่ยื่นคำร้องขอรับใบผ่านภาษีอากรและเดินทางออกจากประเทศไทยจะต้องเสียเงินเพิ่มร้อยละ 20 ของเงินภาษีอากรที่จะต้องเสียทั้งสิ้น
บุริมสิทธิคืออะไร	บุริมสิทธิคือสิทธิที่ผู้ทรงบุริมสิทธิมีเหนือทรัพย์สินของลูกหนี้ ซึ่งทำให้ผู้นั้นสามารถได้รับการชำระหนี้ที่ค้างชำระก่อนเจ้าหนี้อื่นๆ ตามที่กำหนดไว้ในประมวลกฎหมาย
หากมีบุคคลหลายคนเป็นเจ้าหนี้และหนี้นั้นไม่สามารถแบ่งชำระได้ ลูกหนี้จะต้องชำระหนี้อย่างไร?	ลูกหนี้จะต้องชำระหนี้ให้ได้ประโยชน์แก่เจ้าหนี้ทุกคนด้วยกัน
หากต้องการเวนคืนสิทธิในที่ดินให้แก่รัฐ ต้องดำเนินการอย่างไร	ยื่นคำขอเวนคืนต่อพนักงานเจ้าหน้าที่ตามมาตรา 71
หากลูกจ้างต้องการการฟื้นฟูสมรรถภาพหลังจากประสบอันตรายจากการทำงาน นายจ้างต้องจ่ายค่าใช้จ่ายอย่างไร	นายจ้างต้องจ่ายค่าฟื้นฟูสมรรถภาพในการทำงานของลูกจ้างเท่าที่จ่ายจริงตามความจำเป็น
การทำลายหลักฐานเพื่อช่วยผู้อื่นมิให้ต้องรับโทษมีโทษอย่างไร	ต้องระวางโทษจำคุกไม่เกินห้าปี หรือปรับไม่เกินหนึ่งแสนบาท หรือทั้งจำทั้งปรับ

Table 16: Set of preset questions and answers for the Chatbot Arena

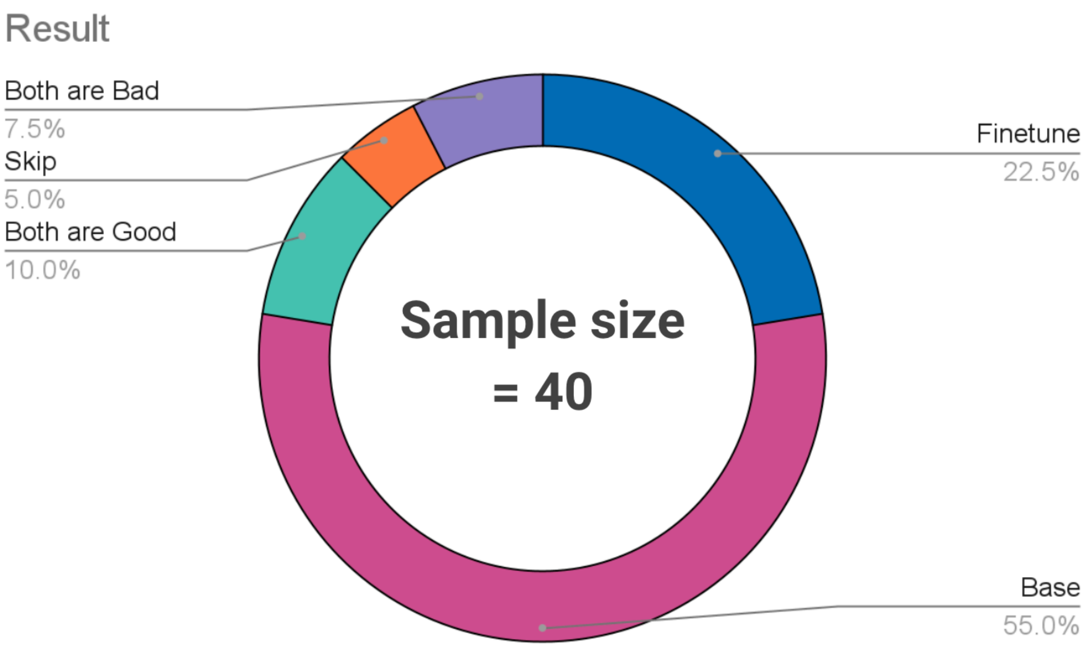

Figure 17: Result from chatbot arena

Figure 17 shows the result from Chatbot Arena with a sample size of 40. Based on the evaluation results in the Chatbot Arena, it is evident that 55% of the total number of queries were directed towards the base model, while 22.5% were allocated to the finetune model. Moreover, 10% indicated that both models performed equally well, whereas 7.5% expressed dissatisfaction with both models. Finally, 5% abstained from making a definitive choice.

Upon investigation, it was found that judges perceived the finetuned model as superior in terms of accuracy and conciseness in its responses. Conversely, the base model was favored for its succinctness and perceived reliability. However, most comments from judges highlighted issues with the finetuned model, such as repeating text and overly lengthy responses, which detracted from its overall effectiveness. These observations suggest that while the finetuned model excels in specific technical aspects, its practical usability is compromised by these response quality issues.

5.Social Impact

The development of a chatbot specifically tailored for the legal domain in Thailand holds significant potential to impact society positively. By harnessing this technology, individuals, particularly those with limited access to legal resources, can obtain timely guidance and information about their legal rights, procedures, and potential actions. This increased accessibility to critical legal information serves to empower citizens, potentially diminishing disparities in legal understanding and promoting more equitable access to justice across various socio-economic groups.

Moreover, such a chatbot can substantially reduce the burden on the legal system. Providing preliminary assistance, allows legal professionals to dedicate more time and resources to handling complex cases, thus increasing the efficiency and effectiveness of legal services. This not only helps streamline the process of legal advice but also ensures that more people can receive help at a preliminary level without necessarily needing to engage a lawyer immediately.

The chatbot also serves as an educational tool, enhancing the public’s understanding of the law. Demystifying legal jargon and processes, helps build a more legally informed community. This can lead to a decrease in legal conflicts and issues as individuals become more aware of their rights and the proper channels through which to address their concerns.

6.Conclusion

Our study has successfully reached the culmination of our project, "Reliable Domain-Specific Chatbot for the Thai Language," where we choose to demonstrate private domain data by implementing a Thai Legal assistant chatbot, available in the LINE application. We have not only completed the development phase but have also meticulously refined and optimized our chatbot's capabilities in many aspects, and we have established a solid foundation for a sophisticated legal assistance tool.

A thorough performance analysis provided us with crucial insights into the information retrieval process and question-answering capabilities, particularly addressing the challenge of answer hallucination to make the chatbot reliable for the end user. The subsequent refinement phase brought forth enhancements across the board, by improving information retrieval capabilities, finetuning the large language models with domain-specific data, optimizing prompts, and optimizing the model inference runtime. These steps have significantly bolstered the accuracy, relevance, and latency of the chatbot's responses.

With the chatbot now live on the LINE platform, we have achieved a significant milestone by providing a valuable tool for those seeking legal assistance. This system sets a new benchmark for domain-specific chatbots, combining technical precision with the flexibility to adapt to user preferences and expectations.

To conclude our project, the end-to-end chatbot evaluation results from both factualness and user preferences aspects demonstrate mixed perceptions regarding the performance of our models. While the finetuned model has shown a commendable improvement in terms of accuracy and conciseness, aligning closely with our goals for contextual correctness, it is evident that we overlooked certain aspects that significantly impact user experience. Notably, issues such as repetitive text and overly lengthy responses have detracted from the overall effectiveness of the finetuned model. These drawbacks have led to a preference for the base model, which, despite its lesser contextual accuracy, was perceived as more succinct and reliable.

Acknowledging these insights, we must address the identified shortcomings. The feedback points us towards the need for a balanced approach in future iterations of the model—enhancing not just the factual accuracy but also refining the delivery of responses to avoid redundancy and verbosity. By doing so, we aim to improve not only the technical capabilities of our chatbot but also its practical usability and reliability, ensuring it meets the nuanced expectations of users seeking legal assistance. Furthermore, future work can also include expanding the chatbot’s legal knowledge base to be able to assist users in more specialized areas of law. Continuous learning mechanisms could also be implemented to allow the chatbot to evolve from user interactions, improving iteratively over time. Moreover, considering the potential increase in user traffic, incorporating load balancing mechanisms becomes crucial to effectively handle a high volume of user requests and ensure optimal performance of the chatbot system.

With these avenues for future exploration, our project concludes with a promising result, and we are looking forward to the advancements that will continue to shape the intersection of AI and legal domain assistance.

7.References

[1] Elastic. (n.d.). Large Language Models. Retrieved November 1, 2023, from https://www.elastic.co/what-is/large-language-models

[2] IBM Research. (n.d.). Retrieval-Augmented Generation (RAG). Retrieved November 1, 2023, from https://research.ibm.com/blog/retrieval-augmented-generation-RAG

[3] Cem Dilmegani. Large Language Model Training. Retrieved November 1, 2023, from https://research.aimultiple.com/large-language-model-training

[4] Sean Michael Kerner. TechTarget. https://www.techtarget.com/whatis/definition/large-language-model-LLM

[5] Deci.ai. (n.d.). Fine-tune LLAMA-2 with LoRA for Question Answering. Retrieved November 1, 2023, from https://deci.ai/blog/fine-tune-llama-2-with-lora-for-question-answering/

[6] Hugging Face. (n.d.). SFT Trainer. Retrieved November 1, 2023, from https://huggingface.co/docs/trl/main/en/sft_trainer

[7] Hayagriva. (n.d.). LoRA and QLoRA: An Efficient Approach to Fine-tuning Large Models - Under the Hood. Medium. Retrieved November 1, 2023, from https://medium.com/@hayagriva99999/lora-and-qlora-an-efficient-approach-to-fine-tuning-large-models-under-the-hood-948468424cd6

[8] GPT4ALL Documentation. (n.d.). Getting the Most of Your Local LLM. Retrieved November 1, 2023, from https://docs.gpt4all.io/#getting-the-most-of-your-local-llm

[9] Albert Ziegler and John Berryman (July 17, 2023), A developer's guide to prompt engineering and LLMs, Retrieved November 2, 2023, from https://github.blog/2023-07-17-prompt-engineering-guide-generative-ai-llms/

[10] Pulkit Sharma (August 8, 2023), Understanding Distance Metrics Used in Machine Learning, Retrieved November 2, 2023, from https://www.analyticsvidhya.com/blog/2020/02/4-types-of-distance-metrics-in-machine-learning/

[11] Anthony Chen, Gabriel Stanovsky, Sameer Singh, et al. (2019). Evaluating Question Answering Evaluation. ACL 2019, Retrieved November 2, 2023, Retrieved from https://aclanthology.org/D19-5817.pdf

[12] Yue Zhang, Yafu Li, Leyang Cui, et al. (2023). A Survey on Hallucination in Large Language Models. ArXiv Preprint ArXiv:2309.01219. Retrieved November 2, 2023, from https://arxiv.org/pdf/2309.01219.pdf

[13] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, Luke Zettlemoyer, et al. (2023). QLoRA: Efficient Finetuning of Quantized LLMs. ArXiv Preprint ArXiv:2305.14314. Retrieved November 2, 2023, from https://arxiv.org/pdf/2305.14314.pdf

[14] Seungjun Lee, Yoonna Jang, Chanjun Park, et al. (2023). PEEP-Talk: A Situational Dialogue-based Chatbot for English Education. ACL 2023. Retrieved November 2, 2023, from https://aclanthology.org/2023.acl-demo.18.pdf

[15] ScienceDirect. Cosine Similarity. Retrieved November 28, 2023, from https://www.sciencedirect.com/topics/computer-science/cosine-similarity

[16] Satanjeev Banerjee and Alon Lavie (2005). METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. Retrieved November 28, 2023, from https://aclanthology.org/W05-0909.pdf

[17] Sthanikam Santhosh. Understanding BLEU and ROUGE score for NLP. Retrieved November 28, 2023, from https://medium.com/@sthanikamsanthosh1994/understanding-bleu-and-rouge-score-for-nlp-evaluation-1ab334ecadcb

[18] Fangyi Yu, Lee Quartey, Frank Schilder. Legal Prompting: Teaching a Language Model to Think Like a Lawyer. arXiv:2212.01326, Retrieved April 25, 2024, from https://arxiv.org/pdf/2212.01326

[19] ONNX AI, Open Neural Network Exchange - The open standard for machine learning interoperability, Retrieved April 25, 2024, from https://onnx.ai/

[20] HuggingFace, Optimum, Retrieved April 25, 2024, from https://huggingface.co/docs/optimum/main/en/index

[21] NVIDIA, A TensorRT Toolbox for Optimized Large Language Model Inference, Retrieved April 25, 2024, from https://nvidia.github.io/TensorRT-LLM/overview.html

[22] iApp, ทนาย AI ระดับจีเนียส - ทะนอย, Retrieved April 25, 2024, from https://iapp.co.th/thanoy

[23] iApp, Chochae Chatbot, Retrieved April 25, 2024, from https://iapp.co.th/chochae-chatboti

[24] Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E. Gonzalez, Ion Stoica, Chatbot Arena: An Open Platform for Evaluating LLMs by Human Preference. arXiv:2403.04132, Retrieved April 25, 2024, from https://arxiv.org/pdf/2403.04132