SU-FMI at SemEval-2024 Task 5: From BERT Fine-Tuning to LLM Prompt Engineering - Approaches in Legal Argument Reasoning

Kristiyan Krumov
FMI, Sofia University
kristiyan.boyanov@gmail.com

Svetla Boytcheva
Ontotext
svetla@uni-sofia.bg
svetla.boytcheva@ontotext.com

Ivan Koytchev
FMI, Sofia University
koychev@fmi.uni-sofia.bg

Abstract

This paper presents our approach and findings for SemEval-2024 Task 5, focusing on legal argument reasoning. We explored the effectiveness of fine-tuning pre-trained BERT models and the innovative application of large language models (LLMs) through prompt engineering in the context of legal texts. Our methodology involved a combination of techniques to address the challenges posed by legal language processing, including handling long texts and optimizing natural language understanding (NLU) capabilities for the legal domain. Our contributions were validated by achieving a third-place ranking on the SemEval 2024 Task 5 Leaderboard. The results underscore the potential of LLMs and prompt engineering in enhancing legal reasoning tasks, offering insights into the evolving landscape of NLU technologies within the legal field.

1 Introduction

Legal texts, including laws, interpretations, arguments, and agreements, are commonly conveyed through writing, resulting in great amount of legal documents. Analyzing these documents, a core aspect of legal work, becomes more intricate as these collections expand. Natural language understanding (NLU) technologies offer potential assistance to legal professionals in this regard. However, their effectiveness hinges on the ability of current state-of-the-art models to adapt to diverse tasks within the legal field.

The legal argument reasoning task (Bongard et al., 2022) of SemEval-2024 represents a significant challenge in the domain of natural language processing (NLP) and an informal addition to the currently existing model evaluation benchmarks such as LexGLUE (Chalkidis et al., 2022b).

Our approach involves fine-tuning pre-trained BERT models and exploring the innovative use of large language models (LLMs) through prompt engineering to address this task.

As a result of our work, we are ranked 3-rd in the SemEval 2024 Task 5[1] Leaderboard out of 20 participating teams. The implementations of the different approaches is available on Github[2] and the fine-tuned models could be accessed in Huggingface[3].

2 Background

Task 5 of SemEval 2024 is novel NLP problem focused on legal argument reasoning within the context of U.S. civil procedure. It contributes a dataset comprised of instances each containing a general introduction to a case, a specific legal question, a proposed solution argument, and a detailed analysis explaining the applicability of the argument. This dataset aims to benchmark the performance of legal language models, posing a significant challenge due to the complexity and nuanced understanding required for legal reasoning. Instances are organized to support a binary classification task: determining the correctness of a given answer to a legal question, aimed at facilitating research on legal argument reasoning.

In the domain of text classification, conventional methodologies often employ "short encoders" such as BERT (Devlin et al., 2018), RoBERTa (Liu et al., 2019), and XLNet (Yang et al., 2019), which have demonstrated commendable efficacy in diverse contexts, ranging from news topic classification to sentiment analysis in movie reviews. Nevertheless, these encoders are constrained by their 512-token processing limit, rendering them less effective for analyzing extensive documents like court judgments. To circumvent this limitation, more advanced approaches, including the Hierarchical Attention Network (HAN) (Yang et al., 2016), a synergy of BERT and CNN, and the combination

[1] https://trusthlt.github.io/semeval24/
[2] https://github.com/frisibeli/semeval-2024-task5
[3] https://huggingface.co/frisibeli

Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024), pages 1652–1658

of XLNet with BiGRU (Chenxi et al., 2022), have been developed, enhancing the semantic understanding of longer texts. Despite these technological strides, the pursuit of an optimal algorithm that can adeptly navigate the complexities of extended documents persists.

The application of automated systems in the legal domain encounters distinct challenges, arising from the specialized language employed and the necessity for intricate multi-step reasoning over extensive texts. Furthermore, the potential of leveraging recent advancements in prompting techniques for legal domain-specific tasks remains largely unexplored. Typically, effective prompting in general NLP tasks has been noted with concise inputs, often limited to a single sentence or a small collection of sentences, accompanied by a restricted array of target labels. This underscores the ongoing quest to adapt and refine NLP techniques to meet the unique demands and intricacies of legal reasoning.

3 System Overview

After analyzing the dataset, we identified that the final system should be capable of handling relatively lengthy contexts and to perform well on reasoning and fact-checking tasks. In this section we separately introduce the different approaches we have experimented with on solving the Legal Argument Reasoning task by dividing them into methods for handling long texts and such for optimizing the NLU capabilities for the legal domain.

3.1 Handling Long Texts

Observing the distributions (Fig. 2) of the token lengths for the dataset entries we could say that a system capable of processing contexts of 2000 tokens would be sufficient to cover the majority of the cases.

3.1.1 Sliding Window (SW)

We leveraged the sliding window techniques as described in (Bongard et al., 2022), as a baseline to overcome the maximum token limit problem. We experimented with **Sliding Window Simple** and **Sliding Window Complex**

3.1.2 Transformer-based models for long text

Transformer-based models encounter difficulty processing lengthy sequences due to their self-attention operation, which exhibits quadratic scaling with sequence length. In response to this constraint, we experimented with the Longformer (Beltagy et al., 2020) model, featuring an attention mechanism that scales linearly with sequence length and increases the maximum input length to 4096 sub-word tokens, which may also improve the performance in understanding legal documents. Additionally, we experimented with Legal-RoBERTa and Legal-Longformer - pre-trained models on legal corpus introduced in (Chalkidis et al., 2023).

3.1.3 Summarizing

A different approach we tried for preprocessing lengthy texts was utilizing summarization models. By condensing extensive content into concise summaries, we not only mitigate the challenges posed by the length limitations of Transformer-based architectures but also streamline subsequent processing stages by reducing the presence of extraneous or tangential content, such as author's thoughts and remarks (Fig. 3).

As part of our solution, we examined several summarization models - BART (Lewis et al., 2019), LexRank (Erkan and Radev, 2004) and ChatGPT4[4].

3.2 Optimising NLU Capabilities for the Legal Domain

Research has demonstrated the efficacy of language model pre-training in enhancing numerous natural language understanding tasks like natural language inference (Devlin et al., 2019). In addition to learning linguistic knowledge, these models are retaining relational knowledge (Petroni et al., 2019) present in the training data which could be beneficial in solving downstream tasks in domains such as the legal one and more precisely - US Civil Procedure where the legal system is based on precedents. In this section we are going to reflect on the methods used by us to improve the performance of the system by enhancing its reasoning capabilities.

3.2.1 Pre-trained Transformer Models on Legal Corpus

As a starting point in addressing the problem, we decided to use Legal-BERT (Chalkidis et al., 2020), being the most successful baseline experiment described in the work of the organizers of the task (Bongard et al., 2022). We fine-tuned it on the task and additionally - on a custom legal dataset 3.2.2. Our contribution continued with the exploration

[4] https://platform.openai.com/docs/models/gpt-4-and-gpt-4-turbo

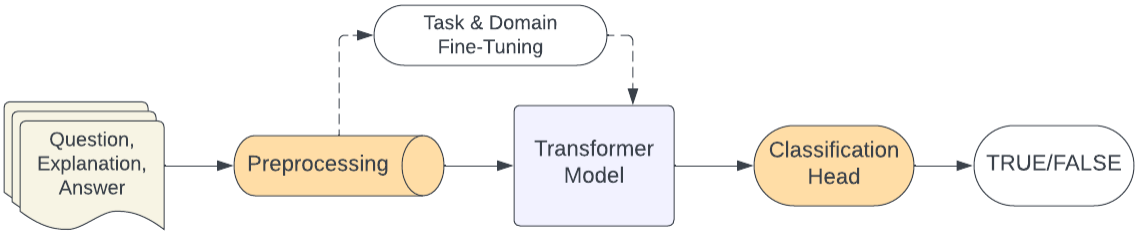

Figure 1: Transformer-based classifier system architecture

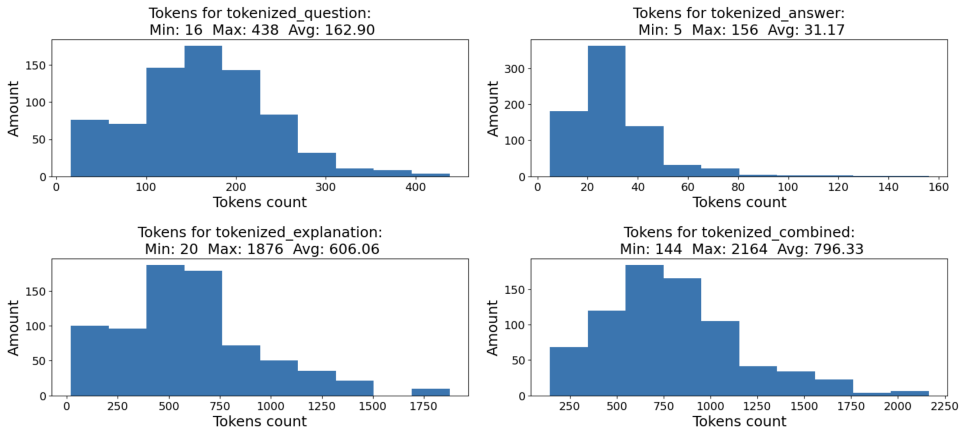

Figure 2: Token count distribution of the dataset per entry parts - Question, Explanation, Answer and Concatenated

of alternative legal transformer models: CaseHold-BERT (Zheng et al., 2021), variants of Legal-BERT (small, large), Legal-RoBERTa (Chalkidis et al., 2023) and InLegalBERT (Paul et al., 2023).

3.2.2 Fine-tuned BERT on Custom Dataset

We additionally fine-tuned the best performing models from 3.2.1 on a custom-tailored dataset of an American civil procedure data (4.3), similar to the entries from the task. The goal with this approach was to strengthen the model's relational knowledge and contextual representations of the language used in the legal domain (Petroni et al., 2019).

3.3 LLM + Legal Prompt Engineering

So far we observed the task as a supervised classification problem, where the models are trained with labeled data to classify inputs into a binary output. Another approach is to use the relatively new method of prompt engineering in combination with some of the currently best-performing generative models (Fig. 4). With prompting, there's generally no need for additional training as the model receives a prompt, which could be a question, examples of input-output pairs (few-shot learning), or task descriptions. This approach allows the model to leverage its pre-trained knowledge to produce outputs for specific tasks in a zero-shot manner, meaning it can generate correct responses without having seen examples of the specific task during its training phase. For this setup, we experimented with several types of LLMs: Mistral-7b-Instruct (Jiang et al., 2023), Llama2-70b (et al., 2023), GPT-3.5-Turbo and GPT-4 [5]; as for most of those models we performed prompt fine-tuning and Legal prompt engineering (Trautmann et al., 2022).

4 Experimental Setup

4.1 Data

For the transformer-based classifier systems (Fig. 1) we performed experiments on the SemEval 2024 Task (Bongard et al., 2022) dataset. We stratified the train partition (750 entries) into train* (88%) and train-dev (12%), ensuring that the distribution of label values was maintained. The dev partition (84 entries) and the test partition were solely utilized for validation to prevent overfitting and bias in the model.

[5] https://platform.openai.com/docs/models

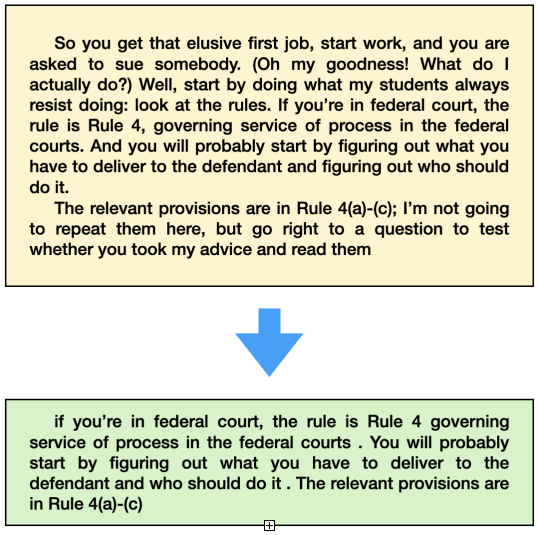

Figure 3: Preprocessing an Explanation from the dataset using T5 for summarization

On other hand, for the generative-based classifier systems (Fig. 4) we used only the dev and test partitions leveraging the generalization capabilities of the large models and inferring in a zero-shot manner.

4.2 Fine-Tuning & Hyperparameters

All experiments related to transformer-based classifier systems (Fig. 1) were conducted using a single A100 40GB GPU, with the following hyperparameters: 5 training epochs and a learning rate of $2e-5$. Additionally weight-decay and early-stopping (patience = 3) were applied.

For the generative-based systems different environments were used:

- Local setup (Apple M2) + OpenAI access for GPT-4, GPT-3.5-turbo
- Local setup (Apple M2) including Ollama[6] for running Mistral-7b and Llama2

We experimented with low temperature hyperparameter values, ranging $0-0.2$, in order to achieve more deterministic results.

4.3 Custom Legal Dataset

For MLM fine-tuning the transformer classifier, a new custom-tailored U.S. Civil Procedure dataset (Ref. 3.2.2) was used. It was collected first by automatically extracting the keywords from each unique explanation+question entry, then manually creating search queries and finally - using the open search API of the Caselaw Access Project[7] downloading relevant cases. The final corpora consists of 1985 different legal texts (cases), sourced by storing each 20 most relevant results for 100 queries.

4.4 Legal Prompt Engineering (LPE)

In (Trautmann et al., 2022), the authors define "Legal prompt engineering (LPE)" as the process of creating, evaluating, and recommending prompts for legal NLP tasks. In the current work as an alternative approach to the transformer-based classifier systems we investigate the performance of LPE on the SemEval 2024 Legal task. We used more than 15 prompts (A.1), as for their creation, we followed some of the 26 principles described in (Bsharat et al., 2024). Modification of a prompt version was done after evaluating how certain changes affect the performance.

The general frame of the prompt was in the form of a task or question, for which the model has to answer only with "TRUE" or "FALSE". An interesting observation is that GPT-3/4 and Llama2 almost always follow that restriction and return one of the two desired outputs with very few times returning something slightly different (e.g. different casing or appending punctuation - "false.", "True"). Contrarily, Mistral-7b-instruct always returns the answer with an additional explanation, which led to a more complex post-processing step for that model.

We used LangChain[8] for prompt template processing, model-agnostic interface unification and easy response post-processing.

5 Results

Table 1 shows the results of the different experiments. The evaluation of different models for the SemEval-2024 Task 5 on Legal Argument Reasoning presents interesting observation on how different models and system types, described in the current work perform on the dev and test dataset partitions. Our baseline approaches, Majority and Random, set the initial benchmarks with Macro-F1 scores significantly lower than those achieved by advanced models, underscoring the complexity of the task. The application of transformer models, including those equipped with a Sliding Window

[6] https://ollama.com/

[7] https://case.law/

[8] LLM framework - https://python.langchain.com/

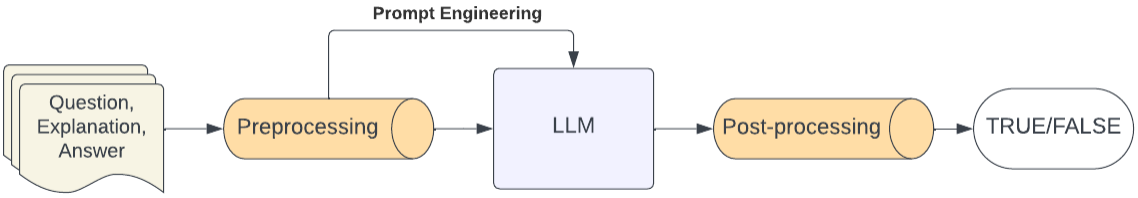

Figure 4: Generative model-based system architecture

Model Name	System Type	Dev Macro-F1	Test Macro-F1
Majority	Baseline	0.44	0.42
Random		0.46	0.46
CaseHold/LegalBERT + SW	Transformer	0.55	-
LegalBERT + SW		0.59	-
LegalBERT-small + SW		0.53	-
InLegalBERT + SW		0.44	-
lexlms/legal-longformer-base		0.50	-
SU-FMI-LegalBERT + SW		0.60	-
lexlms/legal-roberta-large		0.62	0.49
legal-roberta + BART	Classifier + Summary	-	0.50
SU-FMI-LegalBERT + BART		-	0.52
CaseHold/LegalBERT + BART		-	0.54
CaseHold/LegalBERT + GPT-4		-	0.55
CaseHold/LegalBERT + GPT-4 V2		-	0.61
Mistral-7b + LPE	LLM	-	0.58
Llama2-70b + LPE		0.59	0.58
GPT-3.5-turbo + LPE		0.58	0.60
GPT-4 + LPE		0.74	0.7728

Table 1: Model Performance on Development and Test Sets

technique and summarization capabilities, such as BART, showed improvement over the baselines, indicating the value of contextual understanding and content summarization in legal reasoning tasks.

Notably, the integration of Large Language Models (LLMs) with Legal Prompt Engineering (LPE) techniques, particularly with GPT-3.5-turbo and GPT-4, led to a significant leap in performance metrics. These models outperformed traditional transformer models, highlighting the effectiveness of LPE in enhancing the model's ability to interpret and reason over legal texts.

The comparative analysis of model performances on both development and test datasets revealed consistent patterns. Models utilizing LLMs with LPE not only achieved the highest Macro-F1 scores but also demonstrated robustness across different data sets, underscoring their potential for real-world applications in legal reasoning and argumentation.

6 Conclusion

Our participation in SemEval-2024's legal argument reasoning task has yielded valuable insights into the capabilities of transformer-based models and LLMs in processing and reasoning over legal texts. While our methods have shown promise, particularly in leveraging LLMs and prompt engineering, the complexity of legal reasoning poses ongoing challenges.

Further investigation can be done in a solution based on a hierarchical transformer variant such as HIER-BERT (Chalkidis et al., 2022b), (Chalkidis et al., 2019) (Chalkidis et al., 2022a). Our initial experiments with that model architecture did not lead to very high results (0.47 f1-macro on the dev partition) and because of the setup complexity, we decided to leave it for future research opportunities.

7 Acknowledgments

This work was partly supported by the European Union-NextGenerationEU, through the National Recovery and Resilience Plan of the Republic of Bulgaria, project No BG-RRP-2.004-0008.

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.

Sondos Mahmoud Bsharat, Aidar Myrzakhan, and Zhiqiang Shen. 2024. Principled instructions are all you need for questioning llama-1/2, gpt-3.5/4.

Ilias Chalkidis, Ion Androutsopoulos, and Nikolaos Aletras. 2019. Neural legal judgment prediction in English. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4317–4323, Florence, Italy. Association for Computational Linguistics.

Ilias Chalkidis, Xiang Dai, Manos Fergadiotis, Prodromos Malakasiotis, and Desmond Elliott. 2022a. An exploration of hierarchical attention transformers for efficient long document classification.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. Legal-bert: The muppets straight out of law school. *arXiv preprint arXiv:2010.02559*.

Ilias Chalkidis, Nicolas Garneau, Catalina Goanta, Daniel Martin Katz, and Anders Søgaard. 2023. Lex-files and legallama: Facilitating english multinational legal language model development.

Ilias Chalkidis, Abhik Jana, Dirk Hartung, Michael Bommarito, Ion Androutsopoulos, Daniel Katz, and Nikolaos Aletras. 2022b. 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.

Li Chenxi, Feng Jilin, Huang Meng, and Wang Zhonghao. 2022. Research on post earthquake public opinion analysis based on xlnet-bigru-a algorithm. In *2022 IEEE 2nd International Conference on Computer Communication and Artificial Intelligence (CCAI)*, pages 81–84. IEEE.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

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

G. Erkan and D. R. Radev. 2004. Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research*, 22:457–479.

Hugo Touvron et al. 2023. Llama 2: Open foundation and fine-tuned chat models.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2019. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *CoRR*, abs/1910.13461.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Shounak Paul, Arpan Mandal, Pawan Goyal, and Saptarshi Ghosh. 2023. Pre-trained language models for the legal domain: A case study on indian law.

Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H. Miller, and Sebastian Riedel. 2019. Language models as knowledge bases?

Dietrich Trautmann, Alina Petrova, and Frank Schilder. 2022. Legal prompt engineering for multilingual legal judgement prediction.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 1480–1489.

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.

A Appendix

A.1 Prompts

A.1.1 Best Prompt (GPT-4, GPT-3.5, Llama2-70b

System Prompt:

system_prompt = """

You are a legal assistant with a specialization in U.S. Civil Procedure. Your role involves thorough analysis and resolution of cases pertaining to this field. You will encounter three key components in each case:

1. EXPLANATION: This provides additional context and background information about a specific lawsuit.

2. QUESTION: Here, you will be presented with actual facts and details surrounding the lawsuit.

3. HYPOTHESIS: Based on the provided information, a hypothesis will be presented. Your task is to rigorously evaluate this hypothesis in the context of U.S. Civil Procedure and determine its validity. Respond ONLY with 'TRUE' if you conclude that the hypothesis is correct, or ONLY with 'FALSE' if you find it to be incorrect.

*Do no*t provide any reasoning behind your decision.

"""

User Input:

input_template = """

EXPLANATION: {}

QUESTION: {}

HYPOTHESIS: {}

"""

A.1.2 Chain of thoughts

system_prompt = """

You are a legal assistant with a specialization in U.S. Civil Procedure. Your role involves thorough analysis and resolution of cases pertaining to this field. You will encounter three key components in each case:

1. EXPLANATION: This provides additional context and background information about a specific lawsuit.

2. QUESTION: Here, you will be presented with actual facts and details surrounding the lawsuit.

3. HYPOTHESIS: Based on the provided information, a hypothesis will be presented. Your task is to rigorously evaluate this hypothesis in the context of U.S. Civil Procedure and determine its validity.

On User input with EXPLANATION, QUESTION and HYPOTHESIS analyse the legal problem step by step. Explain your thoughts.

"""

final_input = """

Respond ONLY with 'TRUE' if you conclude that the hypothesis is correct, or ONLY with 'FALSE' if you find it to be incorrect.

Do not provide any reasoning and ONLY answer with 'TRUE' or 'FALSE'

"""

A.1.3 Mistral-7b-instruct Best Prompt

You are a helpful civil law assistant. Your answer only with "TRUE" or "FALSE". You answer with "TRUE" if the STATEMENT is correct based on the provided CONTEXT or "FALSE" otherwise. If you don't know the answer - answer with FALSE.

==========

The CONTEXT is {explanation} | {question}

==========

The STATEMENT is: {answer}