Evaluating Large Language Models for Legal Multiple Choice Question and Answering

Elizabeth Lu, Rachel Ren, Neil Tripathi, Brynja Schultz
New York University
New York, USA
{evl5832,rr4000,nt2439,bes9992}@nyu.edu

Abstract

Natural Language Processing (NLP) models play a critical role in automating complex text-based tasks, including applications in specialized fields like legal informatics. This paper investigates the use of NLP for legal case analysis utilizing large language models (LLM) and evaluating them for their effectiveness in processing legal texts. Multiple evaluation metrics are used to assess the models' performance in accurately interpreting and answering these questions. In the domain-specific field, Legal Question Answering (LQA), correctness is especially valued as there are greater repercussions of incorrect choices. We ran three models, BERT-double, Legal-BERT, and Custom-legal BERT. Instead of using traditional metrics to measure how well each model performs on legal multiple choice questions and answering, we propose a novel metric that incorporates question difficulty, model confidence, and correctness to calculate a more comprehensive score. The models and code for our evaluation metrics are provided here: https://github.com/rachelren2025/NLP_Final_Project

Keywords

Natural Language Processing (NLP), Large Language Models (LLM), Artificial Intelligence (AI), Question Answer (QA), Legal Question Answering (LQA), Multiple-Choice Question Answering (MCQA), Evaluation Metrics

1 Introduction

In today's information-driven world, the legal field stands out as one of the most demanding when it comes to accessing and analyzing data. Whether it is legal professionals navigating through intricate case law, students studying to understand legal principles, or individuals researching a specific case, the common challenge lies in the sheer volume and complexity of legal information. Legal texts are often dense, unstructured, and interconnected, requiring significant time and effort to extract relevant insights. This process can be overwhelming, especially for those without extensive experience or resources. In general, the task of analyzing multiple precedents or legislative texts is essential yet time-consuming, leaving less room for strategic decision-making. Students, researchers, and professionals all face a similar struggle when trying to find specific information from vast legal corpora for academic or investigative purposes. Automating parts of this process using Natural Language Processing (NLP) has the potential to transform the field by providing faster, more accurate, and context-aware access to critical information. Given the widespread reliance on AI for question-answering and information retrieval in daily life, we would like to explore whether LLMs can address the unique challenges of the legal field. These models excel at tasks like text summarization, question answering, and contextual understanding. However, questions are raised about the abilities of AIs in the context of the legal domain.

This paper investigates the performance of three LLMs– BERT-double, Legal-BERT, and Custom-legal BERT–in answering multiple-choice legal questions, as a task that tests a model's ability to reason and extract contextually relevant information from complex legal texts. Whilst we don't explicitly use QA systems in their traditional sense, we analyze the models' ability to process natural language prompts, interpret context, and select the most accurate answer from the provided options. We use multiple performance metrics to measure the accuracy of these models in handling domain-specific, structured tasks, and for evaluating their applicability in specialized fields like law. In a high-stakes field like law, correctness is highly valued and can lead to catastrophic consequences if not accounted for. While the potential of LLMs in addressing the challenges of the legal field is

significant, their adoption is often hindered by concerns about their transparency and reliability. A key issue lies in the inherent black-box nature of these models, which complicates their acceptance in high-stakes applications like law.

1.1 The Black Box Problem

LLMs are often considered as black boxes because their decision-making processes are complex and not fully interpretable; we don't know how the algorithm gets to their solution (Brożek et al., 2023). These models consist of billions or trillions of parameters, interacting in non-linear ways with their training often lacking opacity as they are drawn from diverse sources and not fully disclosed to the public. For the users of these LLM models, it is difficult for us to trace specific patterns that lead to their output.

However, we will be focused on an output-based evaluation of using legal and non-legal LLMs as a multiple choice question and answering (MCQA) system in a legal context. First (1.1.1), we will begin by addressing why, in our specific case, the black box problem is relatively minuscule and why an output-based evaluation method is appropriate despite the inherent black box nature of LLMs. Finally (1.1.2), we will explain why precision is prioritized in this context.

1.1.1 The Black Box Problem is Minuscule in Our Case

Researchers in (Brożek et al., 2023) identified four interconnected aspects of the black box problem: opacity, strangeness, unpredictability, and justification. The opacity issue stems from our limited understanding of how LLMs detect patterns or arrive at their answers. While human thought processes are even more opaque, we interpret human behavior as rational and intentional, unlike AI decisions, which seem alien and difficult to trust (Brożek et al., 2023). The unpredictability problem reflects the tension between discomfort with unexpected outcomes and the purpose of AI, which generates insights beyond human capability, often unpredictably. This unpredictability makes AI decisions harder to accept, even when they outperform human judgment (Brożek et al., 2023). The justification problem arises from the need for rational, transparent decisions, especially in fields like law, where AI outputs often lack clarity and seem "mysterious." Addressing this requires AI systems to adopt ex-post justification—explaining decisions after the fact—to meet human standards of rationality and trust (Brożek et al., 2023).

Explainable AI (XAI) offers methods to clarify how AI systems make decisions, addressing aspects of the black box problem (Richmond, 2023). However, researchers in (Brożek et al., 2023) argue that the opacity problem is relatively minor compared to the complexity of the human mind. And in the context of a multiple-choice question-answering (MCQA) system, the black box nature of LLMs is less concerning because the focus lies on measurable outputs—whether the answers are correct or incorrect—rather than on interpreting the reasoning process. Output-based evaluation is the most practical approach for MCQA, as it emphasizes correctness over interpretability. The structured nature of MCQA reduces concerns about unpredictability or opacity, enabling performance to be judged solely on the accuracy of the answers.

1.1.2 Why We Value Accuracy More Than Interpretability

We value accuracy because high-stakes domains such as legal question and answering prioritize accuracy (Monroy et al., 2009; Vold and Conrad, 2021; Khazaeli et al., 2021; Martinez-Gil, 2023; Trautmann et al., 2024). Our MCQA approach aligns with findings that, in high-stakes applications like law, stakeholders prioritize accuracy over interpretability when the two are in conflict (Nussberger et al., 2023). In contexts like law or medicine, the primary concern is ensuring that decisions or outputs are correct and actionable, as inaccuracies can lead to significant real-world consequences. The (Nussberger et al., 2023) study emphasizes that while interpretability is valuable for building trust and transparency, it becomes secondary when precise and reliable outcomes are essential for decision-making. Additionally, researchers in (Gao et al., 2019) further emphasize this by stating that the accuracy of high-stakes applications is far more important than giving predictions of all query samples. This is particularly true in applications where the end-user relies on the AI system as a tool for achieving specific objectives, such as delivering a correct legal ruling or diagnosis, rather than understanding the underlying reasoning.

For MCQA systems in legal contexts, this prioritization of accuracy over interpretability underscores why output-based evaluation is ideal and we can consider the black box problem as minis-

cule in the scope of our project. By focusing on whether the system consistently delivers correct answers, the evaluation addresses the most critical stakeholder requirement—reliable results. While interpretability might enhance user trust in some scenarios, in high-stakes domains, the ultimate test of a system's utility lies in its ability to minimize errors and provide precise answers.

1.2 Research Question

These current metrics are not able to capture the nuanced reasoning and contextual understanding that law practitioners and researchers need. Hence the goal of our paper is to answer:

1. How do we design a new metric that prioritizes the correctness and reliability of the model's outputs?

2. What makes the metric we created better for evaluating LLM performance in answering legal multiple-choice questions?

2 Related Works

2.1 System Types

Question Answering (QA) Systems. QA systems can be said to be one of the most important research areas of Natural Language Processing (NLP), enabling machines to provide precise answers to user queries (Farea et al., 2022). A traditional question and answering system takes an input, usually given in the form of a sentence and produces an output, or an answer (Farea et al., 2022). The design of a QA system may vary, however, researchers from (Farea et al., 2022) have organized QA systems into a general framework of: 1. Question Answering Algorithms, 2. Knowledge Sources, 3. Question Types, and 4. Answer Types. We will revisit this framework at the methodology section where we discuss the intricacies of our design in answering our research questions.

Legal Question Answering (LQA) Systems. A subdivision of QA systems designed to process, analyze, and answer questions within the legal domain. Using NLP and machine learning models, these systems should understand and produce outputs to queries based on legal texts, case law, statutes, and other legal documents.

Multiple Choice Question Answering (MCQA) Systems. In the case of this paper, we are interested in legal-related QA systems, but in the form of a MCQA. These systems have to answer structured questions with predefined options and come with answer keys.

2.2 Metrics

A good evaluation metric measures the performance of a system effectively and aligns with the specific goals of the task. In the context of QA systems, particularly Multiple-Choice Question Answering (MCQA), a good metric should: reflect accuracy and relevance of predictions, consider the difficulty and importance of specific queries, and be scalable and interpretable for real-world applications.

2.2.1 Simple Untrained Automatic Evaluation Scores

Accuracy: A metric that measures the proportion of correctly predicted instances (both true positives and true negatives) out of the total number of predictions made by a model. This metric is simple and intuitive but fails to account for question difficulty or partial correctness.

F1: The harmonic mean of precision and recall, balances these two metrics and is particularly useful for imbalanced datasets, though it can be disproportionately affected by extreme values in either precision or recall.

Macro F1: verages F1 scores across categories equally, ensures equal treatment of all categories but may be skewed by poorly performing small categories (Evidently AI Team, 2024).

TF-IDF: Term Frequency-Inverse Document Frequency is a statistical method that is used to evaluate the importance of a word within a document within a larger corpus. It does so by balancing the word's frequency with how common it is across all documents.

Cosine Similarity: A metric that calculates the similarity between two vectors by measuring the cosine of the angle between them, effectively evaluating how closely aligned the vectors are in a multi-dimensional space.

Geometric Mean: A measure of central tendency that multiplies values and takes the n-th root, making it sensitive to low values. It is sensitive to outliers, a crucial feature that we leverage.

2.2.2 Machine-Trained Evaluation Scores: Bleurt and BertScore

BLEURT and BERTScore are machine-trained evaluation metrics (MTES) designed to measure semantic similarity between generated and reference

texts. BLEURT uses pre-trained transformer models fine-tuned on datasets with human-evaluated examples, enabling it to more accurately reflect human judgment and produce scores that reflect nuanced perceptions of text quality, making it particularly valuable for tasks requiring semantic equivalence. In contrast, BERTScore calculates token-level similarity by leveraging contextual embeddings from models like BERT, focusing on semantic overlap by aligning embeddings and computing cosine similarity to capture the meaning of the text beyond exact lexical matches. While both metrics aim to measure semantic similarity, BLEURT processes candidate and reference texts through a fine-tuned model to generate human-aligned quality scores, whereas BERTScore identifies the most similar tokens between the texts and computes precision, recall, and F1 scores based on these alignments, offering a more computationally efficient alternative (Farea et al. 2022).

BLEURT's strength lies in its fine-tuning on human-annotated data, enabling it to generate scores that closely reflect subjective quality judgments, making it particularly effective for tasks like machine translation and text summarization. However, this comes at a computational cost due to its large model size and significant resource requirements, and its performance may vary outside its training domain. In contrast, BERTScore is lightweight and efficient, making it suitable for tasks requiring rapid semantic evaluation, such as paraphrase detection, summarization, translation, and dialogue generation. While it excels in capturing semantic overlap through embedding-based similarity, BERTScore can overlook document-level coherence and is sensitive to high-frequency tokens, which may distort similarity scores. For evaluating legal text generation, BLEURT's alignment with human judgment is particularly valuable for assessing the semantic accuracy of legal arguments, whereas BERTScore provides a reliable measure of how closely generated legal statements align with the intended meaning of reference statements. Together, these metrics offer a robust framework for evaluating the quality of legal text, tailored to the specific requirements of this domain (Farea et al. 2022).

2.3 Models

In the context of legal question-answering (LQA) systems, various adaptations of the BERT model have been explored to improve performance in domain-specific tasks. For more information about BERT models, we referred to this paper (https://aclanthology.org/N19-1423.pdf). For instance, BERT-Double extends the capabilities of the base BERT architecture by incorporating dual-layer processing, enhancing comparative reasoning and contextual understanding. In the paper provided by (Zheng et al., 2021) BERT-Double is implemented as a variation of the base BERT model (uncased, 110 million parameters). While the original BERT base model was pretrained on the English Wikipedia corpus (general domain corpus) for 1 million steps, BERT-Double extends this pretraining by an additional 1 million steps, resulting in a model pre-trained for a total of 2 million steps (Zheng et al., 2021).

Legal-BERT is a domain-specific adaptation of the base BERT model, pretrained on a comprehensive corpus of U.S. case law sourced from the Harvard Law case database (https://case.law/) (Zheng et al., 2021). This adaptation is specifically designed to address the unique linguistic and semantic challenges inherent in the legal domain, making it better suited for legal Natural Language Processing (NLP) tasks such as legal question answering, case retrieval, and text classification (Zheng et al., 2021).

Additionally, to further refine domain-specific performance, Custom-Legal BERT introduces additional fine-tuning using specialized datasets, such as CaseHOLD, integrating task-specific features like question difficulty and model confidence (Zheng et al., 2021).

These models form the foundation of our study, offering a comparative lens to evaluate general-purpose versus domain-specific approaches to answering legal multiple-choice questions.

Domain-specific pretraining has been extensively explored as a means of enhancing the performance of language models on specialized tasks. Studies on Legal-BERT and Custom Legal-BERT demonstrate the value of adapting pre-trained models to the legal domain using corpora such as the Harvard Law case database (Zheng et al., 2021). For example, CaseHOLD, a highly domain-specific and linguistically complex task, shows substantial performance gains when using Legal-BERT variants compared to general-purpose models like BERT-Double. Custom Legal-BERT, pre-trained from scratch with a legal-specific vocabulary, achieved the highest macro F1 score (69.5%), underscoring the importance of tailoring vocabu-

laries and pretraining processes to domain-specific text. Additionally, comparisons across tasks of varying complexity—Overruling, Terms of Service, and CaseHOLD—highlight how the relevance of domain adaptation increases with task specificity (Zheng et al., 2021). These findings align with prior research indicating that pre-trained models benefit significantly from additional pre training steps, as observed with BERT-Double, and suggest that robust input transformations and rigorous evaluation protocols, such as 10-fold cross-validation, are essential for adapting models to high-stakes domains like law (Zheng et al., 2021).

2.4 Dataset: CaseHOLD

The CaseHOLD dataset, derived from Harvard Law's case law data, comprises over 53,000 multiple-choice questions designed for legal reasoning and summarization tasks. Each question includes one prompt, one correct holding statement and four distractors.The correct holding statement is derived from citations in judicial rulings, specifically from parenthetical text starting with the word "holding." The text preceding the citation serves as the citing text prompt, with a placeholder token <HOLDING> inserted where the holding statement was extracted. To ensure task difficulty, the distractors are selected using TF-IDF similarity between the correct answer and a pool of other holding statements. A similarity threshold of 0.75 is applied to avoid indistinguishable distractors, ensuring the need for deeper legal reasoning beyond simple keyword matching to select the correct statement. The dataset provides a task that evaluates the model's ability to: (1) understand the citing text prompt and (2) Differentiate the correct holding statement from semantically similar distractors. This approach ensures a challenging evaluation framework by emphasizing nuanced legal reasoning over random guessing, making CaseHOLD a benchmark for assessing model performance in handling complex legal texts (Zheng et al., 2021).

3 Methodology

As outlined in the literature review section I. on QA systems, the framework of our specific QA system can be categorized according to the concept proposed by (Farea, et al., 2021), which is structured as follows:

1. Question Answering Algorithm: Our algorithm is a neural based algorithm as we are using BERT.

2. Knowledge Source: The system is operating in a closed domain, as the dataset (CaseHOLD) provides a specific and finite scope of legal knowledge. It is trained and evaluated within the legal domain, not using general or open-ended knowledge from the broader internet.

3. Question Type: Our system uses multiple-choice questions and answers.

4. Answer Type: The answer type is agnostic, as the system selects from predefined multiple-choice options without needing to generate text or extract specific spans.

We aim to address our research question using the CaseHOLD dataset. The dataset consists of a message and five holdings as a potential answer. We processed the dataset through three versions of BERT, generating an output file for each model. Each model produces a predicted answer and a probability score indicating the likelihood of a given holding being correct.

To evaluate the models, we used general metrics such as accuracy, evaluation loss, and macro F1 score. Subsequently, we designed a new evaluation metric and re-evaluated the models to assess its effectiveness. The specifics of our research methodology and the rationale behind this approach are outlined below.

3.1 Model Setup

3.1.1 Input and Outputs

The inputs for the task were adapted to fit the architecture of the BERT-based models by transforming the multiple-choice question into five separate prompt-answer pairs, each structured as [CLS] Prompt [SEP] Answer [SEP]. Here, the prompt represents the citing text from the judicial decision and the answer represents one of the five holding statements. The input pairs are tokenized using a BERT-compatible tokenizer, which splits text into subword units and maps them to numerical IDs. A maximum sequence length of 128 tokens is enforced, with sequences longer than this limit truncated and shorter ones padded. Each of the five tokenized pairs is passed through the model independently. The model outputs logits for each pair, representing the likelihood of each holding statement being the correct answer. The logits for all five

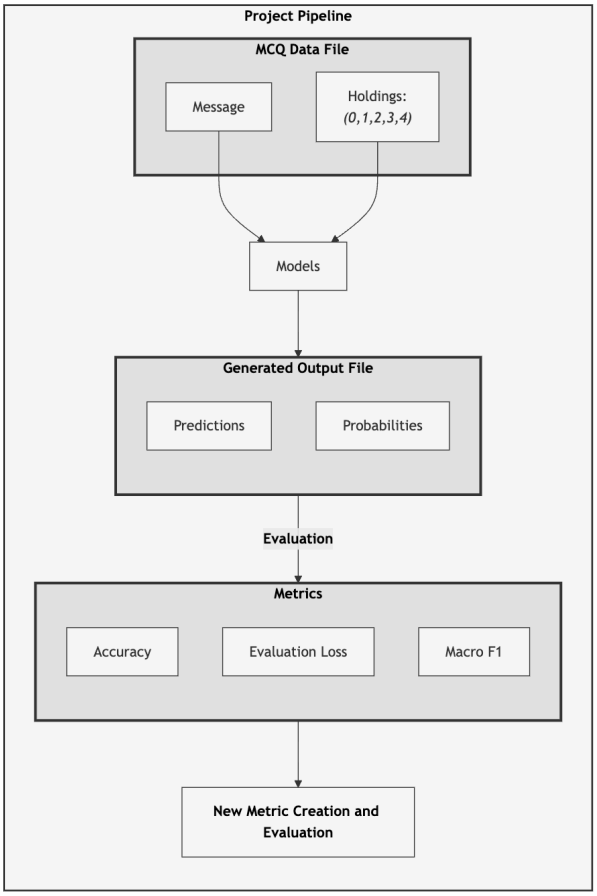

Figure 1: We processed our dataset using three models: BERT-Double, LEGAL-BERT, and Custom Legal-BERT. Each model produced an output file containing predictions and probabilities. The results were evaluated using three standard metrics: Accuracy, Evaluation Loss, and Macro F1. Additionally, we developed a new metric to further evaluate the data

pairs are concatenated, and a softmax function is applied to normalize these logits into probabilities, which are then used to determine the model's predicted answer. The outputs of this process include two files. The first, "predictions.csv," stores the final predictions, where each row corresponds to the predicted answer for a question. The second, "probabilities.csv," contains the softmax-normalized logits for all five answer choices per question, representing the probability distribution over the choices. This approach minimized changes to the original architecture while fully utilizing the strengths of the fine-tuned models (Zheng et al., 2021).

3.1.2 Fine-Tuning the Models

We fine-tuned three BERT-based models—BERT-Double, Legal-BERT, and Custom Legal-BERT—on the CaseHOLD dataset by updating each model's weights on task-specific data. We used the predefined hyperparameters established in the original CaseHOLD paper. Specifically, all models were trained with a learning rate of 5e-6, a batch size of 16, and a maximum sequence length of 128. The training process was conducted for 3 epochs using the AdamW optimizer with mixed precision (fp16) to balance computational efficiency and performance. Dataset splits followed an 80/20 train-test ratio, and performance was monitored every 1000 steps using step-based evaluation.

3.2 Running the Models

The models were fine-tuned and run using the following command:

```
python multiple_choice/run_multiple_choice.py \
    --task_name casehold \
    --model_name_or_path MODEL_NAME \
    --data_dir data \
    --do_train --do_eval --do_predict \
    --evaluation_strategy steps \
    --max_seq_length 128 \
    --per_device_train_batch_size 16 \
    --learning_rate 5e-6 \
    --num_train_epochs 3 \
    --output_dir output \
    --overwrite_output_dir \
    --logging_steps 1000 \
```

```
--fp16
```

Here MODEL_NAME would be replaced with the desired model's hugging face repository path (zlucia/bert-double, zlucia/legalbert, or custom-legalbert). After fine tuning, output files, *predictions.csv* and *probabilities.csv*, were saved in the corresponding output directory to the model run.

3.3 Evaluation Code

We assessed the fine-tuned models' performance using a couple metrics.

Evaluation Loss (eval_loss)

The evaluation loss represents the average cross-entropy loss computed over the evaluation dataset. It measures the divergence between the model's predicted probability distribution and the true labels. Lower loss values indicate better alignment between predictions and ground truth labels. The evaluation loss is calculated for each sample using the normalized logits from probabilities.csv, and the average loss across all evaluation samples is reported. Custom-legal-BERT achieved the lowest evaluation loss, indicating that it produced the most well-aligned predictions among the models.

Macro-F1 Score (eval_f1)

The Macro-F1 score is the unweighted average of F1 scores for each class. It accounts for both precision (accuracy of predictions) and recall (coverage of true labels), treating all classes equally regardless of their frequency. We calculated the F1 scores for each class and then averaged them to get the macro-F1 score. Custom Legal-BERT achieved the highest macro-F1 score, demonstrating a better balance of precision and recall, indicating that the model generalized well across all answer choices.

Accuracy (eval_accuracy)

We use the accuracy_score function from sklearn.metrics to calculate the accuracy score for each model and for the confidence filtered accuracy. Legal-BERT achieved the highest accuracy, indicating that it performed well in identifying correct answers.

3.4 New Metric: Weighted Correctness Score

To evaluate the model performance more rigorously, we propose a weighted correctness score that accounts for three factors: (1) Question Difficulty, (2) Model Confidence, and (3) Correctness. This new metric (WCS) rewards correct answers more for difficult questions and high confidence, while penalizing incorrect answers based on how easy the question is and how confident the model is on the wrong answer.

3.4.1 Difficulty

To calculate the difficulty of a question, we measure the semantic similarity between all answer choices using **BERTScore**. For a question i with answer choices $A_1, A_2, \ldots, A_5$, we compute pairwise similarity scores for all unique pairs:

$$\text{BERTScore}_{p,q} = \text{BERTScore}(A_p, A_q), \quad \forall p \neq q. \tag{1}$$

We then aggregate the pairwise scores using the **geometric mean** to obtain the difficulty score D_i:

$$D_i = \left(\prod_{p \neq q} \text{BERTScore}_{p,q} \right)^{\frac{1}{N}}, \tag{2}$$

where N is the number of unique pairs.

Higher values of D_i indicate greater similarity among answer choices, making the question more difficult to answer correctly. We adjust the score based on both the model's correctness and the difficulty quartile q, where $q1$ is the first quartile, representing the easiest questions, and $q4$ represents the hardest questions. Quartile weights are assigned as follows:

$$w_q = \begin{cases} 0.25, & \text{if correct and } q = q_1, \\ 0.50, & \text{if correct and } q = q_2, \\ 0.75, & \text{if correct and } q = q_3, \\ 1.00, & \text{if correct and } q = q_4, \\ 1.00, & \text{if incorrect and } q = q_1, \\ 0.75, & \text{if incorrect and } q = q_2, \\ 0.50, & \text{if incorrect and } q = q_3, \\ 0.25, & \text{if incorrect and } q = q_4. \end{cases} \tag{3}$$

3.4.2 Confidence

To integrate model confidence into the final score, we define the confidence weight as the maximum probability from the softmax output. The softmax function ensures that the output probabilities sum to 1, forming a valid probability distribution and reflecting the model's certainty in its predictions. A higher value of confidence weight indicates that the model is more confident in its selected prediction.

The softmax function works for confidence because it converts the model's raw logits into probabilities that sum to 1, ensuring the output can be

Model	Eval_loss	Macro-F1	Accuracy	WCS
BERT-Double	0.9361	0.6254	0.6257	-0.44
Legal-BERT	1.1838	0.5176	0.6822	-0.48
Custom Legal-BERT	0.7841	0.6819	0.5177	-0.36

Table 1: Evaluation results of BERT-Double, Legal-BERT, and Custom Legal-BERT models

interpreted as the model's confidence across all classes. The maximum softmax probability represents the model's degree of certainty for its predicted class, where a higher value indicates greater confidence in the prediction.

3.4.3 New Metric Formula

The final metric M is computed as:

$$M = \pm \max(\text{softmax\_probabilities}) \cdot w_q, \quad (4)$$

where the sign is $+$ if the prediction is correct and $-$ if it is incorrect.

Interpretation:

- For **correct predictions**, the model is rewarded proportionally to its confidence and difficulty. Easier questions (low quartile) receive smaller rewards.
- For **incorrect predictions**, the model is penalized more heavily for easier questions, reflecting the expectation that the model should succeed on them.

3.5 Limitations

While our methodology offers significant benefits for evaluating model performance, there are some limitations. These limitations stem from fine-tuning processes, dataset design, and the interplay between question difficulty and semantic similarity.

3.5.1 Catastrophic Forgetting

Catastrophic forgetting refers to a model's tendency to lose previously learned knowledge when fine-tuned on a specific task. In our case, this issue is particularly prevalent due to several factors. Fine-tuning adapts a pre-trained model to a domain-specific dataset, but it risks overwriting general knowledge learned during pre-training. Training over three epochs reinforces the patterns within the fine-tuning data, improving task performance while amplifying the risk of overwriting pre-trained knowledge and causing a decline in general capabilities. Additionally, in our implementation, fine-tuning was applied to all layers of the model, not just the top layers. While this enables deeper adaptation to the legal domain, it also increases the likelihood of catastrophic forgetting, as adjustments across all layers can significantly alter the pre-trained representations. Finally, as the model becomes specialized for legal tasks, its performance on broader or unrelated tasks may deteriorate. These factors combined increase the likelihood of catastrophic forgetting, which can affect the model's ability to generalize beyond the fine-tuning dataset. Addressing this requires careful calibration of training epochs and the incorporation of regularization techniques or knowledge-preserving strategies (Kotha et al., 2024; Lesort et al., 2023).

3.5.2 Dataset Design and Question Difficulty

The dataset we use is constructed to include the correct answer choice and the most semantically similar distractors. Specifically, TF-IDF cosine similarity with a threshold of 0.75 to select holdings that are closest in meaning to the correct answer. In our new metric, we calculate semantic similarity scores using BERTScore to quantify question difficulty. By comparing the embeddings of answer choices, BERTScore provides a nuanced measure of how closely related the distractors are to the correct answer. Questions with higher similarity scores across all answer choices are treated as more difficult, as they require the model to distinguish between highly similar options.

However, this approach raises a potential limitation when determining question difficulty. Since the distractors are explicitly chosen to be semantically similar, the difficulty of a question is inherently influenced by the dataset design. Questions may appear artificially difficult due to the high semantic overlap, rather than the complexity or ambiguity of the legal reasoning required. This poses challenges for accurately distinguishing between genuinely difficult questions and those made challenging by the selection process itself. To mitigate this issue, future iterations could incorporate different measures for question difficulty such as changing how the dataset's distractors are chosen to create questions of different difficulty levels.

3.5.3 Softmax Inability to capture Epistemic Uncertainty

The softmax function, commonly used in classification tasks to convert logits into probabilities, has inherent limitations in representing epistemic uncertainty—the uncertainty arising from model parameters. Softmax outputs can be overconfident, especially when the model encounters inputs that differ from the training distribution. This overconfidence is problematic in legal applications, where understanding the model's uncertainty is crucial for interpreting its predictions. Recent studies have highlighted this limitation. For instance, (Pearce et al. 2021) discuss how neural networks often fail to increase their uncertainty when predicting on data far from the training distribution, leading to overconfident predictions (https://arxiv.org/pdf/2106.04972). This limitation suggests the need for incorporating advanced uncertainty estimation techniques, such as Bayesian neural networks or ensemble methods, to better capture epistemic uncertainty and enhance the reliability of model predictions in the legal domain.

4 Results

Please refer to Table 1 for the scores of each metric. Under our proposed metric, Custom Legal-BERT achieved the highest score, followed by BERT-Double and Legal-BERT, which ranked second and third, respectively.

5 Evaluation

Based on the results from evaluating these models on the CaseHOLD dataset, there is strong evidence that our proposed metric offers a significant advantage over traditional metrics like accuracy by incorporating question difficulty and confidence calibration:

Nuanced Differentiation: Traditional accuracy treats all questions equally, while our metric recognizes the hardness of questions and penalizes confident errors on easy questions more heavily. This ensures the system is more robust and nuanced in its evaluation.

Robustness: Custom Legal-BERT outperformed other models (refer to Table 1), demonstrating its ability to handle harder questions while minimizing overconfident errors. Our metric aligns with existing measures in best-case scenarios (correct, confident answers) but disagrees in worst cases, particularly where models overconfidently predict incorrect answers.

For instance, in (Zheng et al., 2021), the same three models—Custom Legal-BERT, BERT-Double, and Legal-BERT—were evaluated. Under Macro-F1, Custom Legal-BERT performed the best. However, while BERT-Double was their worst-performing model, Legal-BERT was identified as the worst under our metric.

This highlights the robustness of our metric: it agrees on strong performances but identifies weaknesses missed by traditional measures, focusing on model calibration and performance on harder questions.

Practical Reliability: By penalizing confident mistakes on easy questions more heavily, our metric better aligns with real-world expectations, particularly in high-stakes legal settings. **Notably, all scores are negative,** reflecting the metric's strictness in identifying errors. This emphasizes the fact that current models are far from perfectly accurate—often struggling with overconfidence—and mistakes on easier questions are penalized most severely. Such a rigorous evaluation is essential for legal applications, where even minor errors can carry significant consequences, and reliable performance must be prioritized.

6 Future Works

While our proposed metric demonstrates promising results, time constraints limited the scope of our experiments. Future work will focus on:

1. Robustness Testing: Extending evaluations to other legal datasets to assess generalizability across domains and question types.

2. Adversarial Evaluation: Investigating model performance under input perturbations to explore resilience and robustness further.

3. Detailed Statistical Analysis: Analyzing the distribution of question difficulty and its impact on model performance.

4. Model Variations: Testing with larger and more diverse pretrained models to validate the metric's applicability.

7 Conclusion

7.1 Answering Research Questions

We posed two research questions at the beginning of the paper, and here we provide the answers.

To design a new metric that prioritizes correctness and reliability, we integrated correctness, confidence, and difficulty: correct predictions are rewarded, incorrect predictions are penalized, high-confidence correct answers are favored over low-confidence ones, and harder questions—measured using BERTScore similarity—contribute more to the final score. Our metric outperforms traditional measures like accuracy and evaluation loss by penalizing overconfident errors on easier questions, rewarding calibrated confidence on harder ones, and accounting for question complexity, ensuring a more reliable and nuanced evaluation framework for legal multiple-choice tasks.

7.2 Final Thoughts

Our work introduces a novel evaluation metric that integrates difficulty, confidence, and correctness, providing a more nuanced evaluation system for legal question-answering models. Unlike traditional metrics such as accuracy, our approach penalizes confident errors on easy questions more heavily, and similarly rewards high-confidence correct predictions over low-confidence ones to discourage guessing. This ensures models are evaluated not only on correctness but also on calibrated confidence and question difficulty. Results on the CaseHOLD dataset demonstrate our metric's ability to effectively reveal model strengths and weaknesses. While further experimentation is needed to confirm robustness, our metric offers a more reliable framework for high-stakes legal settings.

Contributions

Elizabeth Lu:

- Initial research: lit review on "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding", "LexGLUE", "FairLex"; found existing legal models of Legal-HeBERT, Pol-BERT-Large, Italian-LEGAL-BERT, JuriBERT, Custom-LEGAL-BERT, LEGAL-BERT, LEGAL-GPT-1,2; found datasets CaseHOLD, Case Law, etc. ; research on existing performance indicators
- Project Proposal: Wrote strategy for solving the problem, wrote part of the evaluation plan, and wrote and lit review 5, 6
- Ollama: debugged and tested model code; tested different input commands; set up Cuda, ran Llama 3.2 1B, Llama 3.2 3B, Llama 3.2 Vision 11B, Llama 3.1 8B, Llama 3 8B, Phi3 3.8B, Phi3 14B, Gemma2 2B, Gemma2 9B, Mistral 7B on the dataset, had to rerun all models due to error in the initial implementation of llama; implemented REGEX to clean LLM output; implemented evaluation metrics (accuracy, weighted precision, weighted recall, F1, Macro-F1, Micro-F1) for model results
- Metric Research: lit review on "Evaluating the Knowledge Dependency of Questions", "Evaluation of Question Answering Systems", "Can multiple-choice questions really be useful in detecting the abilities of LLMs?"; researched BLEURT and BERTScore; researched challenges for all types of evaluation scores
- Initial New Metric Idea: modified METEOR score to weight precision higher than recall, weighted metric that punished more harshly on incorrect answers rather than no answer
- Models: figured out how to run the models, ran BERT-Double, Legal-BERT, Custom Legal-BERT on CaseHOLD dataset, fine tuned the three models on the train dataset, reran the three models, debugged code, tested models, coded accuracy and filtered accuracy based on confidence threshold for model results
- New Metric: came up with new metric idea; implemented and tested log loss as potential confidence score; researched and wrote test cases for logloss determining it was unfit for our project; debugged BERTScore code; ran BERTScore model on dataset; wrote and ran the code to split dataset into difficulty quartiles; wrote and ran code to test if accuracy by quartile works; wrote and ran code for testing standard deviation vs geometric mean; debugged new metric code and wrote a detailed guide on how to implement the new metric
- Final Paper: wrote and revised 2.4 Dataset; wrote 3.1.1 Input and Outputs, 3.1.2 Fine-Tuning the Models, 3.2 Running the Models, 3.3 Evaluation Code, 3.4 New Metric: Weighted Correctness Score, 3.4.2 Confidence, 3.5 Limitations, 3.5.1 Catastrophic Forgetting, 3.5.2 Dataset Design and Question Difficulty, 3.5.3 Softmax Inability to capture Epistemic Uncertainty, 4 Results

Rachel Ren:

- Initial research: lit review on "Classification of US Supreme Court Cases using BERT-Based Techniques"; found datasets of PACER, Law Library of Congress, etc.
- Project Proposal: Wrote collaboration plan, expected outcomes, and lit review 1, 4
- Research on Metrics (pros, cons, when to use, application in our project): F1-score, accuracy, Micro-F1, Macro-F1, NDCG, Roc and AUC
- Research/Read: many many papers on Legal MCQ; different types of QA systems, Legal QA system; found granite by ibm (stated in a paper to be best QA system); found h2oGPT as a possible model; original paper on using LLM to create a model based metric; lit review on "The black box problem revisited", "Quantifying Calibration Error in Modern Neural Networks through Evidence Based Theory", "A Survey of Large Language Models"; lit review on TF-IDF and Cosine Similarity; looked at similar QA models and open source models in other domains such as Medical field with google's Med-PaLM; researched BERT/BERTspecific metrics; BERT in QA and what makes BERT Robust
- Initial New Metric Idea: coverage metric; penalty metric
- Final Paper first Draft: wrote entire first draft (scrapped since we shifted directions)
- Final Paper: created and updated Project Pipeline Diagram; wrote Abstract, Keywords, 1 Introduction, 1.1 The Black Box Problem, 1.1.1 The Black Box Problem is Minuscule in Our Case, 1.1.2 Why We Value Accuracy More Than Interpretability, 1.2 Research Question, 2.1 System Types, 2.2 Metrics; accuracy, F1, Macro F1 in 2.2.1 Simple Untrained Automatic Evaluation Scores, 2.3 Models, intro of 3. Methodology, added citations

Neil Tripathi:

- Initial research: lit review on "Named Entity Recognition and Relation Extraction", "A Survey on Recent Advances in Named Entity Recognition from Deep Learning models"; researched BERT, LLMs, transformers; researched potential datasets to use e.g. Caselaw; researched how to run models like BERT, Llama 3 and Gemini; researched Ollama and the code to use it, and proposed to use it to run the models
- Project Proposal: Wrote part of evaluation plan and lit review 3, 7
- Ollama: wrote code for initial implementation of llama on CaseHOLD; wrote code and implemented weighted accuracy as well as vector-based metrics for the model results (but we chose not to use them); wrote initial implementation to run BERT without fine tuning (but was not used)
- Initial New Metric Idea: proposed effort based metric; wrote code and tested cosine similarity script for different new metric idea
- New Metric: reached out to experts in the field to help come up with new metric; proposed confidence and log loss-based idea; worked with TA to finalize new metric; wrote code for running models for BERTScore; wrote code and implemented new metric and combined all data for each model; evaluated new metric for all models; compared and evaluated new metric to other metrics
- Final Paper: wrote geometric mean in 2.2.1, 3.4.1 Difficulty, 3.4.3 New Metric Formula, 4 Results, 5 Evaluation, 6 Future Works, 7 Conclusion 7.1 Answering Research Questions 7.2 Final Thoughts

Brynja Schultz:

- Initial research: found existing legal sources, interviewed legal professionals to discern what would be useful in the field, researched legal data sources, researched existing metrics, lit review on "A Comparative Study of Classifying Legal Documents with Neural Networks"
- Project Proposal: Wrote problem statement, sources, and lit review 2
- Presentation: created and presented slides 1-5 and 9 for the in-class presentation

- Summarized: Legal-BERT paper, “Domain Specialization as the Key to Make Large Language Models Disruptive”, "When Does Pretraining Help? Assessing Self-Supervised Learning for Law and the CaseHOLD Dataset", “Towards an Automated Pointwise Evaluation Metric for Generated Long-Form Legal Summaries”, "Understanding catastrophic forgetting in language models via implicit inference.", "Challenging common assumptions about catastrophic forgetting.", "Evaluating Evaluation Metrics: A Framework for Analyzing NLG Evaluation Metrics using Measurement Theory"

- Final Paper: wrote TF-IDF and cosine similarity in 2.2.1, 2.2.2 Machine-Trained Evaluation Scores: Bleurt and BertScore, and first draft of 2.4 Dataset; did research on catastrophic forgetting and wrote section in the final paper; created final references list

References

Monroy Alfredo, Calvo Hiram, and Gelbukh Alexander. 2009. Nlp for shallow question answering of legal documents using graphs. *SpringerLink*.

Farea Amer, Yang Zhen, Duong Kien, Perera Nadeesha, and Emmert-Streib Frank. 2022. Evaluation of question answering systems: Complexity of judging a natural language. *arXiv.org*.

Vold Andrew and Conrad Jack G. 2021. Using transformers to improve answer retrieval for legal questions. *ACM*.

Nussberger Anne-Marie, Luo Lulu, Celis L. Elisa, and Crockett Molly J. 2023. Public attitudes value interpretability but prioritize accuracy in artificial intelligence. *Nature Communications*.

Brożek Bartosz, Furman Mikołaj, Jakubiec Marek, and Kucharzyk Bartłomiej. 2023. The black box problem revisited: Real and imaginary challenges for automated legal decision making - artificial intelligence and law. *SpringerLink*.

Trautmann Daniel, Ostapuk Nicolai, Grail Quentin, Pol Annerose A., Bonifazi Guido, Gao Shuyang, and Gajek Michael. 2024. Measuring the groundedness of legal question-answering systems. *arXiv.org*.

Chalkidis Ilias, Fergadiotis Manos, Malakasiotis Prodromos, Aletras Nikolaos, and Androutsopoulos Ion. 2020. Legal-bert: The muppets straight out of law school. In *Findings of the Association for Computational Linguistics: EMNLP 2020*.

Devlin Jacob, Chang Ming-Wei, Lee Kenton, and Toutanova Kristina. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv.org*.

Gao Jingwei, Yao Jun, and Shao Yi. 2019. Towards reliable learning for high stakes applications. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Martinez-Gi Jorge. 2023. A survey on legal question–answering systems. *arXiv*.

Richmond Kevin M., Muddamsetty Sai M., Gammeltoft-Hansen Thomas, Olsen Hans Peter, and Moeslund Thomas B. 2023. Explainable ai and law: An evidential survey - digital society. *SpringerLink*.

Khazaeli Soha, JPunuru Janardhana, Morris Chad, Sharma Sanjay, Staub Bert, Cole Michael, ChiuWebster Sunny, and Sakalley Dhruv. 2021. A free format legal question answering system. *ACL Aclanthology*.

Kotha Suhas, Springer Jacob Mitchell, and Raghunathan Aditi. 2024. Understanding catastrophic forgetting in language models via implicit inference. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Evidently AI Team. 2024. Accuracy, precision, and recall in multi-class classification.

Pearce Tim, Brintrup Alexandra, and Zhu Jun. 2021. Understanding softmax confidence and uncertainty. *CoRR*, abs/2106.04972.

Lesort Timothée, Ostapenko Oleksiy, Misra Diganta, Arefin Md Rifat, Rodríguez Pau, Charlin Laurent, and Rish Irina. 2023. Challenging common assumptions about catastrophic forgetting. *arXiv.org*.

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. *Preprint*, arXiv:2104.08671.