XOSA-NOWJ@ALQAC 2024: Efficient Legal Information Processing with the Combination of Semantic–based model and LLMs

[1] Hai-Long Nguyen [*], [3] Duc-Minh Nguyen, [1] Quang-Anh Nguyen, [1] Huy-Quang Chu,
[1] Huu-Dung Nguyen, [1] Khanh-Huyen Nguyen, [1] Cong-Minh Pham,
[1] Thi-Hai-Yen Vuong, [2] Ha-Thanh Nguyen
[1] *VNU University of Engineering and Technology, Hanoi, Vietnam*
[2] *National Institute of Informatics, Tokyo, Japan*
[3] *RMIT University Vietnam*
[*] long.nh@vnu.edu.vn

***Abstract*—In the context of the current development of Large Language Models (LLMs), methods for addressing natural legal language processing problems start to incorporate the knowledge from LLMs. In this study, to tackle the legal documents retrieval task, ranking models based on semantic and lexical features combined with the output of LLMs were implemented and evaluated. For the legal question-answering task, methods for collecting and optimizing prompts on LLMs were employed to create the best prompt pattern for extracting the knowledge and reasoning capability of LLMs. The experimental models demonstrated significant results on the private test set, achieving top-2 for task 1 and top-3 for task 2 in the ALQAC 2024 competition.**

***Index Terms*—Legal AI, Information Retrieval, Question Answering, Large Language Model**

I. Introduction

The Vietnamese legal system is a complex system with numerous layers of documents, from the highest level, the Constitution, to the laws, followed by numerous accompanying documents to guide implementation, amendments, or to consolidate amended documents. This complexity severely limits public access to the law, which also explains why Vietnam is one of the countries with a generally low level of legal awareness. Therefore, solving problems related to Legal Document Retrieval – helping to search for legal documents related to a practical issue, or Legal Question Answering – helping to answer and explain legal questions, is very important. This work helps everyone to easily access legal knowledge, enhances the public's legal awareness, and thus will minimize unnecessary inadvertent legal violations.

ALQAC is an annual competition focused on Legal Information Processing, encompassing two main tasks: Legal Document Retrieval and Legal Question Answering. The experimental dataset includes Vietnamese legal queries and several codes of law within the Vietnamese legal system. Given that Vietnamese is classified as a low-resource language, participating teams have contributed solutions to address this issue through techniques such as data augmentation [1], pre–training language models with legal data [2] or utilizing the translation process and prompting LLMs [3]. The methods presented by the teams each year collectively contribute to improving the overall results of both tasks.

In addressing this year's tasks, this research proposes the following contributions:

- To tackle the retrieval task, we proposed combining lexical–based and semantic–based ranking models. For the semantic–based ranking model, we employed the Cross-Encoder to ensure retrieval accuracy. Additionally, to facilitate fast retrieval models, we used a Bi-Encoder model. Both models achieved relatively high F2 scores, and the ensemble of these two models secured a top-2 result on the private test dataset.
- To address questions and answer tasks, we utilized various automatic prompt optimization techniques to synthesize the optimal prompt to maximize answer accuracy. Despite the simplicity of implementation and fast inference time, the method yields notably better results.

II. Related Works

A. Legal Document Retrieval

Recently, there have been significant advancements in the methods proposed to address the Legal Document Retrieval problem. In ALQAC 2022, the first ranking team [2] had fine–tuned the RoBERTa [4] model with legal data, then employed the appropriate top–k for choosing the negative training sample. Also, the long documents were divided into smaller chunks and processed separately. In ALQAC 2023, the NeCo team [5] employed the multilingual–BERT [6] and fine–tuned it with the legal data from Zalo and crawled from websites such as vbpl.vn, lawnet.vn. Then, they fined-tuned it with the sequence–classification downstream task. In addition to fine-tuning BERT-based models, some recent studies have proposed new model architectures or utilized language models with a larger number of parameters. Nguyen et al. [7] used the LLMs-based ranking models including Tohoku BERT and mono–T5 [8]. Then, they employed a data filter process to

create a new training dataset for the next fine–tuning phases. The solution of that team had achieved the highest F2 score in the COLIEE dataset. For addressing the long documents problem and capturing the full–context of both query and legal articles, the authors of [9] proposed an attentive model based on the sentence embedding model with the CNNs or attention layers on top to combine the context of all sentences.

B. Legal Question Answering

Legal Question Answering problem is a challenging task with various types of questions including yes/no questions, multiple choice questions or factoid questions. Many approaches designed based on the language model have been proposed. In ALQAC 2022, Miko team [2] had tackled the answer extraction problem by employing pre–trained RoBERTa [4] and fine-tuning the model for finding the start and end position. Meanwhile, Nguyen et al. in [10] considered the factoid question as the sentence selection which chose the sentence that contained the gold phrase. The start and end position of the exact phrase is then selected by fine-tuning the the XLM-RoBERTa [11] model. Methods leveraging the power of LLMs to address the Legal Question Answering problem have also begun to emerge. For the ALQAC 2023 dataset, the AIEPU [3] team had combined the fine–tuned RoBERTa model with LLMs. In LLMs–based approach, the data is first translated into English, and then the prompt collection, LLMs execution and label extraction are performed to collect the output of LLMs. This method achieved the first ranking in task 2.

III. Legal Document Retrieval method

For tackling the legal document retrieval task this year, our team has developed several ranking models that based on the semantic and lexical features. Furthermore, the LLMs are also employed for utilizing the general knowledge and basic logical reasoning capability of these models.

A. Lexical–based Ranking model

The lexical–based ranking which is based on the lexical similarity between the query and the candidate articles is a lightweight, efficient and time–saving approach. This ranking method not only helps to rank the candidate documents but also provides good negative samples for the training phase of semantic models, which helps these models distinguish semantically relevant documents among those that are lexically similar to the query. Because of these reasons, this work employed the lexical–based ranking model as the basis for more complex deep-learning-based ranking models. Among many lexical-based ranking models, the Okapi BM25 [12] is a well–known model and is widely used in many retrieval systems. Therefore, this model is chosen to be the lexical-ranking phase.

B. Semantic–based Multi–task model

According to the idea and experiments performed in [13], there is a supportive connection between the "relevancy" and the "affirmation" properties. Utilizing these properties in a multi–task model could help to raise the F2 score of the retrieval task. Therefore, this investigation employs this multi–task model by using the gold–label of legal question answering task. The multi–task model [13] only works with the yes/no question type, so a question transferring phase has been performed to convert the multiple choice and essay questions into yes/no question. For the multiple choice questions, regex patterns were used to replace the question words by the content of each options, then the new generated query would become the yes/no question with "yes" label for the combinations with the correct options and "no" label for the combinations with the wrong options. The special options which mentioned that all other options are correct/wrong, or some options are correct/wrong, are also considered. For the essay question, the question words are replaced with the gold–label answer for making the new query with QA label being "yes".

After transforming the dataset to consist entirely yes/no question, the multi–task model is fine–tuned using this new generated dataset and with the sequence classification downstream task. The negative samples is selected from top–k most relevant articles based on the BM25 model's score.

C. Semantic–based Ranking model using Contrastive Learning technique

1) Contrastive S-BERT: Sentence-BERT (S-BERT) [14] refines the BERT architecture by using siamese or triplet network structures, enhancing its performance in sentence embedding tasks, especially in pairwise comparisons utilizing distance metrics, e.g cosine similarity. This approach enables S-BERT to process texts independently, improving its ability to discern between positive and negative examples more effectively than traditional BERT. This improved discrimination is particularly advantageous in applications requiring detailed textual comparisons, such as legal document retrieval.

For our task, we utilized the 'keepitreal/vietnamese-sbert' model[1], which has been fine-tuned entirely on Vietnamese datasets. This model demonstrates high performance in representing Vietnamese texts, making it suitable for our ALQAC tasks, which often deal exclusively with Vietnamese language data.

The dataset needs to be organized into tuples of (anchor, positive, negative) documents, where the anchor denotes the query sentence, the positive is its confirmed relevant article, and the negatives are selected from the top 30 highest lexical scores obtained during the BM25 phase. This arrangement enables the 'vietnamese-sbert' model, which will later be trained employing triplet contrastive loss, to represent the input data with enhanced precision.

2) ColBERT: Although the above contrastive ranking model utilizing S-BERT is appropriate for the Vietnamese dataset, the majority of articles are approximately 300 tokens long, exceeding the 256-token limit of the chosen pre-trained model. Another factor negatively affecting the retrieval result

[1]https://huggingface.co/keepitreal/vietnamese-sbert

of contrastive learning model is that the pre-trained S-BERT model fine-tuned based on RoBERTa [4] follows the same Masked Language Modeling (MLM) approach as BERT [15] which is not effective for retrieval. Despite further fine-tuning the model in our dataset and employing contrastive learning to enhance retrieval capability, the results were not satisfactory.

Upon identifying that the pre-trained model was inadequate, we adopted ColBERTv2 [16] based on ColBERT [17], a model employs several advanced techniques to achieve state-of-the-art performance in retrieval and search tasks. Central to its design is the late interaction model, where queries and documents are encoded independently, and their interaction occurs post-encoding. This approach optimizes efficiency and scalability by enabling the pre-computation of document embedding, which significantly speeds up the retrieval process. Moreover, the pre-trained model supports a maximum input length of 512 tokens, effectively addressing the length limitation issue. These techniques collectively ensure ColBERTv2's retrieval speed and efficiency, as pre-computed document embedding allows for rapid query comparisons, and only the query needs to be encoded at inference, reducing computational demands.

D. Re–ranking based on LLMs prompting

In the previous phases, we used the BM25 algorithm and encoder models to retrieve legal articles based on lexical and semantic similarity. However, the questions are not always clear and may require logical reasoning to retrieve the right legal articles related to the questions. Large Language Models (LLMs) are deep learning models with up to billions of parameters that can comprehend and generate human language text, among other tasks. Besides, LLMs are trained on large sets of data which gives them a broad domain knowledge, therefore they can be applied to various fields. Notably, LLMs can also learn to handle downstream tasks through instruction inputs. Hence, applying Prompting techniques can leverage basic-level reasoning capabilities of LLMs.

The input content of our prompting method includes the query content and the content and identifiers (e.g. LVC-6, BLDS-400, ...) of all candidate articles obtained from the previous phase, unfortunately, we have only tested this method with the multi-task BERT model + BM25. The output given by LLM includes relevance scores for candidate articles ranging from 0 to 100. The output format specified in the prompting sentence will adhere to JSON format.

Ensemble. During the ensemble phase, we begin by normalizing the output score of each phase using a min-max scaler. Next, a grid search is conducted to find the best weights for the parameters and $threshold$. The final score is calculated as

$$score = \alpha * w_{bert+bm25} + (1 - \alpha) * w_{llm} \quad (1)$$

Where $w_{bert+bm25}$ represents the correlation score from the previous phase, and w_{llm} represents the correlation score obtained from the prompting results. Similar to the previous phase, $threshold$ is used to determine which documents are ultimately selected for the retrieval task.

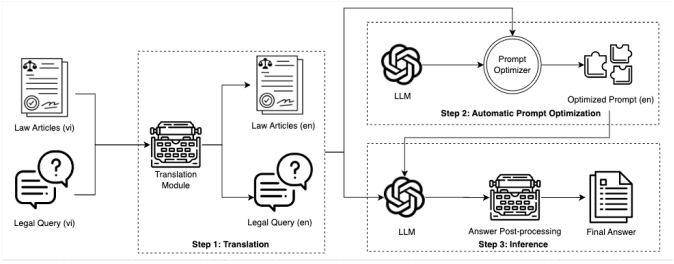

Figure 1. Steps in Question and Answering pipeline

IV. Legal Question Answering Methodology

With continuous development and improved reasoning capabilities, LLMs are expected to bring breakthroughs in solving the textual entailment problem. Despite their strong reasoning ability, models still occasionally produce misleading information or undesired format answers. In an effort to address this problem, applying prompts to the model has been conducted leading to positive changes. Some prompting techniques such as chain of thoughts or few-short learning [18] have brought higher accuracy in answers. Along with that, the optimization of available prompts [19] also paves the way for the increase in accuracy of the model's response. Experiments on utilizing DSPy [20] based on the proposal of MIPRO [21] have resulted in the improvement of reply quality and can be widely applied in tackle obstacles by integrating into LLMs. To combine all of these techniques, a pipeline is proposed at figure 1 for tackling the legal question answering task. The pipeline consists of three main steps. In this pipeline, the focus is on optimizing the prompt in English and use it to answer the questions.

The traditional dataset is translated from Vietnamese to English in step 1. In step 2, this translated data is then used with a LLM to generate an optimized prompt for the Q&A task. Finally, in step 3 the optimized prompt from the previous stage is fed to the LLM to get the answer, followed up by post-processing the answer to get the final answer in the right format.

A. Dataset Translation

Despite the fact that LLMs have developed to accommodate multilingualism, they still encounter certain limitations when processing Vietnamese contexts compared to English. LLMs possess an excellent ability to recognize and understand English situations Therefore, in order to efficiently leverage the superiority of LLMs in the task of reasoning and answering legal problems, we translated the texts from the original language to English for better inference and response.

In particular, the Google Translator package in a Python library called deep-translator is utilized to convert numerous raw contents from Vietnamese into English automatically. Data from the organizers including train set, test set and law set are thoroughly interpreted in English for the next steps of reacting and judging. An example can be seen from table I. In the process of implementation, some contexts in the law dataset exceeded the number of characters that a translator function

Table I
EXAMPLE OF ORIGINAL AND TRANSLATED TEXT

Vietnamese text
Người nghiện ma túy từ đủ 18 tuổi trở lên bị áp dụng biện pháp xử lý hành chính đưa vào cơ sở cai nghiện bắt buộc theo quy định của Luật Xử lý vi phạm hành chính khi bị phát hiện sử dụng chất ma túy một cách trái phép trong thời gian cai nghiện ma túy tự nguyện, đúng hay sai?
Translated English text
Drug addicts aged 18 years or older are subject to administrative measures of being sent to compulsory detoxification facilities according to the provisions of the Law on Handling of Administrative Violations when discovered to be using narcotics illegally. allowed during voluntary drug rehabilitation, right or wrong?

can handle. To serve this problem, we split them into smaller chunks, translated and then recombined.

B. Automatic Prompt Optimization

The pipeline uses MIPRO [21] to synthesize prompts and synthetic few-shot examples based on labelled samples in the training set, the program was implemented in DSPy [20], the best prompts are chosen based on results on evaluation set.

1) True/False questions : We built a pipeline providing the model with a question and relevant articles as hints and instructed it to answer either True or False, based on the hint given. The pipeline's prompt is then optimized using the sample questions in the training set to generate new few-shot samples and better Chain-of-Thought prompts.

2) Multiple choice questions: For Multiple Choice Questions (MCQs), we feed the model question, possible options and relevant articles and instruct it to return strictly to one of the options A, B, C or D.

3) Factoid questions: Because the answer has to be in Vietnamese, we divide the pipeline for factoid questions into two stages. The first stage is to use LLM to generate the factoid answer in English. The second stage is to prompt LLM to translate the answer from the first stage.

C. Answer Post-processing

Raw results from LLM usually contain noise, such as "A.", "B. Content of the choice", "The answer is False.", ... To address this, we implemented a few post-processing logic:

- Remove any unnecessary characters such as "." at the end.
- Try to extract the answer ("True"/"False" or "A", "B", "C", "D") at the start or near the end.
- We did not do any post-processing for factoid answers.

Furthermore, ensembling with majority voting is also a great way to reduce noise in the answers due to the indeterministic nature of LLMs.

V. EXPERIMENTS AND RESULTS

A. ALQAC 2024 dataset Analysis

The ALQAC 2024 dataset consists of two main parts: the "law" set containing 2249 legal articles from 17 different laws. The "full-train" set created by combining the "train" and "unverified-train" datasets. The "full-train" set includes 213 multiple-choice questions, 59 essay questions, and 258 yes/no questions.

To facilitate processing, we combined the article-id and law-id of each legal article in the "law" and each question in the "full-train" into a common format. Finally, we use 80% of the data to train the deep learning model, 10% of the data as the development set (dev) to find the optimal weight combination of models, and the remaining 10% of the data for testing (test).

B. Experimental Setup for Legal Document Retrieval task

1) Dataset Constructed: We constructed the input dataset by reformatting the ALQAC2024 dataset into tuples consisting of (anchor, positive, negative) documents. In this format, the anchor represents the query sentence, the positive corresponds to the gold label of that query, and the negative samples are selected from the top 30 highest lexical scores generated by the BM25 phase.

2) Multi–task BERT: Because the target data is in Vietnamese, the multi–task BERT used the pre–trained multilingual-bert-based-cased [2] as the backbone. Because of the moderate number of training sample in contrast with the model's size, the batch size of 1 is utilized. The fine–tuning process is performed in 5 epochs and with the learning rate of $1e-4$.

3) Contrastive S-BERT: We utilized the sentence-transformers library developed by UKPLAB[3], which provides a robust framework for training and fine-tuning Sentence-BERT models. To efficiently manage and process our Vietnamese dataset, we integrated the datasets library[4] from Hugging Face.

The datasets library allowed us to load and process the Vietnamese legal corpus in a format suitable for contrastive learning. Our constructed dataset can be loaded directly by the library.

The main idea behind contrastive loss function is to help the model to learn effective embeddings of input sequences, by minimizing the distance metrics between anchor and positive documents, while trying to push negative ones away. Specifically, the loss function is defined as follows:

$$\begin{aligned} Loss = \max(&d(anchor, positive) \\ &- d(anchor, negative) + margin, 0) \end{aligned} \tag{2}$$

where d represents the distance metrics, which is cosine similarity in this task. The training process was configured to run for 10 epochs, the initial learning rate was set to $2e-5$, batch size was 16, and the *margin* hyperparameter of the contrastive loss was set to 1.0 by default

4) ColBERT: The ColBERTv2.0 model was utilized for our experiments, facilitated by the use of the RAGatouille[5] library, which offers tools for fine-tuning and training retrieval-augmented generation (RAG) models, ideal for this retrieval

[2]https://huggingface.co/google-bert/bert-base-multilingual-cased
[3]https://sbert.net/
[4]https://huggingface.co/docs/datasets/index
[5]https://github.com/bclavie/RAGatouille

Table II
TASK 2 EVALUATION RESULTS.

	Original Prompt	Bootstrap Fewshot	MIPRO
True/False	70	85	80.0
Multiple choices	93.8	93.8	87.5
Factoid	50	100	100

task. Due to the lack of pre-trained Vietnamese datasets, translated data from the Data Translation phase was used.

For the experiments, the pre-trained model 'colbert-ir/colbertv2.0'[6] was employed, which supports a maximum token length of 512, fitting well with our data requirements. The dataset included one positive sample and thirty negative samples per instance. Consequently, num-new-negatives was set to 0 and mine-hard-negatives was enabled to prevent the model from generating additional negative samples, thus leveraging the existing dataset structure.

Regarding training parameters, the model's default settings were largely adhered to, with specific adjustments to batch size, nbits, learning rate, and maximum token length. A batch size of 4 was chosen to strike a balance between memory usage and training stability, given the moderate size of our dataset. Using 16 nbits, memory usage was reduced while preserving sufficient precision for the embedding. The learning rate was set to 1e-5, enabling more stable and finer updates, essential for minimizing the loss function.

C. Experimental setup for Legal Question Answering task

We used data from the training set of 100 Q&A questions to train for Task 2 and we divided the dataset into 3 categories of questions: True/False, MCQs and Factoid. For each of the question types, we reserve 40% of the questions for evaluation and use 60% for prompt optimization. The test ratio is high due to the lack of labels so larger test ratio ensures it generalizes well to private test set.

The prompts for each task were synthesized using various different algorithms namely Bootstrap Fewshot, COPRO [20] and MIPRO [21]. Table II shows evaluation results for the reserved evaluation set.

It could be observed that by simply optimizing the prompts and selecting the right few-shot examples, the model can achieve notably better results.

D. Result and Analysis

1) Legal documents retrieval: All proposed semantic ranking models are fine–tuned and tested separately. The results of each model on the public test are shown in the Table III. The model the

The results of each run on the private test are shown in the Table IV.

2) Legal question answering: For Task 2, we submitted 3 runs with strategies as follows:

- Run 1 (Ensemble gpt-3.5-turbo): an ensemble of 3 best prompts using gpt-3.5-turbo as LLM to answer.

[6]https://huggingface.co/colbert-ir/colbertv2.0

Table III
TASK 1'S RESULTS ON PUBLIC TEST SET

Run ID	Precision	Recall	F2-macro
Multi-task BERT	0.8589	0.8822	0.8710
Contrastive BERT	0.7238	0.8341	0.7888
COLBERT	0.7641	0.8389	0.8060
Multi-task + Contrastive	0.8453	0.9182	0.8843
Multi-task + ColBERT	0.8370	0.8677	0.8496
Multi-task + LLMs	0.9302	0.9134	**0.9147**

Table IV
TASK 1'S FINAL RESULTS ON PRIVATE TEST SET

Team	Precision	Recall	F2-macro
NOWJ	0.8850	0.8800	0.8771
XOSA-NOWJ (Multi-task + ColBert)	**0.7504**	**0.8367**	**0.7895**
UIT-DarkCow	0.7900	0.7850	0.7856
shikanokonokokoshitantan	0.7210	0.8217	0.7789
se7enese	0.7300	0.87117	0.7132
XOSA-NOWJ (Multi-task + Contrastive)	**0.6181**	**0.8349**	**0.7104**
XOSA-NOWJ (Multi-task + LLMs)	**0.6799**	**0.6566**	**0.6588**
DSCS@ALQAC	0.2567	0.7350	0.5308

- Run 2 (Ensemble gpt-4o): an ensemble with majority voting of 3 best prompts with gpt-4o as LLM.
- Run 3 (Single gpt-3.5-turbo): one single prompt for all 3 tasks with gpt-3.5-turbo as LLM.

For all 3 runs, ensembling with majority voting is only applied for True/False and MCQs as its answers are concrete and easy to combine from different prompts. On the other hand, ensembling free-text answers for factoid questions is more difficult due to the diversity of answer format, so we decided to use only 1 model for these questions. The results on the private set are presented in Table V.

It could be observed from the private test results that ensembling different prompts does not necessarily lead to superior results and potentially decreases the answer quality (Run 1 vs Run 3). Furthermore, the quality of the LLM is a big factor influencing accuracy of the answers.

Table V
TASK 2'S FINAL SUBMISSION ACCURACY ON PRIVATE TEST SET

Team	Overall	True/False & MCQ	Factoid
NOWJ	0.98	0.98	0.88
UIT-DarkCow	0.94	0.95	0.88
XOSA-NOWJ (Ensemble gpt-4o)	**0.91**	**0.92**	**0.77**
shikanokonokokoshitantan	0.89	0.90	0.77
se7enese	0.84	0.87	0.71
XOSA-NOWJ (Single gpt-3.5-turbo)	**0.78**	**0.86**	**0.22**
XOSA-NOWJ (Ensemble gpt-3.5-turbo)	**0.76**	**0.83**	**0.22**
DSCS@ALQAC	0.64	0.67	0.33
Alpha	0.5	0.5	0.5

VI. CONCLUSION

In this competition, our team experimented with new techniques to improve both legal information retrieval and legal question and answer. For the information retrieval task, the integration of ColBERT and the Multi-task learning model showed promising results, securing the top-2 in the private

dataset. However, using LLMs as re-rankers did not yield satisfactory results potentially due to lack of Vietnamese support. For Task 2 question and answer, utilizing automatic prompt optimizers got us top-3 in the private dataset despite relatively simple setups. To conclude, our methods demonstrate the potential of integrating LLMs in prompt engineering and enhancing legal retrieval accuracy, contributing to the facilitation of public access to legal knowledge.

ACKNOWLEDGMENT

Hai-Long Nguyen was funded by the Master, PhD Scholarship Programme of Vingroup Innovation Foundation (VINIF), code VINIF.2023.ThS.075.

REFERENCES

[1] H.-L. Nguyen, D.-Q. Nguyen, H.-T. Nguyen, T.-T. Pham, H.-D. Nguyen, T.-A. Nguyen, T.-H.-Y. Vuong, and H.-T. Nguyen, "Neco@alqac 2023: Legal domain knowledge acquisition for low-resource languages through data enrichment," in *2023 15th International Conference on Knowledge and Systems Engineering (KSE)*, 2023, pp. 1–6.

[2] H. N. Van, D. Nguyen, P. M. Nguyen, and M. Le Nguyen, "Miko team: Deep learning approach for legal question answering in alqac 2022," in *2022 14th International Conference on Knowledge and Systems Engineering (KSE)*. IEEE, 2022, pp. 1–5.

[3] L. Hoang, T. Bui, C. Nguyen, and L.-M. Nguyen, "Aiepu at alqac 2023: deep learning methods for legal information retrieval and question answering," in *2023 15th International Conference on Knowledge and Systems Engineering (KSE)*. IEEE, 2023, pp. 1–6.

[4] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[5] H.-L. Nguyen, D.-Q. Nguyen, H.-T. Nguyen, T.-T. Pham, H.-D. Nguyen, T.-A. Nguyen, H.-T. Nguyen *et al.*, "Neco@ alqac 2023: Legal domain knowledge acquisition for low-resource languages through data enrichment," in *2023 15th International Conference on Knowledge and Systems Engineering (KSE)*. IEEE, 2023, pp. 1–6.

[6] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," *CoRR*, vol. abs/1810.04805, 2018. [Online]. Available: http://arxiv.org/abs/1810.04805

[7] C. Nguyen, P. Nguyen, T. Tran, D. Nguyen, A. Trieu, T. Pham, A. Dang, and L.-M. Nguyen, "Captain at coliee 2023: efficient methods for legal information retrieval and entailment tasks," *arXiv preprint arXiv:2401.03551*, 2024.

[8] R. Nogueira, Z. Jiang, R. Pradeep, and J. Lin, "Document ranking with a pretrained sequence-to-sequence model," in *Findings of the Association for Computational Linguistics: EMNLP 2020*, T. Cohn, Y. He, and Y. Liu, Eds. Online: Association for Computational Linguistics, Nov. 2020, pp. 708–718. [Online]. Available: https://aclanthology.org/2020.findings-emnlp.63

[9] H.-T. Nguyen, M.-K. Phi, X.-B. Ngo, V. Tran, L.-M. Nguyen, and M.-P. Tu, "Attentive deep neural networks for legal document retrieval," *Artificial Intelligence and Law*, vol. 32, no. 1, pp. 57–86, 2024.

[10] H.-L. Nguyen, T.-B. Nguyen, T.-M. Nguyen, H.-T. Nguyen, and H.-Y. T. Vuong, "Vlh team at alqac 2022: Retrieving legal document and extracting answer with bert-based model," in *2022 14th International Conference on Knowledge and Systems Engineering (KSE)*. IEEE, 2022, pp. 1–6.

[11] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, "Unsupervised cross-lingual representation learning at scale," *CoRR*, vol. abs/1911.02116, 2019. [Online]. Available: http://arxiv.org/abs/1911.02116

[12] S. Robertson, H. Zaragoza *et al.*, "The probabilistic relevance framework: Bm25 and beyond," *Foundations and Trends® in Information Retrieval*, vol. 3, no. 4, pp. 333–389, 2009.

[13] N. Hai Long, T. H. Y. Vuong, H. T. Nguyen, and X.-H. Phan, "Joint learning for legal text retrieval and textual entailment: Leveraging the relationship between relevancy and affirmation," in *Proceedings of the Natural Legal Language Processing Workshop 2023*, D. Preoțiuc-Pietro, C. Goanta, I. Chalkidis, L. Barrett, G. Spanakis, and N. Aletras, Eds. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 192–201. [Online]. Available: https://aclanthology.org/2023.nllp-1.19

[14] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019, pp. 3982–3992.

[15] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186. [Online]. Available: https://aclanthology.org/N19-1423

[16] K. Santhanam, O. Khattab, J. Saad-Falcon, C. Potts, and M. Zaharia, "Colbertv2: Effective and efficient retrieval via lightweight late interaction," *arXiv preprint arXiv:2112.01488*, 2021.

[17] O. Khattab and M. Zaharia, "Colbert: Efficient and effective passage search via contextualized late interaction over bert," in *Proceedings of the 43rd International ACM SIGIR conference on research and development in*

Information Retrieval, 2020, pp. 39–48.

[18] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[19] R. Battle and T. Gollapudi, "The unreasonable effectiveness of eccentric automatic prompts," *arXiv preprint arXiv:2402.10949*, 2024.

[20] O. Khattab, A. Singhvi, P. Maheshwari, Z. Zhang, K. Santhanam, S. Vardhamanan, S. Haq, A. Sharma, T. T. Joshi, H. Moazam *et al.*, "Dspy: Compiling declarative language model calls into self-improving pipelines," *arXiv preprint arXiv:2310.03714*, 2023.

[21] K. Opsahl-Ong, M. J. Ryan, J. Purtell, D. Broman, C. Potts, M. Zaharia, and O. Khattab, "Optimizing instructions and demonstrations for multi-stage language model programs," *arXiv preprint arXiv:2406.11695*, 2024.