Legal Knowledge and Information Systems
G. Sileno et al. (Eds.)

doi:10.3233/FAIA230967

LogiLaw Dataset Towards Reinforcement Learning from Logical Feedback (RLLF)

Ha-Thanh Nguyen [a] Wachara Fungwacharakorn [a] Ken Satoh [a]
[a] *National Institute of Informatics, Japan*

Abstract. Large Language Models (LLMs) face limitations in logical reasoning, which restrict their applicability in critical domains such as law. Current evaluation methods often lead to inaccurate assessments of LLMs' capabilities due to their simplicity. This paper presents a refined evaluation method for assessing LLMs' capability to answer legal questions by eliminating the possibility of obtaining correct responses by chance. Furthermore, we introduce the LogiLaw dataset, which aims to enhance the models' logical reasoning capacities in general and legal reasoning specifically. By leveraging the refined evaluation technique, the LogiLaw dataset, and the proposed Reinforcement Learning from Logical Feedback (RLLF) approach, our work aims to open new avenues for research to bolster LLMs' performance in law and other logic-intensive disciplines while addressing the shortcomings of conventional evaluation approaches.

Keywords. LLMs, GPT, legal reasoning, LogiLaw, logic programming

1. Introduction

The legal domain presents a unique set of challenges for artificial intelligence (AI) applications, as it demands high-quality reasoning capabilities, the ability to interpret complex language structures, and accurate decision-making based on legal precedence and context. While the advancement of deep learning techniques has led to the development of sophisticated language models, their current inability to exhibit reliable and consistent logical reasoning limits their applicability for critical functions such as legal advice and case analysis.

Large Language Models (LLMs) like GPT-3 [1] and GPT-4 [2] excel in various language tasks, often providing promising results for areas such as natural language understanding and generation, translation, and question answering. However, their utility in the legal domain is hindered by their weakness in processing complex logic and reasoning requirements.

Conventional AI evaluation approaches are typically simplistic in nature, rendering the assessments of these models' capabilities inaccurate. This issue highlights the need for improved evaluation methods that accurately gauge the logical reasoning ability of LLMs configured for operations in the legal domain.

In this paper, we propose an enhanced evaluation method and introduce the LogiLaw dataset to tackle the challenge of measuring and improving LLMs' logical reasoning capabilities, particularly within the legal domain. Our improved evaluation protocol

aims to eliminate the possibility of LLMs obtaining correct responses by chance, thereby painting a more accurate picture of their true reasoning skills. The LogiLaw dataset is designed to support reinforcement learning to bolster the logical reasoning performance of LLMs when dealing with legal texts and queries.

The main contributions of this work are as follows:

- We propose a refined evaluation method that requires LLMs to generate Prolog code to answer legal questions, followed by a Prolog engine verification process that ensures the models rely on accurate reasoning pathways instead of superficial patterns or luck.
- We introduce the LogiLaw dataset, which captures the generated Prolog code and its verification results to aid in enhancing the logical reasoning abilities of LLMs through reinforcement learning.
- We present the Reinforcement Learning from Logical Feedback (RLLF) approach as a novel method to improve LLMs' logical reasoning performance by leveraging the LogiLaw dataset and focusing on logical feedback rather than relying on subjective human feedback.

The remainder of the paper is structured as follows: Sections 2 and 3 discuss the background, related work, and our improved evaluation method. Sections 4 and 5 present the LogiLaw dataset, the RLLF approach, and experimental results. Lastly, Sections 6 and 7 explore future research directions and conclude the paper.

2. Background and Related Work

The use of artificial intelligence (AI) and machine learning in the legal field has attracted significant attention in recent years, resulting in the development of numerous state-of-the-art models capable of performing various legal text processing tasks, such as contract analysis and case law retrieval [3,4,5,6]. Deep learning models, which offer automated feature extraction, have been utilized not only in similarity matching tasks but also in other semantic matching tasks such as question answering [7,8], machine reading comprehension [9,10], image retrieval [11], and entity matching [12,13]. Thus, these models have proven valuable in the legal domain, where accurate information retrieval and abductive reasoning skills are essential.

Legal retrieval tasks involve identifying relevant legal regulations based on given queries and are fundamental components of intelligent legal counsel systems [3,4,5,6]. Moreover, deep learning and transfer learning methods have successfully addressed the limited data challenges commonly encountered in legal retrieval tasks, leading to the development of various effective approaches [7,14,15,16,17]. Retrieval tasks form the foundation of many legal text processing tasks, such as contract analysis and case law retrieval, which require making inferences and educated guesses about what information is likely to be relevant based on the query and available data. These tasks also align closely with abductive reasoning, aiming to explain observations and arrive at conclusions based on limited information.

The introduction of transformer-based models, such as BERT-PLI [14], Legal BERT [18], BERTLaw [19], and ParaLaw Nets [20], have greatly impacted the performance of legal text processing tasks, including legal document classification, legal text sum-

marization, and legal question answering. These models excel at capturing complex relationships and dependencies in legal texts, demonstrating strong performance across a wide range of tasks.

GPT models, such as GPT-2 [21], GPT-3 [1], GPT-3.5 (ChatGPT) [22], and GPT-4 [2], have also contributed to the evolution of automated language understanding, performing remarkably well on tasks like translation and question answering. However, challenges related to optimizing for user satisfaction and ensuring logical consistency remain, as user satisfaction and logical consistency do not always align. Consequently, reinforcement learning from human feedback (RLHF) might exhibit limitations in improving model alignment with user intent, necessitating the exploration of alternative approaches to tackle semantic challenges and scale learning techniques.

In summary, AI and machine learning have demonstrated great potential in the field of legal text processing. Deep learning models like transformers and GPT models have achieved remarkable performance across various tasks. However, challenges in optimizing for user satisfaction, logical consistency, and limited data availability necessitate more advanced techniques for enhancing legal text processing capabilities, such as improved evaluation methods and datasets designed for logical reasoning, as proposed in this work.

Two recent papers [23,24] further highlight the challenges faced by GPT models in logical reasoning tasks. In one study, various GPT models were tested on the xNot360 dataset [23] to assess their negation detection abilities, and it was found that the models exhibited significant inconsistencies in handling negations, ranging from a high-performing GPT-4 model to the relatively poorer-performing GPT-3.5. This result underscores the need for further refinement in handling negation within the GPT models.

Another study [24] focused on assessing the abductive reasoning capability of state-of-the-art legal reasoning models. It was observed that pre-trained legal models were not necessarily more effective than the original BERT Base model on abductive reasoning, concluding that the performance of these models indicates that abductive reasoning remains a challenging problem. This finding serves as a motivation for our work on developing improved evaluation methods and the LogiLaw dataset to address these limitations in logical reasoning capabilities.

3. Improved Method for Evaluation of Reasoning

Conventional evaluation methods for Large Language Models (LLMs) often focus on simplistic assessments, such as binary classification tasks. While these methods provide a general understanding of a model's performance, they also have limitations in accurately evaluating the model's reasoning capabilities. This is particularly critical when dealing with complex domains like law, where logical reasoning is essential. The enhanced evaluation method proposed in this paper aims to address these limitations by ensuring a more accurate assessment of the LLMs' logical reasoning capacities.

3.1. Shortcomings of Traditional Binary Classification Evaluation

Binary classification evaluation requires the model to output a simple "Yes" or "No" response to a given question. Although this approach can provide a general sense of the

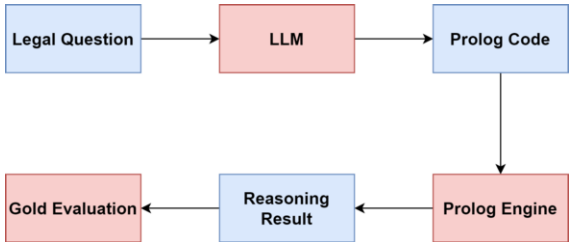

Figure 1. Improved Evaluation Method for LLMs involving the generation of Prolog code and verification using an independent Prolog engine

model's competence, it may not accurately represent the depth of its reasoning abilities. Consequently, models that obtain correct responses by chance or by exploiting superficial patterns in the data, rather than genuinely understanding the underlying logic, can achieve deceptively high performance scores. These shortcomings make it difficult to accurately evaluate the true logical reasoning abilities of LLMs, particularly in the context of the legal domain.

3.2. Requiring the Model to Generate Prolog Code

To overcome the limitations of binary classification evaluation, our improved method requires the LLM to generate Prolog logic programming code in response to a given legal question. Prolog is a powerful and expressive language for representing complex legal arguments and reasoning, and by compelling the model to provide Prolog code as its output, we can ensure a more comprehensive assessment of its logical reasoning capabilities.

This approach has several benefits over binary classification evaluation. First, it allows LLMs to showcase their capacity for understanding and constructing complex logical arguments based on the input data. Second, it eliminates the possibility of models obtaining correct answers by chance or exploiting superficial patterns, as the generated Prolog code must be syntactically and semantically valid to represent the correct logical reasoning.

3.3. Verification of the Generated Code Using an Independent Prolog Engine

After the LLM generates Prolog code in response to a given legal question, we use an independent Prolog engine to verify the correctness of the generated code. The engine executes the provided Prolog code and evaluates its output based on the predefined set of inference rules. If the engine's output matches the expected output based on the ground truth, the model is considered to have successfully provided accurate logical reasoning for the given question.

The use of an independent Prolog engine for verification serves as a vital component of our evaluation method, as it ensures that the model output is not only syntactically valid Prolog code but also provides the correct logical reasoning. Furthermore, this approach decouples the evaluation from the model itself, ensuring a fair and unbiased assessment of the LLM's logical reasoning capabilities.

In Figure 2, we depict the enhanced evaluation method for LLMs, which involves not only converting the legal question but also the related articles into Prolog code and

verifying them with an independent Prolog engine. The process includes the following interconnected components:

1. **Legal Question** (Q) and **Related Articles** (A_i): The input consists of a legal question Q and a set of related articles $\{A_1, A_2, \ldots, A_n\}$. This combination forms the starting point of the evaluation process.
2. **LLM**: The Large Language Model takes the input legal question Q and related articles $\{A_1, A_2, \ldots, A_n\}$, and generates the respective Prolog code as output (P_Q, P_{A_i}).
3. **Prolog Code**: The LLM generates a valid Prolog code representation of the legal question (P_Q) and the related articles (P_{A_i}). This Prolog code represents the LLM's attempt at capturing accurate logical reasoning.
4. **Prolog Engine**: The independent Prolog engine verifies the correctness of the generated Prolog codes (P_Q, P_{A_i}). This ensures the LLM's output demonstrates accurate logical reasoning and prevents relying on chance or superficial patterns.
5. **Reasoning Result**: Denoted as $R(P_Q, P_{A_i})$, this component shows the evaluation outcomes of the LLM's generated Prolog codes, indicating the LLM's competency in providing accurate logical reasoning.
6. **Gold Evaluation**: Represented as $G(P_Q, P_{A_i})$, this component compares the Reasoning Result $R(P_Q, P_{A_i})$ with the ground truth for the given legal question and related articles, assessing the LLM's logical reasoning ability as either correct or incorrect.

Figure 2 visually demonstrates the flow of the evaluation process. The LLM converts the legal question and related articles into Prolog code, which is verified by an independent Prolog engine, and the LLM's logical reasoning capabilities are assessed based on the obtained results. While we demonstrate the use of Prolog as the intermediate representation for logical reasoning, alternative choices for both the language and reasoning architecture can also be employed in the evaluation of LLM's logical reasoning capabilities. For instance, alternative logic programming languages such as Answer Set Programming (ASP) or functional languages like Haskell can be used to capture logical reasoning. Additionally, incorporating graph-based reasoning or neural-symbolic approaches could offer different perspectives on evaluating the LLM's reasoning skills. The key objective is to ensure a fair, unbiased, and decoupled evaluation process that accurately assesses the LLM's ability to provide precise logical reasoning irrespective of the choice of language or architecture. The hybrid approach mitigates concerns over subjectivity, brittleness, and human effort by leveraging the LLM's strengths in learning patterns from legal text corpus for automatic conversion to an intermediate logical representation, which adapts to changing regulations and case law, and by employing an independent reasoning engine for consistent, objective evaluation of logical reasoning capabilities.

4. LogiLaw Dataset

The LogiLaw dataset is developed to enhance the logical reasoning abilities of Large Language Models (LLMs) in the legal domain. By incorporating the generated Prolog code and corresponding verification results from the independent Prolog engine, this

dataset serves as a valuable resource for researchers and AI developers aiming to improve LLMs' performance in legal reasoning and other logic-intensive disciplines.

4.1. Dataset Purpose and Motivation

We did not find any existing datasets with adequate information for training LLMs in complex logical reasoning tasks, especially in fields such as law. The LogiLaw dataset seeks to bridge this gap by providing a comprehensive collection of legal questions, related articles, Prolog code representing the legal arguments, and the results obtained from verifying the Prolog code.

By utilizing the LogiLaw dataset, reinforcement learning algorithms can adapt the models to provide more accurate and reliable logical reasoning in the legal domain and other critical areas. This is paramount in applications where logical reasoning is essential for correct decision-making and compliance with established policies and regulations.

4.2. Details of Dataset Creation

The dataset creation process for LogiLaw involves several steps as follows:

1. The COLIEE dataset [3,4], is used as the base dataset. It is a collection of legal questions and related articles, providing a suitable foundation for generating Prolog code and corresponding verification results.
2. GPT-4 is employed to generate Prolog code from the legal questions and related articles in the COLIEE dataset. The generated code aims to capture the underlying logical reasoning required to answer the questions accurately.
3. GPT-3.5 is used to remove redundant details from the generated Prolog code. Although GPT-3.5 performs poorly in complex tasks, it is relatively competent in executing normalization tasks, making it suitable for this step.
4. The generated Prolog code is executed using a Prolog engine. This engine verifies the correctness and logical reasoning of the Prolog code by comparing its output against the ground truth.
5. The dataset combines all this information (i.e., legal questions, related articles, generated Prolog code, and verification results) to form the comprehensive LogiLaw dataset. This dataset then serves as a resource for training and evaluating LLMs in legal reasoning and related tasks.

4.3. Reinforcement Learning from Logical Feedback (RLLF)

The dataset plays a crucial role in the development of the RLLF approach. RLLF mimics the principles of Reinforcement Learning from Human Feedback (RLHF) [25,22], wherein the model receives rewards from human-generated feedback. However, RLLF focuses on logical feedback rather than human feedback to eliminate biases and promote accurate logical reasoning.

As illustrated in Figure 2, training a reward model with the LogiLaw dataset enables the model to receive rewards based on logical feedback from both the Prolog Engine and the Gold Evaluation instead of human judgments. This approach reduces the influence of subjectivity, uncertainty, and inconsistency present in human feedback. The combination of RLLF with the LogiLaw dataset thus contributes to a more efficient and reliable training process, resulting in LLMs with improved logical reasoning capabilities in critical domains like law.

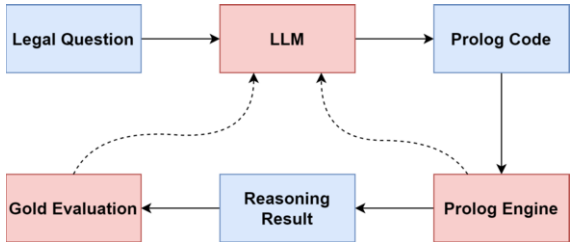

Figure 2. Architecture with feedback signals from both the Prolog Engine and Gold Evaluation used to improve the LLM, represented by dashed lines

5. Experiments and Results

To demonstrate the value of the LogiLaw dataset, we first need to evaluate the current performance of state-of-the-art models like GPT-4 on the COLIEE dataset using Prolog code. If GPT-4 could accurately answer each question in the COLIEE dataset using Prolog code, it would suggest no room for further optimization. However, our experimental results indicate otherwise.

We evaluate GPT-4's performance by generating Prolog code for the legal questions in the COLIEE dataset and running this code through a Prolog engine. The engine output allows us to determine whether the model provided correct answers or not.

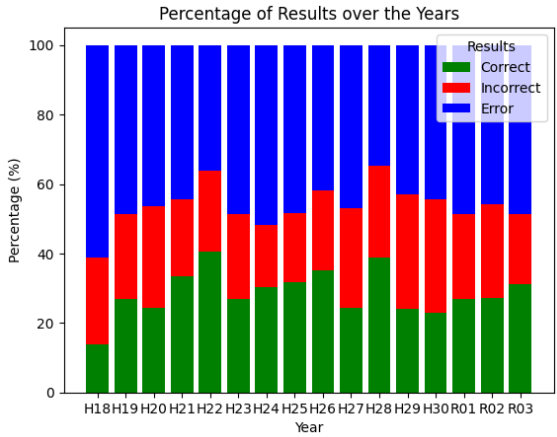

Figure 3. Percentage of correct, incorrect, and error outcomes across different years

Analysis of the plots in Figures 3 and 4 shows the following trends:

- The percentage of correct outcomes varies significantly in different years, ranging from as low as 13.9% (H18) to as high as 40.5% (H26).
- Incorrect outcomes have a somewhat stable distribution, fluctuating between 22.2% (H18) and 32.9% (H30), with minor exceptions.
- Error outcomes account for a large portion of the results, with some years having over 50% of error responses, such as H18 (61.1%) and R01 (48.6%).
- In terms of raw counts, R03 had the highest number of correct outcomes (34), followed by R01 (30) and H26 (26).

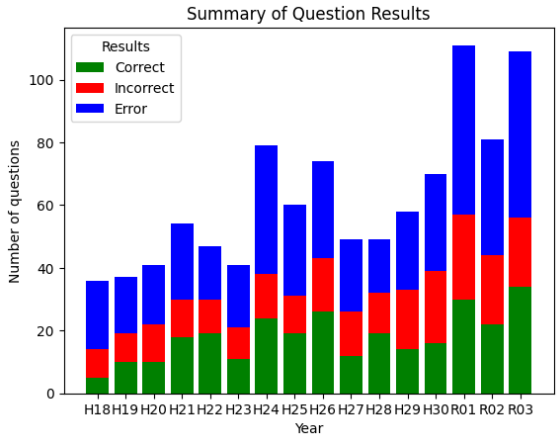

Figure 4. Raw counts of correct, incorrect, and error outcomes across different years

These results indicate substantial room for improving GPT-4's logical reasoning abilities in the legal domain. The correct outcome percentages suggest that GPT-4 is unable to accurately answer many questions in the COLIEE dataset, and a significant number of errors occur when interpreting Prolog code.

Utilizing the LogiLaw dataset for further research and model training provides an opportunity to fine-tune LLMs and enhance their logical reasoning capabilities in the legal domain. By leveraging the insights gained from these experiments, researchers can work towards addressing the weaknesses in state-of-the-art models like GPT-4 and develop more accurate and reliable reasoning systems in critical disciplines like law.

6. Discussion and Future Work

The results of our experiments on GPT-4 using the COLIEE dataset highlight the limitations of state-of-the-art models in logical reasoning in the legal domain. By proposing an improved evaluation method and introducing the LogiLaw dataset, we offer a promising direction for future research aimed at enhancing logical reasoning capabilities in LLMs.

Future work could focus on expanding the LogiLaw dataset to cover a broader range of legal questions and scenarios, thereby strengthening the models' exposure to different legal reasoning challenges. Additionally, investigating novel reinforcement learning techniques, such as Reinforcement Learning from Logical Feedback (RLLF), presents an opportunity to develop more effective training methods that prioritize logical insights over human feedback, reducing subjectivity and promoting accuracy. Also, logical feedback can be applied for requirement checking, including checking formal requirements by verifying with reasoning engines with requirements expressed as logical constraints, or non-formal requirements by interacting with humans as debugging-like dialogues.

With the Prolog code part, the LogiLaw dataset reveals several reasons that make GPT-4 produce incorrect answers, mostly involving lack of background knowledge. Another promising direction involves the integration of knowledge graphs and other external sources of structured information into the legal reasoning process. This approach

could enable LLMs to leverage a rich and diverse set of relationships and dependencies to further enhance their ability to reason about complex legal matters.

Finally, research on incorporating fairness, accountability, transparency, and explainability (FATE) principles into LLMs is crucial to foster trust and ensure responsible AI applications in legal settings. Future work in this direction will contribute to the broader deployment of AI-based legal reasoning systems in critical, real-world scenarios.

7. Conclusions

In conclusion, our work addresses the limitations in the logical reasoning capabilities of LLMs by proposing an improved evaluation method and introducing the LogiLaw dataset. By leveraging these resources, we aim to support research and development efforts for more accurate and reliable LLMs in the legal domain and other logic-intensive fields. Our study serves as a foundation for future work in advancing the performance of LLMs and promoting responsible AI applications in critical areas such as law.

Acknowledgements

This work was supported by the AIP challenge funding related with JST, AIP Trilateral AI Research, Grant Number JPMJCR20G4.

References

[1] Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, et al. Language models are few-shot learners. Advances in neural information processing systems. 2020;33:1877-901.

[2] OpenAI. GPT-4 Technical Report; 2023.

[3] Rabelo J, Kim MY, Goebel R, Yoshioka M, Kano Y, Satoh K. A Summary of the COLIEE 2019 Competition. In: New Frontiers in Artificial Intelligence. Online: Springer International Publishing; 2020. p. 34-49.

[4] Rabelo J, Goebel R, Kim MY, Kano Y, Yoshioka M, Satoh K. Overview and Discussion of the Competition on Legal Information Extraction/Entailment (COLIEE) 2021. The Review of Socionetwork Strategies. 2022 feb;16(1):111-33.

[5] Thanh NH, Quan BM, Nguyen C, Le T, Phuong NM, Binh DT, et al. A Summary of the ALQAC 2021 Competition. In: 2021 13th International Conference on Knowledge and Systems Engineering (KSE). Bangkok, Thailand: IEEE; 2021. p. 1-5.

[6] Nguyen C, Bui MQ, Do DT, Le NK, Nguyen DH, Nguyen TT, et al. ALQAC 2022: A Summary of the Competition. In: 2022 14th International Conference on Knowledge and Systems Engineering (KSE). Nha Trang, Vietnam: IEEE; 2022. p. 1-5.

[7] Kien PM, Nguyen HT, Bach NX, Tran V, Nguyen ML, Phuong TM. Answering Legal Questions by Learning Neural Attentive Text Representation. In: Proceedings of the 28th International Conference on Computational Linguistics. Barcelona, Spain (Online): International Committee on Computational Linguistics; 2020. p. 988-98. Available from: https://aclanthology.org/2020.coling-main.86.

[8] Asai A, Kasai J, Clark J, Lee K, Choi E, Hajishirzi H. XOR QA: Cross-lingual Open-Retrieval Question Answering. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Online: Association for Computational Linguistics; 2021. p. 547-64. Available from: https://aclanthology.org/2021.naacl-main.46.

[9] Nie Y, Wang S, Bansal M. Revealing the Importance of Semantic Retrieval for Machine Reading at Scale. 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). Hong Kong, China: Association for Computational Linguistics; 2019. p. 2553-66. Available from: https://aclanthology.org/D19-1258.

[10] Lee J, Yeung CY. Text Retrieval for Language Learners: Graded Vocabulary vs. Open Learner Model. In: Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021). Held Online: INCOMA Ltd.; 2021. p. 798-804. Available from: https://aclanthology.org/2021.ranlp-1.91.

[11] Krojer B, Adlakha V, Vineet V, Goyal Y, Ponti E, Reddy S. Image Retrieval from Contextual Descriptions. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Dublin, Ireland: Association for Computational Linguistics; 2022. p. 3426-40. Available from: https://aclanthology.org/2022.acl-long.241.

[12] Lin BY, Lee DH, Shen M, Moreno R, Huang X, Shiralkar P, et al. TriggerNER: Learning with Entity Triggers as Explanations for Named Entity Recognition. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online: Association for Computational Linguistics; 2020. p. 8503-11. Available from: https://aclanthology.org/2020.acl-main.752.

[13] Ritter A, Clark S, Mausam, Etzioni O. Named Entity Recognition in Tweets: An Experimental Study. In: Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing. Edinburgh, Scotland, UK.: Association for Computational Linguistics; 2011. p. 1524-34. Available from: https://aclanthology.org/D11-1141.

[14] Shao Y, Mao J, Liu Y, Ma W, Satoh K, Zhang M, et al. BERT-PLI: Modeling Paragraph-Level Interactions for Legal Case Retrieval. In: Bessiere C, editor. Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20. Yokohama, Japan: International Joint Conferences on Artificial Intelligence Organization; 2020. p. 3501-7. Main track. Available from: https://doi.org/10.24963/ijcai.2020/484.

[15] Nguyen HT, Nguyen MP, Vuong THY, Bui MQ, Nguyen MC, Dang TB, et al. Transformer-Based Approaches for Legal Text Processing. The Review of Socionetwork Strategies. 2022 jan;16(1):135-55.

[16] Louis A, Spanakis G. A Statutory Article Retrieval Dataset in French. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Dublin, Ireland: Association for Computational Linguistics; 2022. p. 6789-803. Available from: https://aclanthology.org/2022.acl-long.468.

[17] Vuong YTH, Bui QM, Nguyen HT, Nguyen TTT, Tran V, Phan XH, et al. SM-BERT-CR: a deep learning approach for case law retrieval with supporting model. Artificial Intelligence and Law. 2022 aug.

[18] Chalkidis I, Fergadiotis M, Malakasiotis P, Aletras N, Androutsopoulos I. LEGAL-BERT: The muppets straight out of law school. arXiv preprint arXiv:201002559. 2020.

[19] Nguyen HT, Vuong HYT, Nguyen PM, Dang BT, Bui QM, Vu ST, et al. Jnlp team: Deep learning for legal processing in coliee 2020. arXiv preprint arXiv:201108071. 2020.

[20] Nguyen HT, Nguyen PM, Vuong THY, Bui QM, Nguyen CM, Dang BT, et al. Jnlp team: Deep learning approaches for legal processing tasks in coliee 2021. arXiv preprint arXiv:210613405. 2021.

[21] Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I, et al. Language models are unsupervised multitask learners. OpenAI blog. 2019;1(8):9.

[22] Ouyang L, Wu J, Jiang X, Almeida D, Wainwright C, Mishkin P, et al. Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems. 2022;35:27730-44.

[23] Nguyen HT, Goebel R, Toni F, Stathis K, Satoh K. A negation detection assessment of GPTs: analysis with the xNot360 dataset. arXiv preprint arXiv:230616638. 2023.

[24] Nguyen HT, Goebel R, Toni F, Stathis K, Satoh K. How well do SOTA legal reasoning models support abductive reasoning? arXiv preprint arXiv:230406912. 2023.

[25] Ziegler DM, Stiennon N, Wu J, Brown TB, Radford A, Amodei D, et al. Fine-tuning language models from human preferences. arXiv preprint arXiv:190908593. 2019.