eagerlearners at SemEval2024 Task 5: The Legal Argument Reasoning Task in Civil Procedure

Hoorieh Sabzevari, Mohammadmostafa Rostamkhani, Sauleh Eetemadi
Iran University of Science and Technology
h_sabzevari@elec.iust.ac.ir, mo_rostamkhani97@comp.iust.ac.ir, sauleh@iust.ac.ir

Abstract

This study investigates the performance of the zero-shot method in classifying data using three large language models, alongside two models with large input token sizes and the two pre-trained models on legal data. Our main dataset comes from the domain of U.S. civil procedure. It includes summaries of legal cases, specific questions, potential answers, and detailed explanations for why each solution is relevant, all sourced from a book aimed at law students. By comparing different methods, we aimed to understand how effectively they handle the complexities found in legal datasets. Our findings show how well the zero-shot method of large language models can understand complicated data. We achieved our highest F_1 score of 64% in these experiments.

1 Introduction

Becoming skilled at presenting a legal case is essential for aspiring lawyers. It requires understanding not only the relevant legal areas but also using advanced reasoning tactics like making analogies and spotting hidden contradictions. (Chalkidis et al., 2022). Despite efforts to set standards for modern NLP models in legal language understanding, there still aren't complex tasks focusing on argumentation in legal matters.(Bongard et al., 2022)

The dataset utilized in this study was gathered from *The Glannon Guide To Civil Procedure* by Joseph Glannon (Glannon, 2018) in English. Each sample within the dataset comprises a question, a solution, and an introduction elaborating on the provided solution. The objective is to identify whether the given answer, derived from the introduction text, accurately addresses the question.

This paper explores various approaches to address the challenge of handling lengthy and intricate data, which can be challenging for human comprehension. Initially, we evaluated two models—Longformer (Beltagy et al., 2020) and Big Bird (Zaheer et al., 2021) —known for their effectiveness in classifying data with large input tokens. Subsequently, we assessed the performance of two pre-trained models, Legal-RoBERTa (Chalkidis* et al., 2023) and Legal-XLM-RoBERTa (Niklaus et al., 2023) for legal data using the original code. Finally, we compared the performance of three large language models —GPT 3.5, Gemini, and Copilot— using the zero-shot method on the test dataset.

We recognized the significant impact of leveraging the capabilities and extensive capacity of large language models on analyzing data, especially those focusing on specific topics or lengthy content. Looking ahead, our goal is to improve prompts further to achieve superior results not only for this task but also for similar works. Further details regarding the implementation can be found in this GitHub repository.

2 Background

2.1 Task Setup

As previously mentioned, the original dataset for this task is sourced from *The Glannon Guide To Civil Procedure*. Each sample within the dataset comprises the following components:

1. Question 2. Answer 3. Label 4. Analysis 5. Complete Analysis 6. Explanation

It's noteworthy that "Analysis" and "Complete Analysis" were absent in the test data. The data split involves allocating the initial 80% of questions from each chapter to the training set, the subsequent 10% to the validation set, and the final 10%—typically more challenging questions—to the test set. The final dataset consists of 848 entries.

In this task, the inputs to the model consist of the "Question," "Explanation," and "Answer" values, while the output of the model is represented by the "Label." If the model determines that the

Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024), pages 922–926

answer provided for the question aligns with the explanation, the output label will be 1; otherwise, it will be 0. The objective of this task is to assess the model's reasoning capabilities, particularly its ability to analyze legal issues effectively.

2.2 Related Work

2.2.1 Pre-trained Legal Language Models

Numerous studies have been conducted in the realm of legal issues. Given the challenges posed by comprehending lengthy texts within this domain using existing models, pre-trained language models have been tailored to address this need. One such model is the Legal-BERT model (Chalkidis et al., 2020), which is also employed in this paper. This study introduces a specialized model aimed at facilitating NLP-based legal research by fine-tuning the original BERT model for legal applications. (Li et al., 2023) introduces SAILER, a novel pre-trained linguistic model with a unique architecture designed for legal case retrieval. Furthermore, (Cui et al., 2023b) provides a comprehensive survey of existing Legal Judgment Prediction (LJP) tasks, datasets, and models within the legal domain, encompassing an overview of 8 pre-trained models across 4 languages as part of the LJP. Moreover, an end-to-end methodology is introduced by (Louis et al., 2023) for generating long-form answers to statutory law questions, addressing limitations in existing Legal Question Answering (LQA) approaches.

2.2.2 Domain-Specific LLMs in Law

A Large Language Model (LLM), such as ChatGPT, is remarkable for its ability to handle general-purpose language generation and a variety of other NLP tasks. Domain-specific LLMs are versatile models optimized to excel at specific tasks defined by organizational standards. They further empower lawyers to expand their understanding and explore specialized legal domains. For instance, (Colombo et al., 2024) is the first LLM designed explicitly for legal text comprehension and generation with 7 billion parameters. To empower the legal field, (Cui et al., 2023a) presents ChatLaw, an open-source legal LLM built with a high-quality, domain-specific fine-tuning dataset. The focus of (Savelka et al., 2023) is evaluating GPT-4's effectiveness in generating explanations for legal terms – specifically, whether they are accurate, clear, and relevant to the surrounding legislation.

3 System Overview

3.1 Preprocessing Data

Our dataset comprises 666 samples for the training set, 84 samples for the validation set, and 98 samples for the test set. Initially, we excluded the columns "Analysis" and "Complete Analysis" from both the training and validation datasets as they were absent in the test data. Afterward, we analyzed to determine the distribution of class labels 0 and 1, revealing a notable class imbalance, with the number of instances belonging to class 0 nearly three times higher than those of class 1.

To address this issue, various approaches can be employed. In this study, we opted to mitigate the class imbalance using the focal loss function as our loss function. Our investigation demonstrates the efficacy of focal loss in rectifying class imbalance, enhancing the performance of classes with limited training samples, offering adaptability in adjusting the learning process, and attenuating the impact of noisy data.

3.2 Model

3.2.1 Pre-trained Models

In this study, we tackled the challenge of dealing with long sets of data. To overcome this, we looked into using two models designed to handle large inputs: the Longformer and Big Bird models. These models can handle up to 4096 tokens, which is much more than the BERT model. We also used two pre-trained models specifically trained for legal data: Legal-RoBERTa and Legal-XLM-RoBERTa. These models, like Legal-BERT (Chalkidis et al., 2020), were trained on various legal documents and cases.

Our aim is straightforward: to compare the performance of these models and understand how using pre-trained models with larger input sizes impacts their effectiveness. Through this analysis, we aim to gain insights into the optimal approaches for managing complex legal datasets.

To address the issue of unbalanced data, we implemented the focal loss function, a method that has shown promising outcomes in previous research (Lin et al., 2018) (Wang et al., 2022). The Focal Loss function formally incorporates a factor of $(1 - p_t)^\gamma$ into the standard cross-entropy criterion. This adjustment diminishes the relative loss for accurately classified examples ($p_t > 0.5$), thereby intensifying the focus on challenging instances that

are misclassified.

$$CE(p_t) = -log(p_t) \quad (1)$$

$$FL(p_t) = -(1-p_t)^{\gamma} log(p_t) \quad (2)$$

There is a tunable focusing parameter $\gamma \geq 0$. Figure 1 illustrates the varied impact of this parameter across a range of values.

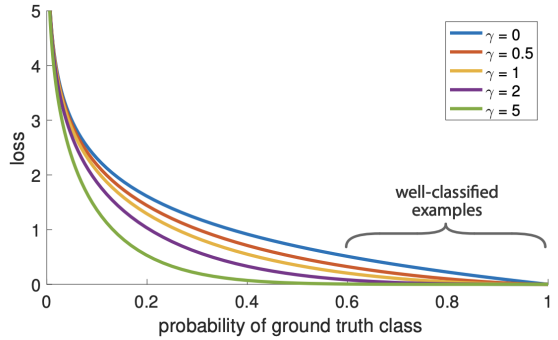

Figure 1: Comparison between cross entropy and focal loss

3.2.2 Large Language Models

In addition, we leveraged three popular large language models to assess their performance within the given task: OpenAI's GPT 3.5 model, Google's Gemini model, and Bing's Copilot model. Later, we formulated the task in the form of prompts, refined them through prompt engineering techniques, and presented them to the large language models. Our objective was to extract the desired answer (0 or 1) from the models' responses, thereby assessing their performance and capabilities in handling the task. We conducted extensive tests on numerous prompts to identify the most effective ones. Employing prompt engineering methodologies alongside large language models, we refined these prompts to enhance their performance and effectiveness in achieving our objectives.

4 Experimental setup

We utilized a dataset of 848 data points, dividing it into 80% for the training dataset, 10% for the validation dataset, and 10% for the test dataset. Our experimental approach involved exploring three main inquiries:

1. Exploration of Longformer and Big Bird models, tailored to effectively process lengthy data inputs.

2. Utilization of pre-trained models for legal contexts, including Legal-RoBERTa and Legal-XLM-RoBERTa, to ascertain their efficacy in legal text analysis.

Hyperparameter	Value
Optimizer	AdamW
Learning Rate	5e-5
Epochs	3
Batch Size	1
Loss Function	Focal Loss

Table 1: Hyperparameter values

Model	Accuracy	$\mathbf{F_1}$
Longformer	**0.79**	0.44
Big Bird	**0.79**	0.44
Legal-BERT	0.75	**0.58**
Legal-RoBERTa	**0.79**	0.44
Legal-XLM-RoBERTa	0.78	0.43

Table 2: Accuracy and macro F1-score of first and second parts' models on the validation set

3. Implementation of the Zero-shot methodology with prompt feeding across three distinct models: GPT 3.5[1], Gemini[2], and Copilot[3], aimed at exploring their adaptability and performance across diverse tasks.

In the initial phase, we employed the AdamW optimizer function with a learning rate set at 5e-5 for both models, conducting training over 3 epochs. The specific values chosen for each hyperparameter are listed in Table 1.

Then, in the second phase, we adapted the original code from the article with the necessary modifications.[4] The original code featured the utilization of Sliding Window Simple (SWS) and Sliding Window Complex (SWC) methods with the Legal-BERT model. We made adjustments to certain sections of the code to ensure compatibility with the test input data. Throughout our evaluation, we utilized the macro F_1 score as the primary evaluation metric.

In the final phase of our experiments, we used the API keys provided by OpenAI and Google. Unfortunately, despite our efforts, we encountered challenges locating the official API for Bing. Given its superior accuracy compared to other models, we decided to manually record the results of the test dataset. Throughout this phase, we employed a variety of prompts and iteratively refined them.

[1] https://chat.openai.com/
[2] https://gemini.google.com/app
[3] https://www.bing.com/chat
[4] https://github.com/trusthlt/legal-argument-reasoning-task

Model	Accuracy	F_1
GPT 3.5	**0.69**	0.59
Gemini	0.44	0.44
Copilot	0.67	**0.64**

Table 3: Macro F1-score of large language models on the test set

Our investigation revealed that incorporating terms such as "step by step" or asking for explanations of the inference steps to achieve the desired outcome positively impacted the model's performance. Furthermore, we encountered limitations in prompt completeness due to the token size constraints inherent in the input models. For instance, Bing's Copilot supported a maximum input size of 4000 characters, which proved insufficient for processing our long samples. Here is an example of an input prompt:

I will provide a question, an answer, and an explanation. Your task is to determine if the answer is correct based on the explanation provided. After reading the explanation, please respond with 'yes' if the answer is correct, or 'no' if it is incorrect.

Question: {question}
Proposed Answer: {answer}
Explanation: {explanation}

Is the proposed answer correct based on the explanation? (yes or no)
Please provide your detailed reason for your choice.
Then, reevaluate and check whether the selected answer is logical or not.
Please use the following format:
<selected_answer>: yes/no
<reason>: your reason for the initial choice
<reason for logical check>: your reason for reevaluation

5 Results

After completing the aforementioned three phases, our investigation revealed that despite fine-tuning, existing models struggled to effectively analyze lengthy data within challenging legal contexts, encountering training process issues and yielding suboptimal outputs. Moreover, the most promising result emerged from fine-tuning the Legal-BERT model, serving as the baseline. While we continue to analyze the underlying reasons for this outcome, initial observations suggest that the learning challenge may be linked to the specific characteristics of the dataset employed. Table 2 presents the performance metrics of the models from the initial two phases, evaluated on the validation dataset.

When it comes to evaluating the results of the zero-shot method, we identified its considerable potential and Bing's Copilot model emerged as the top performer, surpassing expectations. Following suit, the GPT 3.5 model presented moderate performance, while the Gemini model fell short of expected levels. The success of the Copilot model lies in its ability to address previous challenges associated with GPT models by leveraging real-time information accessible through the internet. Table 3 presents the results achieved from employing this method on the test dataset, representing the unofficial results submitted during the post-evaluation phase.

6 Conclusion

In this paper, we present methods designed for classifying lengthy legal cases. We divided our exploration into three main parts:

Firstly, we looked into models with large input token sizes such as Longformer and Big Bird. Secondly, we examined pre-trained models specifically fine-tuned for legal data, such as Legal-RoBERTa and Legal-XLM-RoBERTa. Lastly, we tested the zero-shot method across three major language models.

Among these methods, we found that the zero-shot technique and Bing's Copilot model showed the most promising performance. As for future works, we can explore techniques like data summarization, collaborative approaches such as the round table technique, trying various hyperparameters, and refine prompts to further enhance model performance. These efforts have the potential to advance the effectiveness of classification tasks in legal contexts.

As a future work, it would be valuable to explore additional large language models. These models offer extensive capabilities, especially in summarizing lengthy datasets, which could help evaluate various models' performance. However, it is necessary to note that during summarization, some important details might be overlooked.

Another avenue for future research involves testing the effectiveness of a multi-model approach.

(Chen et al., 2023) This method entails bringing together different large language model agents in a round table conference format. This setup encourages diverse perspectives and discussions to foster consensus. By adopting this approach, researchers can tap into the combined intelligence of multiple models, potentially enriching analysis across various tasks and domains. In this competition, our team achieved the 17th rank out of 21 groups. Our only submission during the evaluation phase utilized the basic prompt and the GPT model. However, significant improvements were made during the post-evaluation phase, resulting in a much higher level of accuracy.

References

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer.

Leonard Bongard, Lena Held, and Ivan Habernal. 2022. The Legal Argument Reasoning Task in Civil Procedure. In *Proceedings of the Natural Legal Language Processing Workshop 2022*, pages 194–207, Abu Dhabi, UAE. Association for Computational Linguistics.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. Legal-bert: The muppets straight out of law school.

Ilias Chalkidis*, Nicolas Garneau*, Catalina Goanta, Daniel Martin Katz, and Anders Søgaard. 2023. LeXFiles and LegalLAMA: Facilitating English Multinational Legal Language Model Development. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, Toronto, Canada. Association for Computational Linguistics.

Ilias Chalkidis, Abhik Jana, Dirk Hartung, Michael Bommarito, Ion Androutsopoulos, Daniel Katz, and Nikolaos Aletras. 2022. LexGLUE: A benchmark dataset for legal language understanding in English. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4310–4330, Dublin, Ireland. Association for Computational Linguistics.

Justin Chih-Yao Chen, Swarnadeep Saha, and Mohit Bansal. 2023. Reconcile: Round-table conference improves reasoning via consensus among diverse llms.

Pierre Colombo, Telmo Pessoa Pires, Malik Boudiaf, Dominic Culver, Rui Melo, Caio Corro, Andre F. T. Martins, Fabrizio Esposito, Vera Lúcia Raposo, Sofia Morgado, and Michael Desa. 2024. Saullm-7b: A pioneering large language model for law.

Jiaxi Cui, Zongjian Li, Yang Yan, Bohua Chen, and Li Yuan. 2023a. Chatlaw: Open-source legal large language model with integrated external knowledge bases.

Junyun Cui, Xiaoyu Shen, and Shaochun Wen. 2023b. A survey on legal judgment prediction: Datasets, metrics, models and challenges. *IEEE Access*, 11:102050–102071.

J.W. Glannon. 2018. *Glannon Guide to Civil Procedure: Learning Civil Procedure Through Multiple-Choice Questions and Analysis*. Glannon Guides Series. Aspen Publishing.

Haitao Li, Qingyao Ai, Jia Chen, Qian Dong, Yueyue Wu, Yiqun Liu, Chong Chen, and Qi Tian. 2023. Sailer: Structure-aware pre-trained language model for legal case retrieval.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2018. Focal loss for dense object detection.

Antoine Louis, Gijs van Dijck, and Gerasimos Spanakis. 2023. Interpretable long-form legal question answering with retrieval-augmented large language models.

Joel Niklaus, Veton Matoshi, Matthias Sturmer, Ilias Chalkidis, and Daniel E. Ho. 2023. Multilegalpile: A 689gb multilingual legal corpus. *ArXiv*, abs/2306.02069.

Jaromir Savelka, Kevin D. Ashley, Morgan A. Gray, Hannes Westermann, and Huihui Xu. 2023. Explaining legal concepts with augmented large language models (gpt-4).

Cheng Wang, Jorge Balazs, György Szarvas, Patrick Ernst, Lahari Poddar, and Pavel Danchenko. 2022. Calibrating imbalanced classifiers with focal loss: An empirical study. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 145–153, Abu Dhabi, UAE. Association for Computational Linguistics.

Manzil Zaheer, Guru Guruganesh, Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and Amr Ahmed. 2021. Big bird: Transformers for longer sequences.