Phi-3 Meets Law: Fine-tuning Mini Language Models for Legal Document Understanding

Nguyen Huu Khanh[1], Nguyen Van Viet[2], Nguyen The Vinh[2], Nguyen Huu Cong[1]

[1] Thai Nguyen University, Thai Nguyen, Vietnam [2] Thai Nguyen University of Information and Communication Technology, Thai Nguyen, Vietnam

Correspondence: Khanh H. Nguyen, khanhnh@tnu.edu.vn

Received: 08/07/2024, revised: 26/08/2024, accepted: 16/09/2024

Digital Object Identifier: 10.32913/mic-ict-research-vn.v2024.n3.1281

Abstract: This study explored the application of Microsoft's Phi-3 model to legal document understanding. We fine-tuned Phi-3-mini-4k-instruct, a compact yet powerful language model, on CaseHOLD's 53,000+ multiple-choice questions designed to test legal case holding identification. Employing Low-Rank Adaptation (LoRA) and Quantized Low-Rank Adaptation (QLoRA) techniques, we optimized the fine-tuning process for efficiency. Our results demonstrated that the fine-tuned Phi-3-mini-4k-instruct model achieved an F1 score of 76.89, surpassing previous state-of-the-art models in legal document understanding. This performance was achieved with only 6000 training steps, highlighting Phi-3's rapid adaptability to domain specific tasks. The base Phi-3 model, without fine-tuning, achieved an F1 score of 64.89, outperforming some specialized legal models. Our findings underscored the potential of compact language models like Phi-3 in specialized domains when properly fine-tuned, offering a balance between model size, training efficiency, and performance. This research contributes to the growing body of work on adapting large language models to specific domains, particularly in the legal field.

Keywords: *Phi-3, CaseHold, fine-tuning, QLoRA, supervised fine-tuning*

Tiêu đề: **Phi-3 với Luật pháp: Tinh chỉnh mô hình ngôn ngữ nhỏ trong việc hiểu văn bản luật**

Tóm tắt: Nghiên cứu này khám phá ứng dụng mô hình Phi-3 của Microsoft vào việc hiểu văn bản pháp lý. Chúng tôi đã tinh chỉnh Phi-3-mini-4k-instruct, một mô hình ngôn ngữ nhỏ gọn nhưng mạnh mẽ, trên hơn 53.000 câu hỏi trắc nghiệm của CaseHOLD được thiết kế để kiểm tra khả năng nhận dạng vụ án pháp lý. Sử dụng các kỹ thuật thích ứng bậc thấp (LoRA) và thích ứng bậc thấp lượng tử (QLoRA), chúng tôi đã tối ưu hóa quy trình tinh chỉnh để đạt hiệu quả. Kết quả của chúng tôi cho thấy mô hình Phi-3-mini-4k-instruct tinh chỉnh đạt điểm F1 là 76,89, vượt qua các mô hình tiên tiến trước đây về khả năng hiểu văn bản pháp lý. Hiệu suất này đạt được chỉ với 6000 bước đào tạo, làm nổi bật khả năng thích ứng nhanh chóng của Phi-3 với các tác vụ cụ thể theo miền. Mô hình Phi-3-mini-4k-instruct cơ bản, không cần tinh chỉnh, đã đạt điểm F1 là 64,89, vượt trội hơn một số mô hình pháp lý chuyên biệt. Kết quả thực nghiệm của chúng tôi nhấn mạnh tiềm năng của các mô hình ngôn ngữ nhỏ gọn như Phi-3 trong các miền chuyên biệt khi được tinh chỉnh đúng cách, mang lại sự cân bằng giữa quy mô mô hình, hiệu quả đào tạo và hiệu suất. Nghiên cứu này góp phần vào khối lượng công việc ngày càng tăng về việc điều chỉnh các mô hình ngôn ngữ lớn cho các miền cụ thể, đặc biệt là trong lĩnh vực pháp lý.

Từ khóa: *Phi-3, CaseHold, fine-tuning, QLoRA, supervised fine-tuning*

I. INTRODUCTION

Understanding legal documents is important in today's judicial and regulatory climate [1]. A legal document is any text that has legal significance or is used in a legal context to communicate rights, responsibilities, or procedures [2]. Legal practitioners, politicians, and researchers increasingly rely on digital texts, making it critical to comprehend legal papers swiftly and accurately. However, the language used in these documents is often characterized by specialized vocabulary, complex sentence structures and formal. Statutes, case law, contracts, and regulatory filings are well-known examples for their complex structure, specialized terminology, and sophisticated logic. Moreover, accurately interpreting legal language requires not only linguistic proficiency but also an understanding of legal principles, rules of interpretation and precedent. Thus, understanding legal documents is a challenging task for both human and

machines to correctly comprehend [3–5].

Numerous studies were conducted to address the challenges of understanding legal documents using various approaches. For example, pioneer techniques relied on rule-based systems [6–8], which were accurate but time-consuming and difficult to scale. More recent advances used machine learning techniques, namely as pre-trained language models, to automate and improve the comprehension of legal documents. These pre-trained models, including K-BERT [9], Lawformer [10], JuriBert [11], jurBert [12], Paraformer [13], XLM-RoBERTa [14] and their derivatives, although exhibited promising performance in the legal domain, they all shared the same challenge as lack of understanding of natural language due to being trained on small datasets. This problem could be alleviated by utilizing large language models (e.g., Llama, GPT). However, these oversized models are primarily accessible to big tech companies (e.g., Google and Microsoft) where computational resources are available, making it challenging for researchers to conduct experiments [15]. This issue is not limited to the legal domain but extends to other similar research settings [16]. Therefore, there is a strong need for a mid-to-mini-sized pre-trained model that can effectively balance natural language understanding and model size [16, 17], particularly when specialized in the legal documents domain.

In response to this need, Microsoft recently released Phi-3 [18] on March 22, 2024, as pre-trained mid-to-mini-sized models. However, to the best of our knowledge, there is no evidence demonstrating the performance of Phi-3 in the legal document domain. Therefore, the purpose of this research is to address this gap by conducting an experiment to evaluate the efficacy of Phi-3 on legal documents.

To meet that objective, we fine-tuned Phi-3-mini-4k with the CaseHOLD [19] dataset, a benchmark intended exclusively for assessing legal document understanding. Figure 1 presents overview of process in this study. Our goal is to experiment Phi-3-mini-4k-instruct and evaluate its potential for comprehending and processing legal texts. This study intends to add to the body of knowledge in legal Natural Language Processing (NLP) by giving empirical evidence on the performance of a mini size pre-trained model, thus opening the way for more efficient and accessible legal document analysis applications.

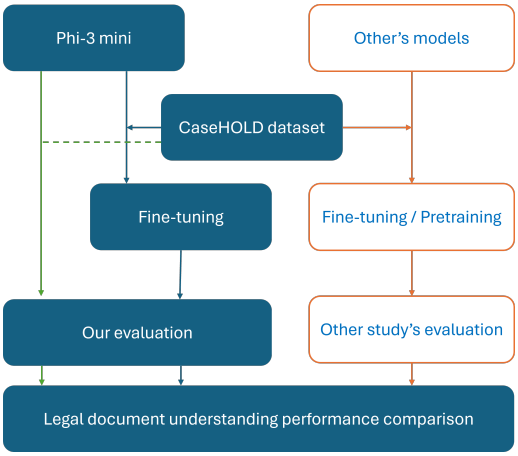

Figure 1. Overview of process in this study

II. MATERIALS AND METHODS

1. The base model

In our study, we used Phi-3 (introduced by Microsoft on March 22, 2024) as the base model for the fine-tuning task [18]. Compared to many state-of-the-art models, Phi-3 is notable for its high-quality parameters and size. The term "high-quality parameters" refers to high-quality data curation, extensive post-training, and reinforcement learning from human feedback (RLHF). Although the number of parameters is relatively small compared to other large language models, evidence [18] shows that Phi-3's performance surpasses that of larger models (e.g., Mixtral 8x7B and GPT-3.5). There is currently no standardized definition for terms like "mini" or "small" models in the literature. However, in the field of AI or as in our study, models can be classified based on the number of parameters: "tiny" refers to models with fewer than 1 billion parameters, "mini" ranges from 1 to 5 billion parameters, and "small" covers models with 5 to 10 billion parameters, following the terminology used by Microsoft. Phi-3 was released in several versions including Mini (3.8B parameters), Small (7B parameters), Medium (14B parameters) and Vision (4.2B parameters). In this study, we chose Phi-3-mini-4k-instruct model from the Phi-3 family due to several compelling factors. First, despite its compact size, Phi-3-mini-4k-instruct has demonstrated impressive performance comparable to significantly larger models, suggesting strong language understanding capabilities. This balance of size and performance is particularly attractive for resource-constrained settings. Second, Phi-3's training methodology aligns well with the requirements of legal document analysis, where accurate and nuanced language understanding is paramount. We hypothesize that these characteristics would translate well to the legal domain, making Phi-3-mini-4k-instruct a strong candidate for fine-tuning on the CaseHOLD dataset.

2. CaseHold Dataset

The CaseHOLD dataset serves as a benchmark for evaluating legal document under-standing in the U.S. legal domain. It comprises over 53,000 multiple-choice questions

Table I
EXAMPLE OF CASEHOLD DATA ITEM

Citing_prompt	CURIAM. Affirmed. See Adams v. State, 76 So.3d 367 (Fla. 3d DCA 2011) (<HOLDING>); accord Little v. State, 77 So.3d 722 (Fla. 3d
Holding_0	holding that section 89313 as amended by section 893101 florida statutes 2002 is constitutional
Holding_1	holding that the legislature amended section 8120142d florida statutes now renumbered as section 8120143c in 1992 to omit habitual offender penalties for the crime of felony petit theft
Holding_2	recognizing that the 1992 amendment to section 8930216 florida statutes changed the law
Holding_3	holding florida sexual predators act section 77521 florida statutes 2000 to be unconstitutional as violating procedural due process
Holding_4	holding that arbitration is not a civil action as that term is used in section 76873 florida statutes 1991
Label	0

designed to identify the relevant holding of a cited case. Each question presents a citing context from a judicial decision as the prompt (named "Citing_prompt"), five answer choices derived from holding statements within legal citations (named "holding_[0-4]") and an indicator for correct answer (named “label”). Table I shows an example of CaseHold data item. The correct answer corresponds to the citing text, while the four incorrect options are selected based on their similarity, making the task challenging. Constructed from the Harvard Law Library case law corpus, this dataset offers a robust framework for fine-tuning and evaluating advanced language models. The precision in extracting holding statements and the strategic selection of distractors highlight its value in advancing legal AI research.

3. Experimental setup

a) Data preprocessing

The CaseHOLD dataset, comprising fields like Citing_prompt, five holdings, and a label, required careful preprocessing to align with Phi-3’s fine-tuning instructions. Our first step involved creating a function to structure the data into a conversational format, integrating all fields effectively. This structured data was then passed to the Phi-3 tokenizer’s apply_chat_template function, adhering to the "<|user|>Question<|end|><|assistant|>Answer<|end|>" format for training, example of chat format is shown in Table II. During evaluation, we adjusted the process by sending only the Citing_prompt and the holdings, omitting the label to allow the model to predict the correct answer. The instruction for apply_chat_template in this phase was simplified to "<|user|>Question<|end|><|assistant|>", facilitating an accurate assessment of the model’s predictive capabilities in a realistic legal question-answering context.

Table II
EXAMPLE CHAT FORMAT FOR TRAINING

<s><|user|> Choose the best option to fill <HOLDING> based on this context, answer only with the number of option(for example: 0): CURIAM. Affirmed. See Adams v. State, 76 So.3d 367 (Fla. 3d DCA 2011) (<HOLDING>); ac-cord Little v. State, 77 So.3d 722 (Fla. 3d Options to choose: 0. holding that section 89313 as amended by section 893101 florida statutes 2002 is constitutional 1. holding that the legislature amended section 8120142d florida statutes now renumbered as section 8120143c in 1992 to omit habitual offender penalties for the crime of felony petit theft 2. recognizing that the 1992 amendment to section 8930216 florida statutes changed the law 3. holding florida sexual predators act section 77521 florida statutes 2000 to be unconstitutional as violating procedural due process 4. holding that arbitration is not a civil action as that term is used in section 76873 florida statutes 1991 <|end|><|assistant|> 0 <|end|>

b) LoRA

There are various fine-tuning approaches in the literature, ranging from full fine-tuning to partial fine-tuning [20]. Among these, Low-Rank Adaptation (LoRA) [20] is an innovative fine-tuning technique widely adopted by the AI community. Compared to other techniques that are resource-intensive and introduce latency, LoRA addresses these issues by training only the "intrinsic dimension" and then updating the original weights accordingly. Mathematically, LoRA could be expressed as follows [20]:

$$W = W_0 + \Delta W = W_0 + BA$$

$$B \in \mathbb{R}^{d \times r}, \quad A \in \mathbb{R}^{r \times k}$$

Where:

- W is the updated weight matrix.
- W_0 is the original pre-trained weight matrix.
- ΔW is the low-rank update applied to the original weights.
- B and A are low-rank matrices.
- r is the rank, a much smaller dimension than d or k.

The true value of LoRA lies in training matrix A and B instead of full training the original pretrained weight W_0. As such, LoRA significantly reduces the computational resources (i.e., training on low-rank matrices) required for fine-tuning while maintaining performance (i.e., through W). This approach enables efficient adaptation allowing researchers to explore specialized applications such as legal document understanding without the need for extensive hardware resources. In our experiment, B and A are low-rank matrices with a rank (r) of 16, which means the updates only happen in a smaller 16-dimensional space. We applied an alpha value of 16, which scales the updates

to ensure they have a noticeable effect on the model's performance. To prevent the model from overfitting, we used dropout with a rate of 0.05 (5%), meaning a small portion of updates was randomly ignored during training to help the model generalize better. The LoRA updates were applied to key parts of the model, including the attention mechanism (key (k_proj), query (q_proj), value (v_proj), and output (o_proj) projections) and the feed-forward networks (gate ('gate_proj'), down (down_proj), and up (up_proj) projections). This setup balances adaptation capacity with parameter efficiency, allowing for effective fine-tuning of the Phi-3-mini-4k-instruct model on the CaseHOLD legal dataset.

c) QLoRA

To speed up the training process, we utilized Quantized Low-Rank Adaptation (QLoRA) technique [21]. The basic idea of QLoRA is to reduce the precision of the model's parameters (weights) by converting them from high precision (e.g., 32-bit floating point) to a lower precision format (e.g., 8-bit or even lower). As such, it reduces the memory required to store the model and speeds up computation. In our study, we set the model to implement 4-bit quantization (load_in_4bit=True) with the "nf4" quantization type, balancing compression and accuracy. We set the compute dtype to bfloat16 for improved training stability and employed double quantization (bnb_4bit_use_double_quant=True) to further optimize memory usage. This setup allowed us to leverage the full potential of Phi-3 while minimizing resource requirements.

d) Supervised fine-tuning

Supervised fine-tuning (SFT) was originally proposed by Ouyang et al. [22]. It is an important stage in aligning LLMs that involves training them on a supervised dataset of high-quality outputs. This procedure involves fine-tuning the model with a next token prediction aim, similar to pretraining, but with a focus on simulating desired behaviors. SFT is computationally efficient and improves the model's performance in terms of instruction compliance, coherence, and accuracy. Given our limited computational resources and the supervised nature of the CaseHOLD dataset, we determined that SFT is particularly well-suited for our study. Accordingly, we employed supervised fine-tuning using an easy-to-use API of Hugging Face, named Supervised Fine-tuning Trainer (SFTTrainer). This API facilitated the creation of our SFT model, optimizing the use of our available resources.

e) Training hyperparameters

In fine-tuning Phi-3-mini-4k-instruct on the CaseHOLD dataset, we carefully selected a set of hyper parameters to optimize the model's performance in legal document understanding. We employed a relatively low learning rate of 0.0001 to ensure stable training, coupled with small batch sizes to accommodate the model's size and memory constraints. The Adam optimizer was used with standard beta values, while a linear learning rate scheduler with a 10% warm-up ratio helped manage the learning rate throughout training. We ran the fine-tuning process for 6000 steps, balancing between sufficient learning and avoiding over-fitting.

f) Computational Resources and Training Environment

The fine-tuning process was conducted using an NVIDIA RTX 3060 GPU with 12GB of VRAM, an Intel Core i7 12700F CPU, and 32GB of RAM. The training spanned 32 hours. The software environment included PEFT version 0.11.1, Transformers version 4.41.2, Pytorch version 2.3.0+cu121, Datasets version 2.19.1, and Tokenizers version 0.19.1. These configurations ensured efficient processing and optimal model performance.

g) Evaluation metrics

Fig. 2 showcase the process to evaluate a model's response, we considered three possible answer formats: (C1) a single digit indicating the correct answer position (e.g., '1'), (C2) the exact text of the correct answer (e.g., 'correct answer') and (C3) the correct answer preceded by '1.' (e.g., '1. correct answer'). For the first type, evaluation is straightforward with a direct comparison to the label field. However, for the second and third formats, we employed the ROUGE metrics to account for potential variations in text representation.

The ROUGE (Recall-Oriented Understudy for Gisting Evaluation) metrics offer a flexible means of comparing textual data based on overlapping units such as n-grams and longest common subsequences. Specifically, we used ROUGE-1, which calculates the overlap of unigrams (single words), ROUGE-2, which assesses the overlap of bigrams (two consecutive words), ROUGE-L, which measures the longest common subsequence between two texts, and ROUGE-Lsum, which is a variant tailored for summarization tasks by comparing sequences across the entire text.

Using ROUGE allows us to capture partial matches between the predicted answer and the ground truth, accommodating slight variations in phrasing or formatting, which may not affect the correctness of the answer. For example, the ground truth might contain two spaces between words due to mistyping (e.g., correct answer), while the model produces a response with only one space (correct answer). In such cases, a strict text-to-text comparison would incorrectly mark the answer as wrong, even though the content is identical. The ROUGE metric, however, evaluates the

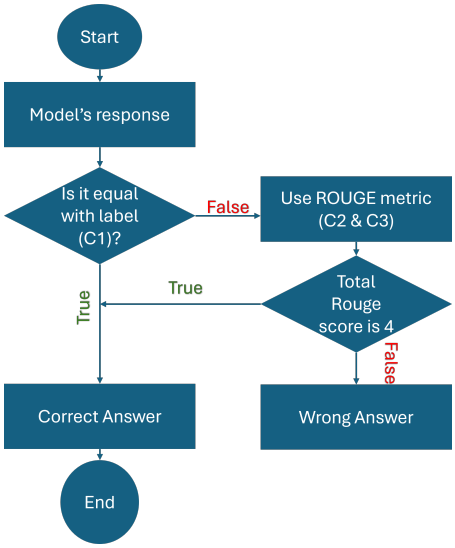

Figure 2. Process to evaluate a model's response

level of overlap between the prediction and ground truth, thereby handling such minor formatting inconsistencies and ensuring that the model is not penalized for trivial errors.

An answer is considered correct if all four ROUGE scores ROUGE-1, ROUGE-2, ROUGE-L, and ROUGE-Lsum each yield a perfect score of 1.0. Since each ROUGE metric outputs a value between 0 and 1, a perfect match across all metrics results in a total sum of 4.0. This strict criterion ensures that the model's response matches the ground truth exactly in terms of content, while still being robust to insignificant formatting differences like extra spaces or punctuation variations.

Finally, the model's overall performance was measured using the F1 score. By using the F1 score in the final step, we compared the performance between the Phi-3 models (fine-tuned and not fine-tuned) in this study and other models from Zheng's [19] and Chalkidis's [23] results, which also used the CaseHOLD dataset. This comparison provided a comprehensive measure of the model's accuracy and effectiveness.

III. RESULTS AND DISCUSSION

The results of our study, along with the comparative results from Zheng et al. [19] and Chalkidis et al. [23], are summarized in Table III. The base Phi-3-mini-4k-instruct model, with 3.8 billion parameters, achieved an F1 score of 66.84 without any fine-tuning. This performance is notably higher than the standard BERT model from [19]'s results (61.30) and even the BERT (double) model with twice the training steps (62.30). However, it falls short when compared to more specialized legal models such as Legal-BERT and Custom Legal-BERT, which achieved F1

Table III
EXPERIMENTS'S RESULTS

Model	Params	Training steps	Vocab size	F1 score
Zheng et al's result[19]				
BERT	110M	1M	30K	61.30
BERT (double)	110M	2M	30K	62.30
Legal-BERT	110M	2M	30K	68.00
Custom Legal-BERT	110M	2M	32K	69.50
Chalkidis et al's result[23]				
BERT	110M	1M	32K	70.80
RoBERTa	125M	100K	50K	71.40
DeBERTa	139M	1M	50K	72.60
Longformer	149M	65K	50K	71.90
Bigbird	127M	1M	32K	70.80
Legal-BERT	110M	1M	32K	75.30
CaseLaw-BERT	110M	2M	32K	75.40
Our result				
Phi-3 mini 4K (base)	3.8B	-	32K	66.84
Fine-tuned Phi-3 mini 4K	3.8B	6K	32K	75.76

scores of 68.00 and 69.50, respectively. This is indicative of the base model's general-purpose capabilities, which, without fine-tuning, do not fully leverage the domain-specific nuances required for optimal performance in legal document understanding. Upon fine-tuning with 6,000 training steps, the Phi-3-mini-4k-instruct's performance significantly improved, achieving an F1 score of 75.76. This result surpasses all models in [19] and [23] results, including Legal-BERT (75.30) and CaseLaw-BERT (75.40). This improvement highlights the effectiveness of fine-tuning in enhancing the model's understanding of legal documents and underscores the importance of domain-specific fine-tuning.

A comparative analysis of parameter efficiency and training steps reveals that the base Phi-3-mini-4k-instruct model, with its 3.8 billion parameters, is significantly larger than the models used by [19] and [23], which typically have around 110M to 149M parameters. Despite the larger parameter size, the base model's performance was initially modest, highlighting the necessity of fine-tuning. Fine-tuning with only 6,000 training steps, compared to the millions of steps used by other models, demonstrates the efficiency and rapid adaptability of Phi-3-mini-4k-instruct in acquiring domain-specific knowledge. Several key points emerge from our analysis: the scalability and fine-tuning potential of the Phi-3-mini-4k-instruct model, driven by its large parameter count, enable it to capture complex dependencies and patterns within the data. The dramatic performance improvement post-fine-tuning indicates that while the base model is competent, its true potential is

realized when it is adapted to specific domains through fine-tuning.

An interesting observation emerges when analyzing the impact of vocabulary size across the different studies. Results [19] demonstrated a marginal improvement in F1 score for Legal-BERT when increasing the vocabulary size by 2,000 tokens (from 30K to 32K). However, a similar vocabulary increase for the BERT model in [23] resulted in a substantial performance boost (from 61.3 to 70.8 F1 score). This discrepancy suggests that vocabulary size alone does not dictate performance gains and its impact is likely intertwined with other factors such as training data, model architecture, and fine-tuning methodologies. This observation raises the question of whether a larger vocabulary size, for instance, 50K for Phi-3, could further enhance its performance. Further investigation is needed to disentangle the interplay between vocabulary size and other model parameters for optimizing legal document understanding.

While our findings are encouraging, this study has limitations. Firstly, both the CaseHOLD dataset and the Phi-3 model are English-centric, restricting the generalizability of our conclusions. Future work should explore multilingual legal datasets and models to address this constraint. Secondly, our evaluation focused solely on multiple choice question answering using CaseHOLD. Investigating Phi-3's capabilities on diverse legal tasks and datasets, such as LexGlue, COLIEE or CAIL, would provide a more comprehensive assessment of its strengths and weaknesses. Finally, our focus remained on zero-shot performance and a single model architecture. Exploring few-shot prompting techniques and evaluating other promising open-source models, like LLaMa-3.1 or Gemma 2, could reveal further performance gains and provide valuable insights into the optimal model and training paradigm for legal document understanding.

IV. CONCLUSION

Our study demonstrated the effectiveness of fine-tuning Phi-3-mini-4k-instruct for legal document understanding. The model's performance improvement from an F1 score of 66.84 to 75.76 after fine-tuning highlights its adaptability and potential in specialized domains. This research showcases the capabilities of compact yet powerful language models in tackling complex tasks such as legal text analysis. While our study demonstrated the efficacy of Phi-3-mini-4k-instruct in legal document understanding, several promising avenues for future research emerge. Firstly, exploring the impact of a larger vocabulary size on Phi-3's performance could reveal further optimization potential. Secondly, expanding the evaluation to encompass diverse legal tasks and datasets, such as those within LexGLUE, COLIEE, or CAIL, would provide a more holistic view of the model's capabilities and generalizability across different legal domains. Thirdly, investigating the effectiveness of few-shot prompting techniques with Phi-3 could uncover new avenues for performance enhancement in resource-constrained scenarios. Lastly, benchmarking against other cutting edge open-source models, including LLaMa-3.1 and Gemma2, would contribute valuable insights into the comparative strengths and weaknesses of different architectures and pre-training paradigms for legal NLP applications.

REFERENCES

[1] Y. Li, G. Hu, J. Du, H. Abbas, and Y. Zhang, "Multi-task reading for intelligent legal services," *Future Generation Computer Systems*, vol. 113, pp. 218–227, 2020.

[2] S. Yurii, P. Nataliia, T. Tetiana, P. Tetiana, and H. Stanislav, "Official document as a legal act: Essential aspects," *J. Legal Ethical & Regul. Isses*, vol. 22, p. 1, 2019.

[3] F. u. Hassan, T. Le, and X. Lv, "Addressing legal and contractual matters in construction using natural language processing: A critical review," *Journal of Construction Engineering and Management*, vol. 147, no. 9, p. 03121004, 2021.

[4] S. R. Ahmad, D. Harris, and I. Sahibzada, "Understanding legal documents: classification of rhetorical role of sentences using deep learning and natural language processing," in *2020 IEEE 14th International Conference on Semantic Computing (ICSC)*. IEEE, 2020, pp. 464–467.

[5] G. Marino, D. Licari, P. Bushipaka, G. Comandé, T. Cucinotta *et al.*, "Automatic rhetorical roles classification for legal documents using legal-transformeroverbert," in *CEUR WORKSHOP PROCEEDINGS*, vol. 3441. CEUR-WS, 2023, pp. 28–36.

[6] I. Timmer and R. Rietveld, "Rule-based systems for decision support and decision-making in dutch legal practice. a brief overview of applications and implications," *Droit et Societe*, vol. 103, pp. 517–534, 11 2019.

[7] D. Bourcier and G. Clergue, "From a rule-based conception to dynamic patterns. analyzing the self-organization of legal systems," *Artificial Intelligence and Law*, vol. 7, pp. 211–225, 1999.

[8] K. Pal and J. A. Campbell, "An application of rule-based and case-based reasoning within a single legal knowledge-based system," *ACM SIGMIS Database: the DATABASE for Advances in Information Systems*, vol. 28, pp. 48–63, 9 1997. [Online]. Available: https://dl.acm.org/doi/10.1145/277339.277344

[9] W. Liu, P. Zhou, Z. Zhao, Z. Wang, Q. Ju, H. Deng, and P. Wang, "K-bert: Enabling language representation with knowledge graph," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, pp. 2901–2908, 4 2020. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/5681

[10] C. Xiao, X. Hu, Z. Liu, C. Tu, and M. Sun, "Lawformer: A pre-trained language model for chinese legal long documents," *AI Open*, vol. 2, pp. 79–84, 2021. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S2666651021000176

[11] S. Douka, H. Abdine, M. Vazirgiannis, R. E. Hamdani, and D. R. Amariles, "Juribert: A masked-language model adaptation for french legal text," *Natural Legal Language*

Processing Workshop 2021, pp. 95–101, 2021. [Online]. Available: https://aclanthology.org/2021.nllp-1.9.pdf

[12] M. Masala, R. Iacob, A. S. Uban, M.-A. Cidotã, H. Velicu, T. Rebedea, and M. Popescu, "jurbert: A romanian bert model for legal judgement prediction," *Natural Legal Language Processing Workshop 2021*, pp. 86–94, 2021. [Online]. Available: https://aclanthology.org/2021.nllp-1.8.pdf

[13] 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, pp. 57–86, 3 2024. [Online]. Available: https://link.springer.com/10.1007/s10506-022-09341-8

[14] H. N. Van, D. Nguyen, P. M. Nguyen, and M. Nguyen, "Miko team: Deep learning approach for legal question answering in alqac 2022," *2022 14th International Conference on Knowledge and Systems Engineering (KSE)*, pp. 1–5, 2022. [Online]. Available: https://arxiv.org/pdf/2211.02200

[15] M. Kejriwal, "Ai in industry today," in *Artificial Intelligence for Industries of the Future: Beyond Facebook, Amazon, Microsoft and Google*. Springer, 2022, pp. 47–73.

[16] J. Sauvola, S. Tarkoma, M. Klemettinen, J. Riekki, and D. Doermann, "Future of software development with generative ai," *Automated Software Engineering*, vol. 31, no. 1, p. 26, 2024.

[17] X. Han, Z. Zhang, N. Ding, Y. Gu, X. Liu, Y. Huo, J. Qiu, Y. Yao, A. Zhang, L. Zhang *et al.*, "Pre-trained models: Past, present and future," *AI Open*, vol. 2, pp. 225–250, 2021.

[18] M. Abdin and S. A. Jacobs, "Phi-3 technical report: A highly capable language model locally on your phone," 4 2024. [Online]. Available: http://arxiv.org/abs/2404.14219

[19] L. Zheng, N. Guha, B. R. Anderson, P. Henderson, and D. E. Ho, "When does pretraining help?: assessing self-supervised learning for law and the casehold dataset of 53,000+ legal holdings," in *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law*. ACM, 6 2021, pp. 159–168.

[20] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models," 6 2021. [Online]. Available: http://arxiv.org/abs/2106.09685

[21] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, "Qlora: Efficient finetuning of quantized llms," in *Advances in Neural Information Processing Systems*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, Eds., vol. 36. Curran Associates, Inc., 5 2023, pp. 10 088–10 115. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2023/file/1feb87871436031bdc0f2beaa62a049b-Paper-Conference.pdf

[22] L. Ouyang, J. Wu, and X. Jiang, "Training language models to follow instructions with human feedback," in *Advances in Neural Information Processing Systems*, vol. 35. Curran Associates, Inc., 3 2022, pp. 27 730–27 744. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2022/file/b1efde53be364a73914f58805a001731-Paper-Conference.pdf

[23] I. Chalkidis, A. Jana, D. Hartung, M. Bommarito, I. Androutsopoulos, D. M. Katz, and N. Aletras, "Lexglue: A benchmark dataset for legal language understanding in english," *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, vol. 1, pp. 4310–4330, 10 2022. [Online]. Available: https://arxiv.org/abs/2110.00976v4

Nguyen Huu Khanh has graduated with a Master's degree in Computer Science from the University of Information and Communications Technology - Thai Nguyen University since 2022 and is currently a PhD student here since 2023. His main research interests are Computer Science, Natural Language Processing and Computer Vision.
Email: khanhnh@tnu.edu.vn

Nguyen Van Viet received the Bachelor's Information Technology at Thai Nguyen University in 2009 and Master's degree Information Technology at Manuael S. Enverga University Foundation, Lucena City, Philippines in 2012. He worked as a lecturer at the Faculty of Information Technology, School of Information and Communication Technology, Thai Nguyen University from 2009. Now, he is a researcher at the Thai Nguyen University of Information and Communication Technology, Thai Nguyen, Vietnam. Office address: University of Information and Communication Technology, Thai Nguyen University, Thai Nguyen, Vietnam.
Email: nvviet@ictu.edu.vn

Nguyen The Vinh received his PhD in Computer Science from Texas Tech University in 2020. He is currently a lecturer in Software Engineering at the University of Information and Communication Technology - Thai Nguyen University. His main research interests are Computer Science and AI.
Email: vinhnt@ictu.edu.vn

Nguyen Huu Cong received his PhD in automatic control from Hanoi University of Science and Technology in 2003 and was promoted to Associate Professor in 2007. His main research interests are automatic control and optimal control for objects with distributed and slowly changing parameters.
Email: conghn@tnu.edu.vn