LLMs for LLMs: A Structured Prompting Methodology for Long Legal Documents

Strahinja Klem[1] and Noura Al Moubayed[1]

Durham University, Durham, UK,
strahinja.klem@durham.ac.uk

Abstract. Large Language Models (LLMs) and other transformer models have found applications in medicine, financial technology, education and many others. However, their adoption into the legal sphere has proven complex and challenging due to the unique structure and nature of legal texts as well as issues of reliability, transparency and accountability. We present a pipeline which outperforms the previous method [1] (that uses DeBERTa-large) by up to 9%, addresses the aforementioned issues better than fine-tuning and highlights the need for better automatic evaluation metrics.

Keywords: Machine Learning, Prompt Engineering, Law

1 Introduction

We address two ubiquitous issues with LLMs which are especially potent within law. The Long Document Problem, when a document cannot fit within the context window of a model, and Information Retrieval from one more many documents. To do this, we introduce a pipeline centred around a prompt engineering framework which improves the performance of the foundation model QWEN-2 on the CUAD legal question-answering (QA) dataset [1] and allows for increased consistency and control. The documents first go through a segmentation and augmentation step, where inputs are chunked and combined to add redundancy, protecting against information loss at chunking thresholds. After producing a prompt via the framework, segments are queried in batches, reducing the search space for the model, increasing accuracy and reliability for each individual prompt. Finally, we use two hand-crafted, interpretable heuristics to weigh the outputs and select the most likely answer, adding transparency to the black-box model.

2 Methodology

Dataset and Model: The Contract Understanding Atticus Dataset [1] (CUAD) is designed for legal QA, containing 13,000 examples of expert curated data and high-quality annotations. The model we opted to use was the open-source QWEN2-7B as a proof-of-concept, as larger models will perform better based on scaling laws, and its accessibility aligns with our stated goals.

Structured Prompt Engineering (SPE): This is a framework for searching through prompts by iteration and testing to reliably maximise the performance of a LLM on a given task, akin to software engineering. We firstly split the data into a small training set, used for prompt creation and refinement, and a test set to show that the prompt generalises. Secondly, we construct a template which matches the QA input structure, and use it to generate a set of base prompts, via synonyms and paraphrasing. The small training set allows for extensive accuracy testing on both datasets, and the top performing candidate is selected as the base prompt for the next stage. Finally, using the same testing structure, we systematically create valid combinations of known techniques[2].
Heuristics: Distribution-Based Localisation (DBL) uses previous data to create a distribution of likelihoods of answer locations over document segments for each question, then weighs chunks based on their position within it to approximate the correct one, relying on the assumption that similar documents keep the same structure as they scale. Inverse Cardinality Weighing (ICW) clusters answer embeddings using DBSCAN with cosine distance as the metric, and weighs them inversely-proportionally to the size of their respective groups. This follows our empirical observation that the proportion of incorrect to correct answers tends to be high. The performance of the heuristics will vary depending on the validity of these assumptions. However, our experiments show they predominantly hold.

3 Results and Conclusions

During the study, both human evaluation and automatic metrics (ROUGE, METEOR and Cosine Similarity) were applied. We find the metrics are less effective the shorter a text becomes, which limits effectiveness in QA. They also penalise responses which do not match exactly, disadvantaging the generative approach. Hence, we rely on human evaluation for accuracy when comparing between methods on limited data, but on metrics for the factorial design. Our comparative experiment is split into results containing all answers and those which remove true negatives, since the latter tends to be less relevant in legal practice. We find that, whilst QWEN outperforms DeBERTa on all questions by 9% (6%), it still exhibits the same distribution across them - performance deteriorates with more technical questions regardless of the inclusion of true negatives. In our 2x2 factorial experiment, we notably find that, whilst prompting significantly increases performance, augmentation seems to perpetuate pre-existing tendencies.

References

1. Dan Hendrycks, Collin Burns, Anya Chen, and Spencer Ball. Cuad: An expert-annotated nlp dataset for legal contract review. *arXiv preprint arXiv:2103.06268*, 2021.
2. Jules White, Quchen Fu, Sam Hays, Michael Sandborn, Carlos Olea, Henry Gilbert, Ashraf Elnashar, Jesse Spencer-Smith, and Douglas C Schmidt. A prompt pattern catalog to enhance prompt engineering with chatgpt. *arXiv preprint arXiv:2302.11382*, 2023.