Poster ID: 10

Rumesh Weerakoon, 2nd Year Master's Student
Supervisor: Prof. Incheon Paik
Intelligent Data Analytics Lab, The University of Aizu

Poster Session at Postgraduate Forum

Simplifying Legal Verbiage: Improving Legal Corpus Communication through Ontology-assisted Language Models

Overview

- Legal documents are often complex and difficult for non-specialists to understand.
- This creates barriers to legal access, complicating interactions between lawyers, clients, and the public.
- Simplifying this language can enhance comprehension, reduce misunderstandings, and promote access to justice.

Research Importance

- Addressing complexity in legal language
- Enhancing understanding of legal documents
- Advancing legal technology
- Comprehensive legal knowledge representation
- Bridging Legal-NLP gap for practical applications

Ordinary Documents

- Created for everyday purposes
- Use everyday language
- Flexible in structure
- Do not carry significant legal consequences (Do not carry legal significance)

vs

Legal Documents

- Created for special purposes
- Utilize specialized language
- Follow a structured format
- Have legal implications and consequences (Carry legal significance)

Application

- **Legal document processing**: automated classification, summarization, and extraction of key information
- **Legal chatbots and assistance**: offer precise, context-aware responses
- **Legal compliance monitoring**: identify risks or potential violations
- **Translation for legal language**: precise translation conserving legal meaning

Research Challenges

The challenges can be both technical and domain-specific due to the nature of legal language and the complexity of NLP tasks.

Complexity of Legal Language: Ambiguity and Vagueness, Highly Structured Jargon.

Ontology Design and Integration: Domain Adaptation, Dynamic Nature of Law

Preserving Legal Accuracy: Balancing Simplification with Precision, Semantic Loss.

Technical Challenges in NLP: Contextual Understanding, Handling Complex Sentence Structures.

Approach

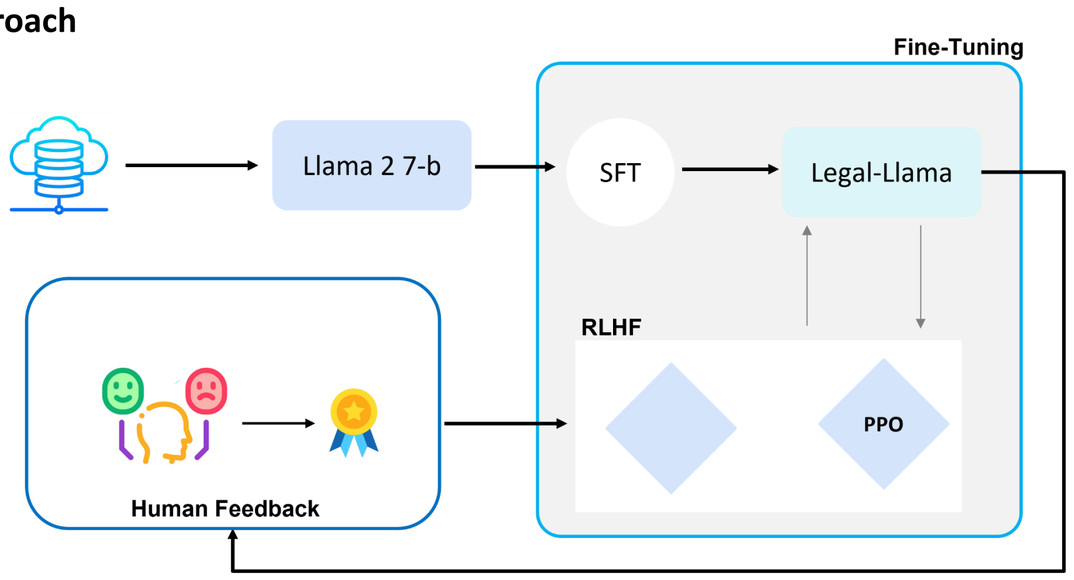

Figure 1: Methodological approach of Proximal Policy Optimization (PPO)

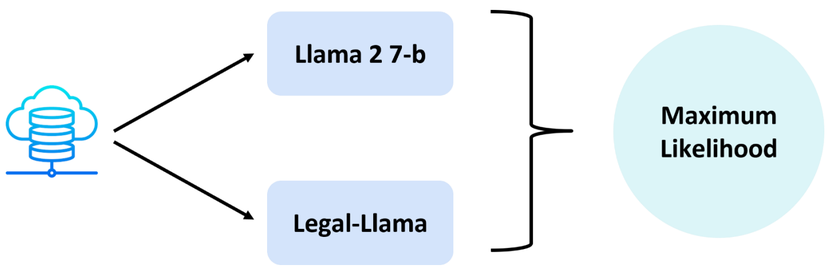

Figure 2: Methodological approach of Direct Preference Optimization (DPO)

Table 1: Detailed methodological approach

Step	Approach
Ontology construction	Develop a comprehensive ontology for the specified legal domain.
Data Collection and Preprocessing	Gather a dataset of legal texts. Preprocess the data, including cleaning, tokenization, and structuring.
Initial Pretraining	Pretrain the LLM on data using a language modeling objective to provide a basic understanding of legal language patterns.
Fine-Tuning	Fine tune with Direct Preference Optimization (DPO). Base model with Legal data + simplified legal corpus. Base model with ontology assisted legal corpus.
Evaluation	Evaluate the final fine-tuned model and its performance.

Current Status

1. Evaluated the base model (Llama2-7b) with the legal corpus
2. Finetuned the base model with legal text and simplified corpus
3. Evaluated the Finetuned model

Results

Table 2 : Evaluation: base model with original legal corpus

Steps	Training Loss	Validation Loss
200	2.203	2.1782
400	2.1447	2.1700
600	2.1292	2.1642
800	2.1653	2.1615
1000	2.1479	2.1595

Table 3: Evaluation: base model with simplified legal corpus

Steps	Training Loss	Validation Loss
200	1.8571	1.8469
400	1.8196	1.8125
600	1.7887	1.7709
800	1.7494	1.7247
1000	1.7256	1.7187

Table 4 : Evaluation metric scores

Evaluation metric	Base Model	Fine-tuned Model
BLEU	0.1046	0.1528
ROUGH-1	0.353	0.4049
ROUGH-2	0.2181	0.2487
ROUGH-L	0.3432	0.3891

Proposed metrics for the final model evaluation

- Summarization Performance: ROUGE-1, ROUGE-2, ROUGE-L
- Performance Evaluation: Accuracy, Precision, Recall, F1 Score
- Language Modeling: Perplexity Measurement
- Evaluating automatic text simplification: SARI metric

Generated prompts for evaluation

Base model with original text

prompt: Explain the term Incapacitation in the field of Criminal Law.

The term incapacitation is used in the field of criminal law to describe the process by which a person who is convicted of a crime is no longer able to commit a crime. Incapacitation is typically used as a form of punishment for those who have committed serious crimes. There are several different types of incapacitation that can be used as a form of punishment. The most common type of incapacitation is incarceration. Incarceration is when a person is placed in a prison or other type of correctional facility. Incarceration is typically used as a form of punishment for those who

Finetuned model with original text and simplified text

prompt: Explain the term Incapacitation in the field of Criminal Law.

The term incapacitation is used in the field of Criminal Law to describe the condition of a person who is unable to commit a crime. Incapacitation may be accomplished by imprisonment, or by mental or physical incapacity. The term is also used to describe the condition of a person who is unable to commit a crime due to age. For example, a child is incapacitated from committing a crime because he or she is under the age of criminal responsibility. In the United States, incapacitation is a requirement for the death penalty. In the United States, the death penalty is reserved for those who are incap

Future Works

1. Finalize the ontology annotation
2. Use a more diverse set of legal corpus, covering various domains (contracts, patents) to assess the base model's generalization ability
3. Finetune the model with ontology assisted and simplified legal corpus
4. Fine-tune the model separately for specific legal subdomains (contracts vs. case law)
5. Evaluate the model performance

References

- Mumford, J., Atkinson, K., & Bench-Capon, T. (2023). Combining a Legal Knowledge Model with Machine Learning for Reasoning with Legal Cases. In Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law (pp. 167–176). Association for Computing Machinery.
- D. Song, S. Gao, B. He and F. Schilder, "On the Effectiveness of Pre-Trained Language Models for Legal Natural Language Processing: An Empirical Study," in IEEE Access, vol. 10, pp. 75835-75858, 2022, doi: 10.1109/ACCESS.2022.3190408.
- R. Qin, M. Huang and Y. Luo, "A Comparison Study of Pre-trained Language Models for Chinese Legal Document Classification," 2022 5th International Conference on Artificial Intelligence and Big Data (ICAIBD), Chengdu, China, 2022, pp. 444-449, doi: 10.1109/ICAIBD55127.2022.9820466.
- Y. H. Gu, X. Piao, H. Yin, D. Jin, R. Zheng and S. J. Yoo, "Domain-Specific Language Model Pre-Training for Korean Tax Law Classification," in IEEE Access, vol. 10, pp. 46342-46353, 2022, doi: 10.1109/ACCESS.2022.3164098