End-to-End Development of Chatbots Using Large Language Models: A Use Case in Thai Legal Documents

Anonymous ACL submission

Abstract

Large Language Models (LLMs) are increasingly applied in various domains, including chatbots. This paper presents the development of "TamTanai", a domain-specific chatbot for the Thai legal sector, using LLMs. The development process is structured into three main stages: (1) legal dataset construction, (2) LLM refinement, and (3) LLM deployment. Initially, data from Thai legal websites was collected and converted into question-answering pairs. Subsequently, several Thai-supporting LLMs-Typhoon, OpenThaiGPT, and SeaLLM-were evaluated and fine-tuned using the QLoRA method. We utilized the Retrieval-Augmented-Generation (RAG) technique to integrate retrieved legal cases, thereby enhancing response reliability. Additionally, our reranking module, which combines keyword and contextual searches, further refined the retrieval process. To simulate attorney-like interactions, we utilized a Legal Reasoning Prompt (LRP) technique (Yu et al., 2023). Deploying the chatbot using TensorRT-LLM (NVIDIA, 2023) involved significant challenges, particularly in parameter tuning for optimal performance. Comparative experiments demonstrate that "TamTanai" outperforms competitors like Thanoy and GPT-3.5 across several metrics, including BERT-score and ROUGE indices, while maintaining a response time of under ten seconds on a conventional server.

1 Introduction

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 the 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 LLMs is the phenomenon of hallucination, where the models generate responses that, while seemingly coherent and articulate, are 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 accuracy declines in specialized fields with limited training data, such as legal domains. This issue is even more pronounced in the Thai language, where domain-specific nuances add complexity.

In this paper, we aim to overcome these challenges by developing a Thai language chatbot called TamTanai tailored for specialized domains, particularly Thai law. The objective is to create a chatbot that delivers precise responses in these complex areas, advancing the capabilities of chatbots to handle intricate, domain-specific interactions in Thai, and setting a new standard for reliability and accuracy in language-specific chatbot technology. Moreover, we contribute in several key areas. Initially, we developed the chatbot using the Retrieval Augmented Generation (RAG) method, enhancing its ability to provide contextually relevant responses, especially in Thai language and legal contexts. Subsequently, we create a comprehensive Thai law dataset to fine-tune LLM with the QLoRA adapter (Dettmers et al., 2023), ensuring accurate legal responses. Additionally, we refine the chatbot to address hallucination issues by enhancing the LLM, retrieval system, and prompt design. After deployment on the LINE Official Account platform, we evaluate its performance against benchmarks like GPT-3.5, setting new standards for reliability and accuracy in Thai legal chatbots.

The remainder of the paper is organized as follows. Section 2 shows related works. As for our work, the details of dataset construction, model refinement, and model deployment are demonstrated in Sections 3 to 5. Then, the experiments and results are shown in Section 6. Finally, the conclusion is presented in Section 7.

2 Related Works

2.1 Thai-Supported LLMs

Three open-source Thai generative large language models have also recently been developed. First, Typhoon (Pipatanakul et al., 2023), was developed specifically for the Thai language with 7 billion parameters, where its first version is adapted from Mistral-7B (Jiang et al., 2023). Typhoon is evaluated on various benchmarks and shows that Typhoon is the state-of-the-art open-source Thai large language model. Secondly, OpenThaiGPT (openthaigpt-1.0.0-beta-7b) (Viriyayudhakorn et al., 2024) is based on Llama2 (Touvron et al., 2023) with 7 billion parameters. Its tokenizer extends Llama2s tokenizer to include 24,554 additional Thai tokens to improve generation efficiency. OpenThaiGPT continues pre-training Llama-2 on Thai data and performs instruction fine-tuning on the translated instruction datasets. Finally, SeaLLMs (Nguyen et al., 2023), an innovative series of language models that specifically focuses on Southeast Asian (SEA) languages. SeaLLMs are built upon the Llama-2 model with 7 billion parameters and further advanced through continued pre-training with an extended vocabulary, specialized instruction, and alignment tuning to better capture the intricacies of regional languages.

2.2 Prior Domain-Specific Chatbots

In the realm of domain-specific chatbots, particularly within the legal sector, several notable examples have demonstrated the efficacy of tailored conversational agents in addressing specific user needs and requirements. These chatbots are often designed to provide legal assistance, information retrieval, or guidance on legal matters, leveraging domain knowledge and expertise to deliver accurate and relevant responses.

One prominent example is the “LawBot” chatbot (Giri, 2024), designed to provide legal information and guidance specifically for students and young people. Developed by a team of researchers at the University of Cambridge, LawBot offers personalized advice on various legal topics, ranging from consumer rights to employment law, through an intuitive chat interface. By tailoring its responses to the needs and concerns of its target audience, LawBot aims to bridge the gap between legal knowledge and everyday life, empowering users to make informed decisions and assert their rights effectively.

Another noteworthy example is the “Thanoy” LINE chatbot (iApp, 2024a), developed by iApp, which utilizes ChoeChaeGPT (iApp, 2024b) as its underlying LLM. However, the other underlying details of “Thanoy” are not available to the public yet.

3 Legal Dataset Construction

In this research, we focus on the Thai legal domain, aiming to develop a domain-specific chatbot. The primary data source for this endeavor is a comprehensive set of official legal documents obtained from a dedicated Thai legal documentation website[1]. To ensure the relevance and importance of the data, we selected 31 key documents frequently referenced in legal consultations. This selection process was rigorously guided by consultations with legal experts, who helped identify the critical areas most pertinent to the needs of the Thai legal system. Such a focused approach ensures

[1] https://www.ocs.go.th/

that the dataset covers the essential elements required for effective legal assistance.

Once the key legal documents were collected, the data preparation phase involved several crucial steps to convert the raw data into a format suitable for training a large language model (LLM). Initially, each document underwent a detailed cleaning process. This included the removal of irrelevant sections and standardization of document formatting to eliminate textual inconsistencies. Subsequently, we segmented the documents into smaller units based on their specific subtopics. Each unit represents a single, coherent thought or legal principle, facilitating more targeted training and application.

To further enhance the dataset, we utilized the state-of-the-art LLM, GPT-4, to generate additional training and testing data. This synthetic dataset was then subjected to a rigorous human verification process to ensure its accuracy and relevance. By cleaning, reformatting, and enriching the data, we prepare a robust dataset that supports both the training needs of the LLM and its effective deployment in real-world legal consultations. A total of 4,121 examples have been created in this process.

4 Model Refinement

4.1 Model Fine-Tuning

Our study focuses on fine-tuning Language Model (LLM) adapters using a combination of our specialized question-answering dataset and proprietary legal domain data from the Thai legal system. This crucial step aims to augment the LLM's proficiency in specific areas of Thai law, thereby amplifying its question-answering accuracy and relevance within this domain.

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.

LLM	**w/o context↑**	**w/ context↑**
Typhoon-7B	0.7147	**0.7624**
OPT-7B	0.6835	0.7210
SeaLLMs-7B-v2	**0.7158**	0.7536

Table 1: Performance comparison in terms of BERT F1 among different LLMs: Typhoon, OpenThai GPT (OPT), and SeaLLMs. There are two columns refer to w/o and w/ legal context embedded in a prompt. Boldface refers to the winner.

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 1 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 tended 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, we aimed to encourage Typhoon-7B to provide shorter, more practical responses without repetitive text. However, neither approach proved effective.

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.

These hyperparameters were carefully selected to strike a balance between model adaptability, computational efficiency, and prevention of overfitting, ultimately contributing to the effectiveness of our fine-tuning process. With these configurations, we fine-tuned the adapter for 50 epochs. 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.

4.2 Retrieval-Augmented-Generation (RAG)

The development process will focus on constructing the core functionalities of the chatbot using the Retrieval Augmented Generation (RAG) method (Gao et al., 2024). This approach combines retrieval-based and generation-based techniques to create a hybrid system as illustrate in figure 1, significantly enhancing the chatbot's ability to provide improved and contextually relevant responses. By leveraging RAG, we ensure that the chatbot can accurately handle the complexities of the Thai language and legal terminology. This method allows the system to draw from a rich database of pre-existing knowledge while generating responses dynamically, leading to more precise and contextually appropriate interactions.

Figure 1 illustrates the architecture of our information retrieval system. The system utilizes a combination of three techniques: contextual search, keyword search, and reranking. These techniques are chosen to effectively address the specific requirements of legal queries. The contextual search and keyword search are executed parallelly.

- Contextual Search: The retriever will search for documents in the database that are similar to a given query by comparing the embedding vector (generated by multilingual-e5-small model (Wang et al., 2024)) of the question and the source documents using cosine similarity.

- Keyword Search: In the legal domain, certain questions include specific section numbers, making keyword searches particularly effective for accurately identifying relevant documents in these instances. On the other hand, when questions do not include section numbers, contextual search is employed to understand and retrieve documents based on the overall context of the query.

- Reranking: After retrieving documents using both keyword and contextual search methods, we combine the results and use a reranking technique to prioritize the most important documents (Liu et al., 2023) for prompting to the LLM. The Cohere/rerank-multilingual-v2.0 model has been used to rerank the retrieved documents. This process ensures that the 3 most relevant documents are selected, thereby enhancing the chatbot's accuracy and contextual relevance in responding to legal inquiries.

4.3 Prompt Engineering

In this step, we aimed to tweak and optimize a prompt to ensure the most coherent and

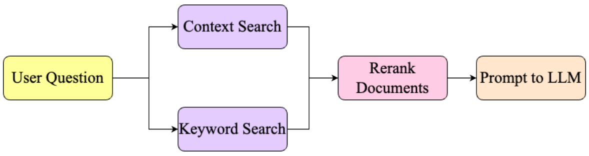

Figure 1: Architecture of the information retriever

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 (Yu et al., 2022). 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.

5 Model Deployment

5.1 Model 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.

In the complete optimization process, we initially merge the weights of our fine-tuned QLoRA on the Thai legal dataset with the original base SeaLLMs-7B-v2 to form a single model. Subsequently, we perform int8 weight quantization (originally float16) on the merged weights to reduce memory requirements on the computing units. This process effectively reduces the model size from 14 GB to 7 GB in memory. Although this reduction in model size results in a slight loss in precision, the trade-off is considered acceptable due to the significant decrease in memory usage and the resulting 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. 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.

5.2 System Deployment

Our chatbot will be deployed within the LINE application using the Line Bot SDK for Python, integrated with the LINE Messaging API, and connected via a Webhook implemented with FastAPI. This setup ensures efficient interaction and integration with various services, validating the system in a real-world environment and collecting data for continuous improvement.

6 Experimentation and Results

6.1 Comparison with Existing Solutions

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, Thanoy (iApp, 2024a). 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 TamTanai, as GPT-3.5 may not have access to the latest Thai legal information.

Chatbot model	BERT F1↑	ROUGE-1↑	ROUGE-2↑	ROUGE-L↑
TamTanai (Ours)	**0.8070**	**0.3698**	**0.3255**	**0.3380**
Thanoy	0.7666	0.1152	0.0813	0.0964
GPT-3.5	0.7736	0.2633	0.1670	0.2101

Table 2: Factualness evaluation results between our chatbot (TamTanai) with other models. The latency cannot be compared among chatbots since they deployed on different servers. Boldface refers to the winner.

Table 2 shows end-to-end evaluation in the factualness aspect. Our chatbot (TamTanai) 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.

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.

6.2 Results of Fine-Tuning

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 containing 413 rows.

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 3 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.

In addition to the time concern, Figure 2 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.

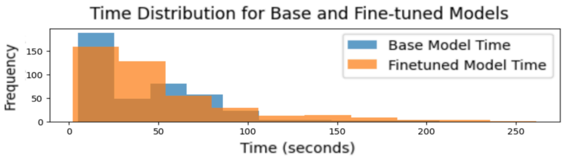

Figure 2: Time distribution for base and fine-tuned models

6.3 Results of Retriever

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.

Fine-tuned model	Latency (seconds)↓	BERT F1↑	ROUGE-1↑	ROUGE-2↑	ROUGE-L↑
w/o QLoRa	**42.46**	0.7964	0.3992	0.3273	0.3549
w/ QLoRa	49.01	**0.8197**	**0.4046**	**0.3526**	**0.3660**

Table 3: Effect of model fine-tuning using QLoRA on SeaLLMs-7B-v2. Boldface refers to the winner.

IR model	Latency (seconds)↓	Accuracy↑	Precision↑	Recall↑	F1↑	BERT F1↑
Baseline	**0.01**	0.6877	0.2300	0.6884	0.3448	0.6706
Ours	2.95	**0.8596**	**0.6840**	**0.8596**	**0.7618**	**0.7476**

Table 4: Result of comparison between baseline and improved version (ours) of information retriever mechanism. Boldface refers to the winner.

Prompt	Correct answer↑	BERT F1↑	A vs B test↑
Prototype	40	0.8042	7
Optimized	**45**	**0.8071**	**43**

Table 5: Result comparison between original prompt and optimized prompt.The "correct answer" and "A vs B test" columns are based on total of 50 questions. Boldface refers to the winner.

Table 4 shows the comparison of the original version that utilized only contextual search (L2 distance) to the improved version that utilized a hybrid search mechanism. Despite increasing latency, every other aspect of the retriever has been significantly improved. The new retrieval mechanism enhances the system's accuracy in finding relevant documents and reducing noise. The higher BERTScore indicates that better quality documents lead to more correct LLM answers. Our hybrid approach, combining keyword extraction, contextual search, and reranking algorithms, successfully narrows the search space and prioritizes relevant documents, improving the chatbot's precision and authority in responses.

6.4 Results of Prompt Engineering

We evaluated our prompt engineering changes by sampling 50 random questions from the test dataset, assessing their effectiveness through improvements in accuracy, relevance, and contextual appropriateness of the generated responses. Table 5 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.

6.5 Results of Model Optimization

After optimizing our model, we conducted a detailed evaluation to ensure the modifications did not compromise its core functionality. The primary goal was to compare the performance of the optimized model against the original in terms of latency and contextual correctness. Table 6 shows a comparison between the original model, using the Transformer library, and the optimized model, utilizing the NVIDIA TensorRT-LLM engine. Although the optimized model exhibited slightly lower contextual scores (BERT and ROUGE) due to 8-bit weight quantization and other techniques, the inference runtime improvement is substantial, reducing the average response time from 49 seconds to approximately 3 seconds. This trade-off, resulting in a slight decrease in contextual accuracy, is justified by the vastly improved responsiveness, crucial for real-world applications where immediate interaction is prioritized. Therefore, our runtime-optimized model is ready for deployment.

7 Conclusion

In this paper, we successfully implemented a Thai legal assistant chatbot called "Tam-

Model	Latency (seconds) ↓	BERT F1↑	ROUGE-1↑	ROUGE-2↑	ROUGE-L↑
Original	49.01	**0.8197**	**0.4046**	**0.3526**	**0.3660**
Runtime-optimized	**3.03**	0.8070	0.3698	0.3255	0.3380

Table 6: Results comparison between original fine-tuned model and runtime-optimized model with NVIDIA TensorRT-LLM. Boldface refers to the winner.

Tanai" via the LINE application. We refined and optimized TamTanai's capabilities by focusing on performance analysis, information retrieval, and addressing answer hallucination to enhance response reliability. Notable improvements include fine-tuning language models with domain-specific data and optimizing prompt and inference runtime. To achieve this, we constructed a comprehensive Thai legal dataset. Our experiments demonstrated that TamTanai outperformed other chatbots, including GPT-3.5 and Thanoy (another Thai legal chatbot), across all measures (BERTScore F1 and ROUGE).

8 Acknowledgement

We thank Learning Innovation Center (LIC) for research funding. We thank LANTA HPC and TARA HPC for computing resources. We thank MR. Matichon Maneegard for helpful suggestions.

References

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *Preprint*, arXiv:2305.14314.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. 2024. Retrieval-augmented generation for large language models: A survey. *Preprint*, arXiv:2312.10997.

Mandaar Mukesh Giri. 2024. Lawbotpro.

iApp. 2024a. ทนาย ai ระดับจีเนียส - ทะนอย.

iApp. 2024b. Chochae chatbot.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2023. Lost in the middle: How language models use long contexts. *Preprint*, arXiv:2307.03172.

Xuan-Phi Nguyen, Wenxuan Zhang, Xin Li, Mahani Aljunied, Qingyu Tan, Liying Cheng, Guanzheng Chen, Yue Deng, Sen Yang, Chaoqun Liu, Hang Zhang, and Lidong Bing. 2023. Seallms – large language models for southeast asia. *Preprint*, arXiv:2312.00738.

NVIDIA. 2023. Tensorrt-llm.

Kunat Pipatanakul, Phatrasek Jirabovonvisut, Potsawee Manakul, Sittipong Sripaisarnmongkol, Ruangsak Patomwong, Pathomporn Chokchainant, and Kasima Tharnpipitchai. 2023. Typhoon: Thai large language models. *Preprint*, arXiv:2312.13951.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models.

Kobkrit Viriyayudhakorn, Sumeth Yuenyong, Thaweewat Rugsujarit, Jillaphat Jaroenkantasima, Norapat Buppodom, Koravich Sangkaew, Peerawat Rojratchadakorn, Surapon Nonesung, Chanon Utupon, Sadhis Wongprayoon, Nucharee Thongthungwong, Chawakorn Phiantham, Patteera Triamamornwooth, Nattarika Juntarapaoraya, Kriangkrai Saetan, and Pitikorn Khlaisamniang. 2024. Openthaigpt 1.0.0.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Multilingual e5 text embeddings: A technical report. *Preprint*, arXiv:2402.05672.

Fangyi Yu, Lee Quartey, and Frank Schilder. 2022. Legal prompting: Teaching a language model to think like a lawyer. *Preprint*, arXiv:2212.01326.

Fangyi Yu, Lee Quartey, and Frank Schilder. 2023. Exploring the effectiveness of prompt engineering for legal reasoning tasks. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13582–13596, Toronto, Canada. Association for Computational Linguistics.