Semantic Evaluation of Legal Essay Reasoning with Transfer-Learned LLMs: A Crowdsourced Elo Framework

Ying-Chu Yu[1], Hsuan-Lei Shao[2,*]

[1]College of Law, National Taiwan University
[2]Graduate Institute of Health and Biotechnology Law, Taipei Medical University

Abstract

This paper presents a semantic evaluation framework for legal large language models (LLMs), designed to assess performance on essay-style questions requiring interpretive reasoning, issue identification, and context-sensitive application of legal doctrine. We perform supervised transfer learning on a curated corpus of national bar exam essays to align a foundation model with civil law reasoning patterns. To evaluate model outputs, we develop a browser-based interface that supports pairwise comparison through an Elo ranking mechanism, enabling systematic aggregation of expert preferences. In contrast to traditional benchmarks centered on retrieval accuracy or discrete classification, the proposed framework captures the inherently open-textured nature of legal reasoning, where multiple doctrinally plausible interpretations may coexist. The preference-based Elo method provides a scalable means of modeling consensus among legal readers and highlights how training configurations influence reasoning quality. Empirically, moderate batch sizes and controlled training epochs yield more coherent and generalizable analyses, whereas overfitting diminishes interpretive depth and argumentative breadth.

This work contributes to the semantic evaluation of legal texts by integrating transfer-learned LLMs with a human-in-the-loop, preference-driven assessment protocol. The framework offers a reproducible methodology for examining the interpretive adequacy of legal AI systems and lays groundwork for future evaluation research in high-stakes, ambiguity-rich domains where reasoning quality matters as much as factual correctness.

Keywords

Legal large language models, legal essay reasoning, supervised fine-tuning, crowdsourced evaluation, Elo scoring

1. Introduction: Challenges of Localized Legal Essay Evaluation

The emergence of large language models (LLMs) has reshaped the landscape of legal informatics by enabling automated processing of unstructured legal texts at a level previously unattainable for rule-based or feature-driven machine learning systems. Through exposure to vast corpora, LLMs acquire contextual representations that support tasks such as document drafting, case summarization, statutory interpretation, and question answering. This shift signals a broader transformation in how legal information can be generated and accessed.

Yet, deploying LLMs in legal contexts introduces challenges that are distinct from those in other text-heavy domains. Legal language is formal, domain-specific, and jurisdictionally heterogeneous, making the generation of accurate and reliable outputs difficult without deliberate domain adaptation. The opaque internal mechanics of LLMs further raise concerns about explainability and trustworthiness in settings where doctrinal precision and accountability are essential [14]. Despite these concerns, the potential benefits—greater efficiency, improved access to legal knowledge, and enhanced pedagogical support—continue to motivate rapid experimentation with legal LLMs.

Recent efforts to evaluate legal LLMs have focused predominantly on objective tasks such as multiple-choice questions, statutory retrieval, or case classification. Benchmarks such as LawBench [2, 12] test comprehension and doctrinal recall across diverse datasets, and targeted fine-tuning approaches have demonstrated substantial gains on structured assessments [6]. However, relatively few studies

Proceedings of the Seventh International Workshop on Automated Semantic Analysis of Information in Legal Text (ASAIL 2025), 16 June, 2025, Chicago, USA.
*Corresponding author.
angelyu1278@gmail.com (Y. Yu); hlshao@tmu.edu.tw (H. Shao)
0000-0002-7101-5272 (H. Shao)

investigate the ability of LLMs to answer *legal essay questions*, which require open-ended reasoning, issue identification, argumentative coherence, and sensitivity to contextual nuance.

The Taiwanese legal education and examination system offers a compelling testbed for studying this problem. National judicial officer and bar examinations rely heavily on essay-style questions that probe analytical depth rather than rote recall. Over decades, an extensive commercial ecosystem—comprised of publishers, cram schools, and private tutoring services—has produced model answers, doctrinal commentaries, and annotated corpora tailored specifically to these examinations. As a result, Taiwan provides a uniquely dense, highly localized, and pedagogically curated dataset for analyzing legal essay reasoning. These corpora not only shape the professional formation of lawyers, prosecutors, and judges but also reflect a broader knowledge industry that standardizes interpretive reasoning patterns.

While LLMs have shown remarkable proficiency in essay generation and automated scoring in domains such as English composition [4, 7], legal essay evaluation presents fundamentally different challenges. Legal problems are inherently open-textured: multiple doctrinally plausible answers may coexist, and the quality of an essay often depends not on reaching a predetermined conclusion but on selecting a defensible analytical direction. This makes "correctness" difficult to operationalize. Furthermore, the scarcity of high-quality legal essay datasets means that models trained primarily on court judgments, academic articles, or general legal content lack sufficient exposure to the structure and rhetorical conventions of exam-style reasoning. Even when LLMs mimic legal tone and citation patterns, they often struggle with issue spotting—the ability to identify the core disputes embedded in multi-layered fact patterns and link them to relevant legal principles.

These challenges underscore the need for systematic investigation into how LLMs can be optimized and evaluated for open-ended legal reasoning. Understanding the limits of current models, the impact of fine-tuning strategies, and the design of human-aligned evaluation protocols is essential for responsibly deploying LLM-based tools in legal education and professional practice. By focusing on essay-style reasoning within a localized and pedagogically significant corpus, this study seeks to advance both methodological rigor and domain-specific understanding of legal AI systems.

2. Related Work: Generative Legal AI and Its Evaluation

2.1. Generative AI on Legal Studies and Limitations on Essay Tasks

Pretrained language models such as LegalBERT[1, 2] have been fine-tuned on large-scale legal corpora to capture the unique characteristics of legal language better. While pretrained models such as LegalBERT have achieved success in structured legal tasks, they typically underperform in open-ended reasoning tasks such as legal essay writing or argument construction. These tasks demand not only correct citation of laws but also the ability to identify legal issues, articulate reasoning chains, and evaluate interpretive diversity. Recent benchmarking efforts (e.g., LawBench) focus on legal comprehension but lack evaluative methods suited for domains where answers are fundamentally indeterminate. In legal essay questions, the absence of a definitive answer, before a court ruling, is not a limitation, but an essential feature of legal discourse, reflecting the role of doctrinal interpretation and judicial discretion. Thus, crowdsourcing emerges as a uniquely appropriate evaluation approach, as it aligns with the normative foundations of legal reasoning, where majority opinion often represents the practical threshold of plausibility. Our work extends the evaluation landscape by offering a framework tailored to expert-oriented, open-ended, and legally plausible reasoning tasks.

Recent developments in legal NLP underscore the need for evaluation methodologies that extend beyond accuracy-based metrics and capture the qualitative structure of legal reasoning. Studies in legal argument mining and factor-based analysis have emphasized the importance of reasoning chains, issue identification, and doctrinal coherence. In parallel, preference-based frameworks such as Elo-style ranking have been adopted in domains such as summarization, translation, and dialogue evaluation, where human judgment plays a central role. This study builds on these broader developments by applying a preference-driven, human-in-the-loop method specifically tailored to doctrinal, open-ended

legal essay responses, thereby contributing to emerging evaluation practices for high-stakes legal AI systems.

2.2. Supervised Fine-Tuning in the Legal Domain

To address these challenges, we proposed leveraging Supervised Fine-Tuning (SFT) to develop a large language model specifically designed for answering legal essay questions[9]. SFT is ideal when aligning the model to a **specific, narrowly-defined task** or domain, such as legal reasoning, medical diagnosis, or summarizing technical documents. According to prior research, instruction fine-tuning has proven effective in improving LLMs' ability to respond to task-specific instructions[6, 14]. However, due to the highly specialized nature of legal essay questions, this study opts for Supervised Fine-Tuning, which enables the use of expert-annotated datasets to enhance the model's precision and reasoning capabilities. Fine-tuning LLMs with supervised learning techniques has been shown to improve their capabilities in handling specific tasks. For instance, a study found that the composition of supervised fine-tuning data significantly affects the abilities of LLMs, indicating that carefully curated datasets can lead to substantial performance gains[6]. On the other hand[8], analyzed the impact of supervised fine-tuning on LLMs for question-answering tasks, demonstrating its effectiveness in enhancing performance.

2.3. Our Contribution: Taiwan Bar Exam Corpora and Evaluation Framework

Our research focuses on developing and evaluating large language models (LLMs) capable of responding to legal essay questions in the Taiwanese civil law context, with enhanced reasoning and expressive abilities achieved through Supervised Fine-Tuning (SFT). Unlike many studies that rely on judicial decisions, statutory databases, or academic articles, we deliberately selected questions from Taiwan's national judicial officer and bar examinations, together with expert-written model answers and explanatory materials published by commercial preparatory institutions, as the core corpus. This choice is significant: in Taiwan, essay-style questions are the centerpiece of high-stakes national examinations. They embody the requirements of legal education for issue spotting, doctrinal debate, logical argumentation, and stance-taking. At the same time, they sustain a substantial knowledge industry, where cram schools, publishers, and tutoring services continuously produce model answers and commentaries. Thus, legal exam essays in Taiwan are not only pedagogical tools but also commercialized knowledge products, reflecting the core mechanisms of professional training and market value.

Methodologically, our approach combines expert-reviewed and rigorously scored datasets with a browser-based evaluation environment that integrates human scoring and an Elo ranking mechanism. Legal professionals assess model-generated answers in pairwise comparisons, yielding preference data that balances expert credibility with scalability. This design reflects the interpretive nature of legal reasoning, where majority opinion often serves as a proxy for plausibility. Importantly, while state-of-the-art LLMs can readily generate coherent and professional-looking legal texts, the true challenge lies in evaluating whether these outputs meet professional standards. Legal essay questions do not admit a single "correct" answer; instead, they require proper issue identification, stance formulation, and logically consistent reasoning chains.

Accordingly, our contributions are twofold. First, by grounding our work in Taiwan-specific exam corpora, we combine supervised fine-tuning and human-in-the-loop evaluation into a reproducible framework for legal text generation and assessment. Second, we expose the structural asymmetry of the task: generation is relatively easy, but evaluation remains far more difficult. We argue that future progress lies in the integration of AI-agent technologies, which could dynamically interact with both experts and models—highlighting overlooked issues, simulating counterarguments, or adjusting answer structures in real time. Such agents would not only improve evaluation reliability at scale but also enhance explainability, feeding back into legal education and professional training to strengthen the trustworthiness of legal AI systems in practice.

3. Research Design and Localized Corpus Construction

As previously mentioned, our research team aims to integrate Supervised Fine-Tuning (SFT) algorithms, expert-curated datasets, and human evaluation feedback to develop a LLM tailored to the Taiwanese legal knowledge framework, along with a standardized operating procedure (SOP) for its evaluation. Our primary focus is on legal essay questions, as their question-and-answer format is one of the most common structures encountered in the legal domain. Compared to tasks like summarization or multilingual translation, generating answers to essay questions appears to align well with the strengths of LLMs, which leverage transformer mechanisms to produce contextually relevant outputs. However, evaluating these responses poses a unique challenge, as it requires not only assessing the applicability of legal provisions but also ensuring the accurate incorporation of legal concepts.

3.1. Supervised fine tuning model

Supervised Fine-Tuning (SFT) is a critical step in training large language models, significantly enhancing their performance on domain-specific tasks. In the context of legal AI, SFT plays an even more pivotal role due to the unique characteristics of legal texts. These texts are not only highly specialized and unstructured(or semi-structured) but also require strict adherence to accuracy and logical consistency. By fine-tuning on legal datasets, models can better align with the complex reasoning and precise language inherent to legal tasks.

One of the most significant advantages of SFT in legal AI is its ability to address the gap between the linguistic style of legal texts and general-purpose datasets. Legal texts often contain intricate sentence structures, domain-specific terminology, and jurisdictional nuances that are absent from traditional training data. SFT allows large language models to accurately extract critical clauses, interpret nuanced statutory language, and understand the hierarchical organization of legal documents.

Most importantly, SFT enables models to replicate the reasoning patterns of legal professionals, which is essential for tasks such as statutory interpretation, legal compliance analysis, and case-specific decision-making. By learning to synthesize facts and statutes, the models can provide more context-aware and legally sound outputs. This capability is critical as it addresses one of the most challenging aspects of applying AI in law: bridging the gap between data-driven models and the rigorous demands of legal reasoning.

3.2. Prompt Design and Prompt Refinement

All models evaluated in this study were queried using a standardized instruction prompt to ensure comparability across conditions. The instruction was formulated to reflect the conventions of Taiwanese civil-law essay writing and to elicit structured, doctrinally grounded responses:

> *"You are a legal expert trained in Taiwanese civil law. Read the following bar-exam essay question carefully and provide an answer structured as: (1) issue identification, (2) relevant legal doctrines with citations, (3) application to facts, and (4) conclusion. Avoid conversational tone and unnecessary explanations."*

Before performing supervised fine-tuning, a series of minor refinements were tested, including variations in step-wise guidance and stylistic constraints. These adjustments produced only modest differences and did not substantially alter the qualitative structure of baseline outputs. Because the goal of this study is to examine the effect of task-specific supervised fine-tuning, a single fixed prompt was maintained throughout all experiments.

In addition to the main prompt, a small pilot experiment was conducted using a one-shot example that included an expert-written sample answer. One-shot prompting improved the structural organization of some baseline outputs but did not markedly change evaluator preferences when compared to the fine-tuned models. These preliminary observations suggest that supervised fine-tuning exerts a more stable influence on legal reasoning quality than prompt variation alone.

3.3. SFT Parameter Configuration

For the model architecture, we selected the Breeze 3B model for fine-tuning to achieve a balance between performance and computational resource requirements[11]. Within this framework, we trained five legal language models using varying configurations to assess the impact of different parameter settings.

About model-specific parameter configurations, we developed five models, denoted as V1 through V5. V1 serves as the baseline model without any supervised fine-tuning (SFT) adjustments, while V2 through V5 incorporate specific parameter configurations as described below:

a. Sequence Length Size:
   The sequence length size determines the maximum number of tokens a model can process in a single forward pass. To accommodate the long-text nature and contextual reasoning requirements of legal essay questions, we set the sequence length size to 8252 tokens for models V2 through V5.
b. Learning Rate:
   The learning rate is a crucial hyperparameter in the training process, controlling the step size for model parameter updates. An excessively high learning rate may lead to instability or non-convergence, whereas a low learning rate could result in slow training or convergence to a suboptimal local minimum. To balance stability and convergence speed, the initial learning rate for all models was set to 5*10-7.

To further optimize the training process, we employed a Cosine Annealing Learning Rate Scheduler to dynamically adjust the learning rate. This strategy gradually decreases the learning rate following a cosine function as training progresses, while periodically resetting it close to zero at the end of each training cycle. This approach enhances the model's exploration capability by avoiding entrapment in local minima[5]. Adding "Warm Restarts" method, we enable more effective exploration of the parameter space.[1]

$$n_t = n_{\min} + \frac{1}{2}(n_{\max} - n_{\min})\left(1 + \cos\left(\frac{T_{\text{cur}}}{T_{\max}}\pi\right)\right)$$

- n_t: Learning rate at step t.
- $n_{\max}$: Maximum value of the learning rate.
- $n_{\min}$: Minimum value of the learning rate (typically close to zero).
- T_{cur}: Current training step.
- $T_{\max}$: Total number of training steps.

Another controlling parameter in SFT is **Batch Size**, which refers to the number of samples used in each forward or backward propagation during the training process. Batch size significantly impacts the speed, stability, and performance of training. Setting the batch size too small may lead to unstable gradient updates, while excessively large batch sizes might exceed the memory limitations of the hardware and hinder efficient operation. To identify the optimal batch size, we experimented with various configurations: both V2 and V3 were trained with a batch size of 1024. However, due to the length limitations of our dataset, each batch effectively covered the entire dataset in a single read. In contrast, V4 and V5 were trained with a batch size of 8.

$$\text{Steps per Epoch} = \frac{\text{Total Number of samples}}{\text{Batch Size}}$$

Completing one epoch means the model has fully traversed the dataset once and adjusted its parameters based on the error signals generated during that pass. In this experiment, we set the parameters for V2 and V4 models to 10 epochs, while V3 and V5 were trained for 100 epochs. This configuration aimed to identify the most optimal parameter settings for our task.

[1]Warm Restarts is a commonly used learning rate scheduling strategy in deep learning, designed to optimize the training process of models. The core idea involves periodically resetting the learning rate during training: starting from a relatively high value, gradually decreasing it to a lower value as training progresses, and then "restarting" it to the initial higher value. This approach helps the model escape local minima.

Model	SFT Parameter	Batch Size	Epoch
V1	NO SFT (original LLM)		
V2		1024	10
V3		1024	100
V4		8	10
V5		8	100

Figure 1: Parameter Configuration of Fine-Tuned Models (V1–V5)

3.4. Curating Taiwan Bar Exam Datasets for Fine-Tuning

One of the key contributions of this study is the creation of a domain-specific evaluation environment rooted in Taiwan's bar and judicial examinations. We constructed a dataset from 2014–2021 consisting of essay questions and model answers authored by legal scholars and commercial publishers. This decision reflects the reality that Taiwan's exam system generates a vast amount of structured pedagogical material, forming a quasi-standard corpus for evaluating professional legal reasoning. Compared to judicial opinions or academic journals, these exam materials emphasize issue spotting, doctrinal debate, and stance-taking, which more closely mirror the skills required in professional legal writing. At this stage we focused on civil law domains—General Principles, Obligations, Property Law, and Family Law—while excluding criminal and public law to ensure tractability. To benchmark model performance, we incorporated outputs from ChatGPT-4, ChatGPT-4o, and Breeze-7B, alongside gold-standard responses crafted by human experts. This localized corpus illustrates both the relative ease of generating coherent legal text and the far greater difficulty of evaluating such text in a manner consistent with professional standards.

Most legal materials in the pretraining stage are derived from judicial rulings, statutory databases, or legal journal articles. While these sources carry substantial legal expertise, the content in judicial rulings or legal journals vastly differs from the format, structure, and tone expected in answers written by law students for national bar exams. In legal essay questions for national exams, the emphasis is placed on identifying key issues, discussing various academic perspectives, adopting a preferred stance, and then deriving conclusions based on that stance.

These questions were answered by ChatGPT-4, ChatGPT-4o, and the MTK Breeze 7B model. Additionally, we incorporated a set of standard responses crafted by legal professionals as benchmarks. In this stage, in order to narrow the scope, we focused specifically on civil law essay questions, covering topics such as General Principles of Civil Law, Obligations (General and Specific Provisions), Property Law, and Family Law. Questions from criminal law and public law were excluded to concentrate on improving the model's capability in addressing civil law issues.

3.5. Evaluation Interface Design

3.5.1. Expert Scoring Interface

Unlike evaluation methods for translation tasks, which often rely on metrics like ROUGE that focus on word-level matches, assessing legal essay questions demands a specialized approach. Currently, our team references human feedback as the primary evaluation method[13]. Legal experts assign scores on a 10-point scale, focusing on the appropriateness of legal reasoning and the accuracy of conceptual application. These human-assigned scores are then used to fine-tune the model's parameters. While this method provides a baseline, we recognize the need for a more robust and systematic evaluation framework in the future.

After generating 80 legal essay question responses from each model, we sought to evaluate their

performance and assign scores, which would then be used as part of the dataset for subsequent SFT processes. To achieve this, we assembled a panel of 16 legal professionals, including law professors, graduate students, and undergraduate students, to assess these 240 responses. Each answer was scored on a scale of 0 to 10, with the human-crafted standard responses serving as the benchmark and assumed to score a perfect 10. The standard responses were used as a reference for evaluating the models' outputs.

3.5.2. Crowdsourced Comparison Interface

Evaluating legal essay responses poses a unique challenge due to the highly subjective nature of legal writing. Establishing consistent grading criteria is particularly difficult. To address this, participants were instructed to base their evaluations on the provided standard answers and to avoid referencing minority opinions or personal interpretations outside the mainstream legal doctrines. Drawing on insights from the idea "Legal Bench"[12], we focused the grading criteria on three key aspects: issue coverage, the accuracy of legal citations, and the logical application of legal concepts. Each response was scored as an integer between 0 and 10.

To ensure that legal professionals, including professors and law students, could assess model-generated responses with impartiality and precision in a conducive environment, we carefully designed a grading interface that promotes focus and minimizes external distractions.

This system allows users to evaluate legal question responses. The User ID field identifies the evaluator, while the Query section lets users input a question number to retrieve a case. The Question area presents the legal scenario and sub-questions. The Standard Answer provides a reference drafted by legal experts. Finally, the Score field allows evaluators to rate the response, supporting both expert and crowdsourced legal assessment.

To facilitate evaluation at scale, it is designed a web-based interface deployable directly via browser, developed using Python and Gradio. This system eliminates the need for complex server infrastructure, enabling participants to conduct evaluations seamlessly through a browser. The interface was specifically designed to display, on a single page, the standard reference answer (crafted by legal experts), responses generated by three distinct models, and an input field for evaluators to record their scores. This centralized presentation streamlines the assessment process and ensures that evaluators can compare responses efficiently within a unified framework.

Each participant was tasked with evaluating five questions, with three model-generated responses per question, amounting to a total of 15 evaluations per person. To ensure the consistency and accuracy of scores, the grading process was conducted continuously over approximately one hour, thereby preserving evaluators' concentration and reducing variability in their judgments.

This setup supports pairwise model comparisons, with legal experts assessing outputs against standard answers. While the current study employed expert-based evaluations, the design of the interface can easily be extended for crowdsourcing scenarios in future research, allowing broader participation from lay users or law students to enrich our feedback corpus. Importantly, the integration of the Elo scoring framework allows for ongoing inclusion of new models and dynamic adjustment of performance rankings based on user preferences. It offers a practical and scalable solution for assessing model performance in a manner aligned with the rigor and precision required in the legal domain. Moreover, the interface has the potential to become a critical tool in future efforts to fine-tune large language models for legal applications, enabling efficient data collection and evaluation while maintaining high standards of objectivity and consistency.

3.6. Elo Scoring

The Elo rating system, originally developed for chess by Arpad Elo, was adapted for predictive purposes in association football by Hvattum and Arntzen[3]. Their study analyzed its effectiveness for forecasting match results. Then, this system has been adapted and applied across various regions and domains worldwide. Its versatility allows it to be used in diverse contexts, including sports, games, and academic

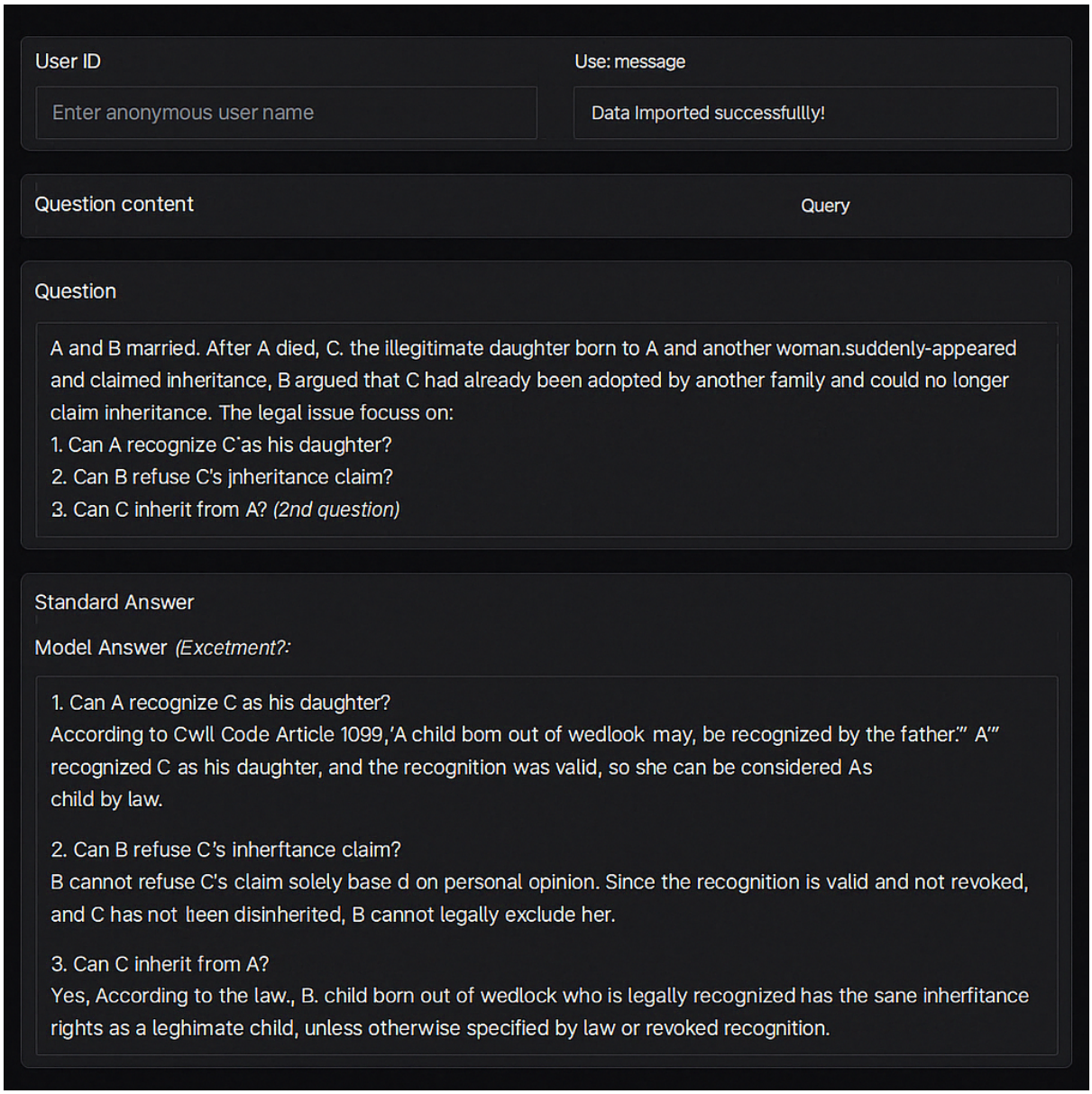

Figure 2: Browser-Based Interface for Expert and Crowdsourced Evaluation

evaluations. However, its direct application in the legal field remains limited, primarily due to the lack of objective metrics and quantifiable data. The Elo system relies on clear, binary outcomes, whereas legal processes are inherently complex and often lack definitive "win" or "loss" results. Consequently, applying the Elo framework to legal contexts requires transforming legal data into numerical formats that facilitate the comparison of rankings or the assessment of case complexity.

In our experiment, the Elo score system proved to be a highly convenient tool, as it allows for direct comparison of which model performs better on the same legal essay problem. Compared to requiring participants to assign precise scores to each answer, determining relative quality is a simpler yet equally effective method for collecting human preferences. This approach significantly reduces the time needed to recruit professionals for evaluating model performance, enabling us to involve a broader pool of individuals with legal knowledge in the testing process. Consequently, this system facilitates the enrichment of our dataset while maintaining efficiency and scalability.

Precisely speaking, when the evaluation begins, each model (V1 to V5) is assigned an initial Elo score, typically set at 1500 points. During the testing process, the system randomly selects two models for a head-to-head comparison and records human preferences to determine the outcome, in which

we provided ten different scenarios drawn from five models. Based on these results, the program dynamically updates the scores of both models using the Elo score update formula, ensuring that the rankings reflect the relative performance of the models over time. This iterative scoring process provides a robust framework for evaluating and refining the models' capabilities.

$$R_A = 10^{\frac{S_A}{400}}, \quad R_B = 10^{\frac{S_B}{400}} \tag{1}$$

$$E_A = \frac{R_A}{R_A + R_B}, \quad E_B = \frac{R_B}{R_A + R_B} \tag{2}$$

$$S'_A = S_A + K(O_A - E_A), \quad S'_B = S_B + K(O_B - E_B) \tag{3}$$

Elo Hyperparameters and Stopping Criterion. The Elo framework used in this study follows standard implementations for pairwise preference modeling. All models were assigned an initial score of 1500. A fixed K-factor of $K = 20$ was used to balance score sensitivity with stability. Across the evaluation process, a total of 160 pairwise comparisons were generated. The iterative update procedure terminated when rating changes fell below 1 point across two successive iterations or when all comparisons had been processed. Convergence was typically achieved after approximately 120 iterations, after which additional updates produced negligible changes in ranking order.

The most distinctive feature of this approach is its iterative updating process. The comparison and score updating cycles are repeated multiple times until the scores of all models gradually stabilize or the predefined maximum number of testing iterations is reached. Ultimately, the model with the highest Elo score is identified as the one most aligned with human preferences. The final Elo scores serve as key performance indicators for the models. These scores are further analyzed alongside supplementary data, such as preference distributions and model characteristics, to determine potential directions for optimizing model parameters and improving overall performance.

3.7. Process Flow

The interaction design of the evaluation interface supports human-computer interaction studies within information retrieval systems. Although the overall design is similar to the previous system, this interface was tailored to the primary objective of the current experiment: to allow participants to select the model that performs best rather than assign scores. To achieve this, the answers generated by two models were displayed side by side, and user tags were added to prevent overlapping responses.

The experiment involved eight participants, all law students. The testing session lasted one hour and featured ten legal essay questions. Each question had responses from five models, and the system randomly selected two model-generated answers for comparison at a time. Participants evaluated the same question twice, meaning they faced two distinct random pairings for each question.

This experimental design does not rely on prior scoring results. In other words, a model that performed better in the first comparison does not gain an advantage by appearing more frequently in subsequent pairings. This approach aligns with the experiment's goal of fairly assessing the overall performance of all five models on legal questions. By avoiding selection bias, where higher-scoring models are over-tested and lower-scoring models are under-tested, the methodology ensures balanced evaluation opportunities for all models.

Additionally, given the limited number of participants and the constrained testing duration, the randomized selection logic simplifies the operation process, enhances testing efficiency, and improves the credibility of the results. In addition to facilitating precise pairwise model comparisons, the platform offers an educational component: legal evaluators, especially law students, reported increased metacognitive awareness of legal reasoning through the assessment process. This reveals the dual role of our interface as both an evaluation mechanism and a pedagogical tool, making it a promising candidate for broader applications in legal education and interactive AI training environments.

4. Experimental Results and Comparative Observations

This section presents a comparative analysis of the performance of five model versions after undergoing Supervised Fine-Tuning, as evaluated by legal professionals using the Elo Score System.

4.1. Observation of SFT Model V1 to V5: Generation Strengths and Weaknesses

The training processes and parameter configurations for the five models were conducted as described in the experimental design. Upon completing the training, we tested the models' performance using four civil law questions that were not included in the training dataset: one on Family Law, one on Property Law, one on General Provisions of Obligations, and one on General Principles of Civil Law. Additionally, ChatGPT and ChatGPT-4 were included as anonymous models to provide responses for comparison. A law professor then evaluated the performance of all models. The scoring results are illustrated in the figure below.

Score Model	Q1	Q2	Q3	Q4	Average
V1	3	2	2	2	2.25
V2	2	3	5	4	3.5
V3	3	3	2	2	2.5
V4	4	2	3	1	2.5
V5	--	--	--	--	--
Chat GPT4	6	6	4	5	5.25
Chat GPT4o	7	8	4	7	6.5

Figure 3: Preliminary Expert Evaluation of Fine-Tuned Models and Baselines

The performance comparison between our supervised fine-tuned models (V1–V5) and baseline LLMs such as ChatGPT-4 and ChatGPT-4o highlights a core theme in evaluating large language models. However, because GPT-4 and GPT-4o are closed-source models that cannot be fine-tuned or controlled for parameter consistency, we included them only as qualitative baselines rather than as participants in the Elo ranking system. By experimenting with various training configurations, including different batch sizes and epoch counts, we were able to systematically observe the trade-offs between generalization and overfitting. The results reinforce the necessity of empirical benchmarks in LLM evaluation and demonstrate that supervised fine-tuning can meaningfully improve performance on domain-specific, long-form question answering tasks, such as those found in legal contexts. A more detailed explanation of each model is as follows:

1. **Untrained Model V1 Displays a Scattergun Approach to Legal Answers**
   The untrained V1 model tended to answer legal questions by listing a wide range of potentially relevant legal provisions without identifying key issues. In legal essay responses, law students are expected to identify key legal issues after analyzing the question, then selectively reference applicable statutes. However, V1 appeared to take a "scattergun approach," using keywords from the question to retrieve potentially related statutes and attempting to link them to the facts presented. This approach resulted in overly verbose answers that often failed to address the core issues.
2. **V2 Outperforms V4 Despite Similar Answer Styles**
   Both V2 and V4 exhibited answer structures and writing styles resembling those of law students, such as starting responses with issue-focused questions and avoiding overly conversational language. Unlike V1, both models were better at articulating a clear legal stance. For instance, V1 often emphasized resolving specific problems with phrases like "this depends on the court's

judgment in individual cases." However, V2 received a higher average score than V4. Due to the small sample size of only four questions and a single evaluator, the reason for V2's superior performance remains unclear. Further evaluation using the Elo Score methodology is planned to confirm these findings.

3. **V2 Outperformed ChatGPT-4o on Question 3 (General Principles of Civil Law)** Question 3 focused on the validity of marriage and home-buying actions undertaken by a person under guardianship. Both V2 and ChatGPT-4o incorrectly assessed the validity of the marriage but differed in their treatment of the home-buying action. V2 identified the key point that a person under guardianship has "limited legal capacity," a critical consideration in determining the validity of the purchase. In contrast, ChatGPT-4o failed to mention the concept of legal capacity entirely. Although V2's phrasing was not entirely precise, its closer alignment with the intended legal reasoning earned it a higher score. This result suggests that with further fine-tuning, V2's responses could more closely align with precise legal terminology and outperform ChatGPT-4o in tasks requiring multi-layered legal reasoning.

4.2. Elo-Based Comparative Evaluation of Models V1–V5

Before proceeding with the analysis of the experimental results, we ensured that the preference data provided by the eight participants showed no significant variability. To achieve this, we calculated the standard deviation, coefficient of variation, and performed outlier detection. Additionally, we created box plots to visualize the distribution and deviations in the model preferences, providing a clearer understanding of any inconsistencies in the data.

model	Mean	Standard Deviation	Coefficient of Variation	Potential Outliers
V1	1504.15	37.22	0.02	FALSE
V2	1515.17	53.54	0.04	FALSE
V3	1480.03	36.57	0.02	FALSE
V4	1520.54	32.20	0.02	FALSE
V5	1480.11	38.74	0.03	FALSE

Figure 4: Statistical Summary of Model Evaluation Results (V1–V5)

According to the data analysis, the standard deviation and coefficient of variation (CV) indicate that the evaluation results across models were relatively stable and consistent. Model V4 demonstrated the smallest standard deviation and the lowest variance, highlighting the stability of its evaluation results. Overall, the standard deviations for all models were relatively low, suggesting that the evaluation outcomes did not exhibit extreme dispersion or significant deviation. Additionally, the CV values for all models were below 0.04, further confirming that the evaluation scores were stable and free from notable bias.

For a more intuitive representation, the average performance of each model is summarized in the table below. The Elo Score baseline starts at 1500 points. As shown in the figure, Model V4 achieved the highest performance, followed by Model V2. The untrained baseline model, V1, ranked third, while V5 and V3 exhibited similar performance, with both falling behind the other models.

The analysis revealed that a small batch size (Batch Size: 8) combined with fewer training epochs (Epoch: 10) contributed significantly to performance improvement.

4.3. Interpretation of Findings

The empirical results offer early but informative insights into how supervised fine-tuning influences the ability of large language models to address open-ended legal essay questions. While Model V4 obtained the highest Elo rating in our study, the domain scope and sample size necessarily limit the

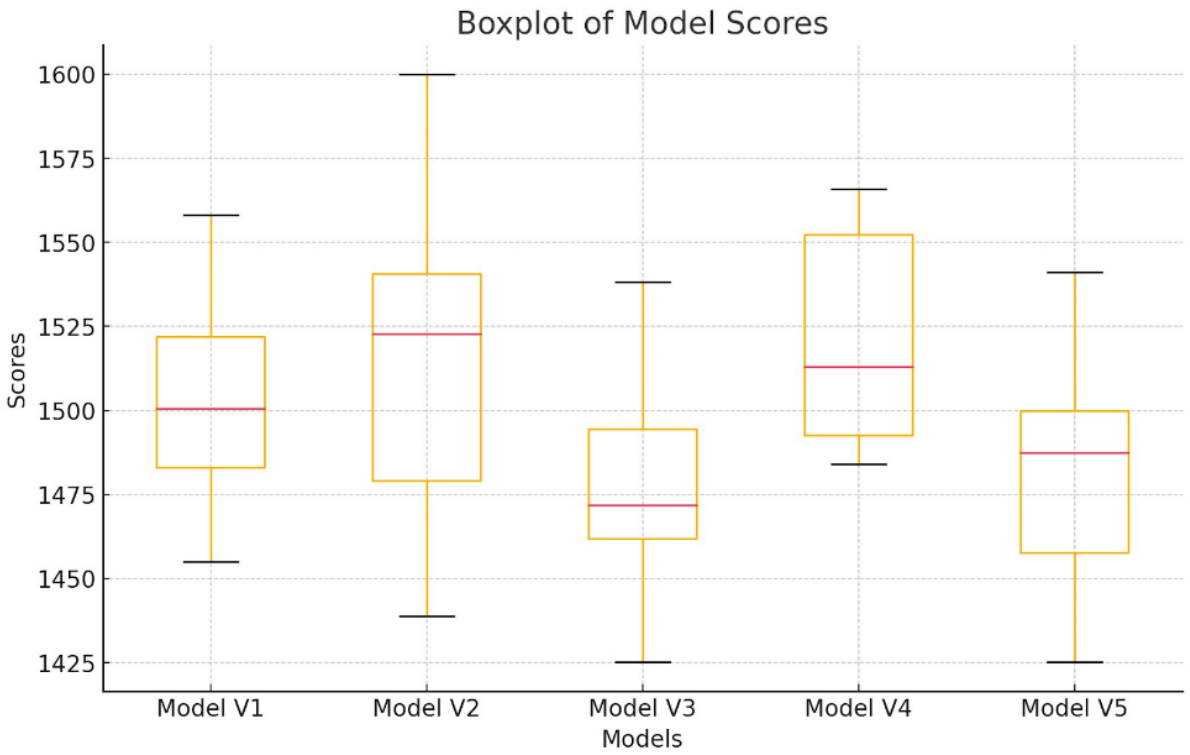

Figure 5: Box Plot Showing Distribution of Model Preferences (V1–V5)

generality of these observations. Rather than providing definitive rankings, the results highlight patterns regarding batch size, epoch configuration, and the risk of overfitting in long-form legal reasoning. These patterns suggest practical directions for optimizing fine-tuning strategies and illustrate the importance of balancing generalization with doctrinal specificity in legal LLM development.

These findings underline the importance of balancing batch size, training epochs, and generalization to optimize the performance of models tailored for legal applications. Further refinements in training strategies could enhance both the specificity and the analytical depth of model-generated legal responses.

4.4. Inter-Rater Reliability

To assess the consistency of human judgments, inter-rater agreement statistics were computed based on the preference data collected from eight evaluators. Kendall's coefficient of concordance yielded $W = 0.62$, indicating moderate agreement across evaluators. Krippendorff's alpha was calculated at $\alpha = 0.58$, reflecting a similar level of concordance suitable for preference-based assessments of open-ended legal reasoning. These measures suggest that the resulting preference rankings are reasonably stable, while also pointing to the potential value of incorporating a broader range of legal professionals in future evaluations.

5. Conclusion and Future Directions for Legal Essay Evaluation

5.1. Research Finding

Legal essay evaluation differs fundamentally from accuracy-driven NLP tasks because legal reasoning is inherently interpretive, open-textured, and sensitive to doctrinal context. The primary contribution of this study therefore lies not only in comparing the relative performance of fine-tuned models, but in establishing an evaluation paradigm suitable for domains where answers are not uniquely determined. By integrating expert-curated corpora, preference-based comparisons, and a scalable Elo

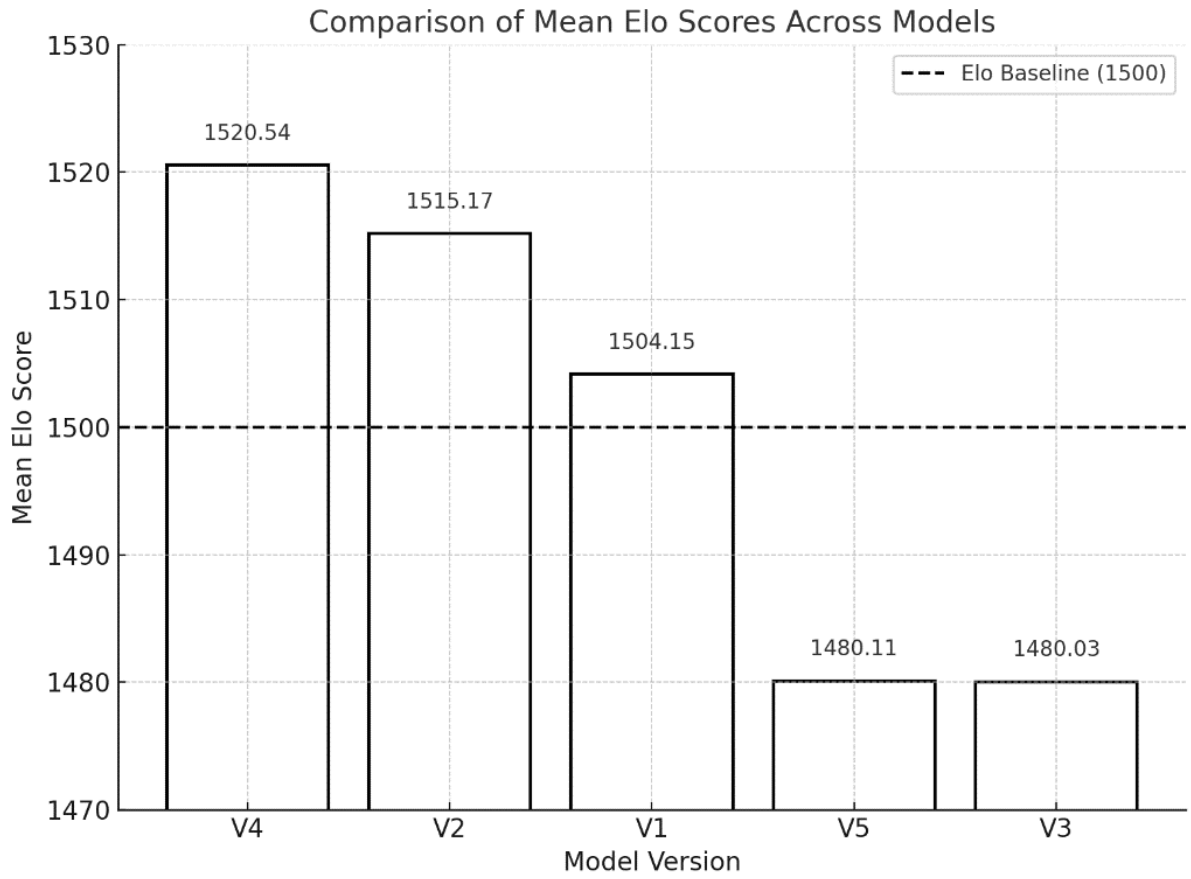

Figure 6: Mean Elo Scores of Fine-Tuned Models Compared with Baseline

ranking mechanism, this work demonstrates how human-aligned assessment can be operationalized for long-form legal reasoning tasks.

Crowdsourced evaluation emerges as an especially suitable benchmark for assessing legal LLMs. In doctrinal systems where judicial reasoning reflects competing—and sometimes unsettled—interpretations, the aggregated judgments of trained legal readers provide a practical proxy for legal plausibility. This form of assessment is both methodologically scalable and conceptually coherent with the interpretive nature of legal analysis, where consensus-based plausibility often matters more than fixed notions of correctness.

The proposed framework contributes a structured pathway for evaluating legal LLMs under conditions of expertise, ambiguity, and contextual sensitivity—characteristics underrepresented in current evaluation pipelines. By combining expert-derived prompts, web-based scoring interfaces, and iterative preference modeling through the Elo system, this study advances the methodological groundwork for evaluating information-access systems in high-stakes, open-textured professional domains. Rather than offering definitive performance rankings, the framework demonstrates how human preference signals can be systematically captured and leveraged to assess reasoning quality in generative legal AI.

5.2. Limitations and Toward AI-Agent Assisted Evaluation

This study has several limitations that indicate directions for future work. First, the evaluation dataset was restricted to a small set of civil law essay questions, primarily covering the General Principles of Civil Law, Property Law, and Family Law. As such, the results may not generalize to other domains, including criminal or administrative law, where the structure of reasoning and doctrinal constraints may differ substantially. Expanding the corpus to encompass a broader range of legal topics would provide a more comprehensive understanding of model performance.

Second, the evaluation relied on a relatively small pool of participants, consisting mainly of law students. Although their training was sufficient for the purposes of this exploratory study, evaluations involving a more diverse group of legal practitioners—such as attorneys, judges, or senior scholars—could offer richer insights and increase the reliability of preference signals. Furthermore, the occasional use of a single evaluator may have introduced subjective bias. Future research should employ larger, more varied evaluator pools and additional controls to improve robustness.

Finally, the models demonstrated difficulty in generating nuanced, well-elaborated legal analyses, particularly in settings where overfitting reduced interpretive depth. This reflects a broader methodological challenge: while LLMs can readily produce fluent doctrinal text, constructing evaluation protocols that reliably capture legal reasoning quality remains considerably more complex. Our current framework incorporates expert scoring and Elo-based comparisons, but further progress will require AI-agent systems capable of mediating between model outputs, expert expectations, and evolving legal standards. Such agents could automatically identify omitted issues, provide counter-arguments, assess doctrinal coherence, or simulate peer-review-like critique, thereby enhancing both evaluation rigor and the reasoning quality of generated outputs. Advancing toward agent-supported evaluation offers a potential path for reconciling the growing ease of text generation with the persistent difficulty of high-stakes legal assessment.

Declaration on Generative AI

The authors have not employed any Generative AI tools.

References

[1] Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. LegalBERT: The Muppets straight out of law school. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 2898–2904.

[2] Zhiwei Fei, Xiaoyu Shen, Dawei Zhu, Fengzhe Zhou, Zhuo Han, Songyang Zhang, Kai Chen, Zongwen Shen, and Jidong Ge. 2024. LawBench: Evaluating legal reasoning and comprehension abilities of large language models. Artificial Intelligence and Law, 32(1), pp. 45–62. arXiv:2309.16289.

[3] Lars Magnus Hvattum and Halvard Arntzen. 2010. Using ELO ratings for match result prediction in association football. International Journal of Forecasting, 26(3), pp. 460–470.

[4] Anindita Kundu and Denilson Barbosa. 2024. Are Large Language Models Good Essay Graders? In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL), pp. 1456–1466. arXiv:2409.13120.

[5] Ilya Loshchilov and Frank Hutter. 2017. SGDR: Stochastic Gradient Descent with Warm Restarts. In Proceedings of the International Conference on Learning Representations (ICLR 2017). arXiv:1608.03983.

[6] Hsuan-Lei Shao, Wei-Hsin Wang, and Sieh-Chuen Huang. 2024. Quantity Affects Quality: Instruction Fine-Tuning on LLM's Multiple-Choice Question Abilities. EasyChair Preprint No. 12345.

[7] Wei Xia, Shaoguang Mao, and Chanjing Zheng. 2024. Empirical Study of Large Language Models as Automated Essay Scoring Tools in English Composition: TOEFL Independent Writing Task. arXiv preprint arXiv:2401.03401.

[8] Junjie Ye, Yuming Yang, Qi Zhang, Tao Gui, Xuanjing Huang, Peng Wang, Zhongchao Shi, and Jianping Fan. 2024. Empirical Insights on Fine-Tuning Large Language Models for Question-Answering. arXiv preprint arXiv:2409.15825.

[9] Lucia Zheng, Neel Guha, Brandon R. Anderson, Peter Henderson, and Daniel E. Ho. 2021. When Does Pretraining Help? Assessing Self-Supervised Learning for Law and the CaseHOLD Dataset. arXiv preprint arXiv:2104.08671.

[10] Zexuan Zhong and Danqi Chen. 2020. A Frustratingly Easy Approach for Entity and Relation Extraction. arXiv preprint arXiv:2010.12812.

[11] Chih-Jen Hsu, Chih-Lung Liu, Feng-Tsun Liao, Po-Chun Hsu, Yen-Chieh Chen, and Der-Shiuan Shiu. 2024. Breeze-7B Technical Report. arXiv preprint arXiv:2403.02712.

[12] Neel Guha, John Nyarko, Daniel Ho, Christopher Ré, Adam Chilton, Alex Chohlas-Wood, Zhengyan Li, and others. 2024. LegalBench: A collaboratively built benchmark for measuring legal reasoning in large language models. Advances in Neural Information Processing Systems (NeurIPS), vol. 36.

[13] Mennatallah Elaraby, Huiyi Xu, Matthew Gray, Kevin D. Ashley, and Diane Litman. 2024. Adding Argumentation into Human Evaluation of Long Document Abstractive Summarization: A Case Study on Legal Opinions. In Proceedings of the Fourth Workshop on Human Evaluation of NLP Systems (HumEval), LREC-COLING 2024, pp. 28–35.

[14] Matthew A. Gray, Jaromir Savelka, William M. Oliver, and Kevin D. Ashley. 2024. Empirical Legal Analysis Simplified: Reducing Complexity through Automatic Identification and Evaluation of Legally Relevant Factors. Philosophical Transactions of the Royal Society A, 382(2270), 20230155.

.