Building LLMs for Legal Reasoning

Elven Shum
CPSC490 Project
Advisors: Ruzica Piskac, Scott Shaprio, Timos Antonopoulos
elven.shum@yale.edu

1 Introduction

1.1 Overarching Motivation:

Large Language Models (LLMs) have emerged as powerful tools across various domains, including the legal sector, where they promise to enhance and streamline legal processes. However, despite their utility in basic legal tasks, these models often struggle with more intricate legal reasoning that involves specialized terminologies and complex legal principles. Trained primarily on general language corpora, conventional LLMs tend to underperform when faced with texts dense in legalese, revealing a gap in their capacity to interpret and apply nuanced legal constructs. [5]

Our project seeks to address this gap by not merely fine-tuning an LLM for improved legal reasoning, but by strategically utilizing prompt engineering and chain-of-thought reasoning to elicit more accurate and contextually grounded outputs. Rather than solely focusing on fine-tuning, we emphasize techniques that refine the model's interpretative capabilities—pushing the boundaries of how LLMs process legal predicates and reach conclusions. Collaborating with Yale Law School professor Scott Shapiro and Yale Computer Science professor Ruzica Piskac, we concentrate on proximate cause cases, as detailed in Section 3.1, to develop a robust framework that enhances the model's ability to handle sophisticated legal reasoning tasks. This approach aims not just for better performance, but for more meaningful decision-making, where the model's reasoning mirrors the critical thinking demanded in real-world legal contexts.

1.2 Content of Report, Distinction Between 488 and 490:

Please note this project is separated into two parts, divvied up between CPSC488's final project and CPSC490. The CPSC488 component of this project consists of the following components: Scraped Raw Legal Case Data, Data Optimization Scripts, Optimized Proximate Cause Dataset. The CPSC490 component consists of (1) the Fine-Tuned Model, (2) Fine-Tuning Scripts + Chain of Thought Prompts, (4) Automated Evaluation Script (3) Real World Case Verdict Prediction Evaluation, (4) Proximate Cause Reasoning Evaluation. As will be discussed in later sections, these evaluations are done by both an LLM and by the author, who is not legally trained. Proper holistic evaluation requires professional evaluation from a legal scholar. Please note that section 3.1-3.3 are intended to be in the CPSC488 report, but are included in this CPSC490 report for better context and continuity; to not offend university dual submission requirements, sections 3.1-3.3 are not considered part of the CPSC490 component, and are merely for reference.

2 Related Work

Recent advances in integrating LLMs with legal reasoning have led to the development of several benchmarks and models. LegalBench remains the most comprehensive benchmark, testing LLMs on tasks ranging from issue-spotting to rule application, offering a thorough evaluation of models' performance in practical legal scenarios [5]. Complementing this, LawLLM was developed specifically for

U.S. legal tasks, improving the prediction of verdicts, case retrieval, and precedent recommendation through advanced fine-tuning techniques [4].

ADAPT, a prompting framework, enhances legal judgment prediction by focusing on multi-task instruction tuning. This framework tackles limitations in traditional LLMs by refining models through specific legal datasets, thus increasing reasoning accuracy across complex legal charges [3]. Similarly, Chain of Logic has demonstrated improved performance over standard methods by applying a rule-based approach to tasks like jurisdiction and contract interpretation, further showcasing the role of specialized prompting techniques in legal applications [1].

Beyond the U.S., DISC-LawLLM focuses on legal reasoning for the Chinese legal system, utilizing fine-tuned models to navigate local laws and legal documents. This highlights the growing importance of adapting LLMs for jurisdiction-specific contexts [2]. Additionally, LegalGPT introduces a multi-agent framework for legal reasoning, leveraging chain-of-thought reasoning to enhance interpretative capabilities across diverse legal tasks [6].

As the field evolves, domain-specific models such as SaulLM are showing superior performance compared to general-purpose models like GPT-4. These legal-specific LLMs provide deeper contextual understanding and higher accuracy in tasks like document review and verdict prediction, underscoring the potential of tailored models in the legal sector [7].

3 Methodology

3.1 Dataset Selection:

Our project revolves around "Proximate Cause" cases in law. These cases are emblematic of systematic legal reasoning, use explicit categorizations of necessary types of legal evidence, and are minimally semantically interprative. Other legal domains, for example constitutional law cases, would be ill suited to applying LLMs, and they're almost entirely devoted to interpreting particular words in a modern context. The systematic nature of proximate cause cases, coupled with their relatively concise length, renders them apt for the operational scope of LLMs. The structured methodology employed in teaching these legal concepts to law students is rather identical; typically, students are tossed a plethora of cases from a large book and expected to determine the abstract reasoning. This further underscores the feasibility of instilling a refined understanding of these concepts in LLMs through fine-tuning. In the USA, different States, while internally consistent, each possess slightly varying laws regarding their handling of Proximate Cause cases. Therefore, we select our cases from only the state of Texas. Furthermore, in order to get a wide variety of Proximate Cause cases, we select from several intervening causes: Lightning, Disease, Criminal acts, Rescue, Suicide, Medical treatment. These keywords were suggested by Professor of Jurisprudence, Scott Shapiro.

3.2 Dataset Acquisition and Cleaning:

For exhaustive dataset, we leveraged the resources of the Yale Law School library, courtesy of Professor Scott Shapiro. Our data acquisition was facilitated through YLS via LexisNexis, a legal database provider. Legal case repositories are not typical pre-filtered or categorized. Therefore 3rd party Legal Databases, where the categorization has already been completed, is an indispensable tool for legal research, like ours. Cases retrieved from LexisNexis are only in Rich Text Format (RTF), laden with unnecessary metadata and formatting elements.

To address this, we developed a data cleaning script, primarily focused on two tasks: converting RTF files to a more manageable plain text (txt) format and excising irrelevant metadata,d formatting, headers, and footers. The resultant text files were then formatted to align with the requirements of our Data Optimizer (discussed in the subsequent section). This streamlined dataset, devoid of extraneous information, ensured that the data fed into the model was of consistent quality.

3.3 Dataset Optimization:

To enhance the proficiency of a Language Model (LM) in not only comprehending legal reasoning but also in deducing verdicts from case facts, our research focused on fine-tuning the model using Proximate Cause legal reasoning cases. It became evident that the utilization of raw case data was suboptimal for several reasons. Primarily, these cases often exceed the limits of many context

windows due to their length. Additionally, they contain extraneous elements such as judge names and case references. Most critically, the ingestion of raw cases in their entirety might not effectively train the LM in legal reasoning; it could instead result in a superficial replication of verdict-like responses. The objective was to enable the LM to internalize the underlying logic and reasoning of the cases, necessitating an understanding of the Legal Reasoning Rule and the ability to distill relevant facts from a case for rule application.

To foster this advanced learning, significant Dataset Optimization was imperative. We employed a preprocessing LM, ChatGPT-4, to transform the text of raw cases into structured (Background, Verdict) pairs. Recognizing that legal cases inherently comprise both background facts and judicial rulings, our training approach differed; for the fine-tuned LM, we provided only the background, tasking it with verdict prediction. This necessitated the creation of (Background, Verdict) pairs specifically for training, which were then used to prompt the LM to generate verdicts based solely on case backgrounds. These pairs were designed with key attributes in mind: brevity compared to the full-length cases, inclusion of only factual elements in the Background, and ensuring that the Verdict not only elucidated the reasoning but also exclusively referenced facts presented in the Background. Through iterative prompt engineering, we identified the phrase *"Please summarize the following case in the form { Background: [background of the case], Verdict: [verdict of the case] }. In Background, do not include the cases's outcome, only facts. Be extremely detailed in the Background, as any facts could be relevant. In Verdict, explain reasoning for the verdict in future tense, as if you were predicting. Do not cite facts in the Verdict that were not in Background. Be very detailed in Verdict."* that yielded best results.

3.4 Fine-tuning Methodology: OpenAI api

To fine-tune the model on proximate cause legal reasoning, we utilized the OpenAI API. The decision to use this API was driven by several factors. First, our goal was to produce the most accurate legal reasoning model possible, specifically targeting proximate cause cases. Using state-of-the-art (SOTA) models was crucial to ensure the highest chance of capturing the nuances in legal reasoning. Open-source SOTA models were considered; however, the computational costs associated with self-hosting these models were prohibitive, given our hardware limitations. Thus, the OpenAI API, particularly the GPT-4o model, was selected for fine-tuning due to its robust performance in non-mathematical reasoning tasks.

We fine-tuned the model on a dataset of 440 legal cases related to proximate cause. The dataset was split into 95% for training and 5% for evaluation, resulting in 22 evaluation cases. This split was designed to test the model's ability to generalize legal reasoning, ensuring that it could pick up the subtle complexities of proximate cause reasoning in unseen cases. The proof of concept aimed to determine whether such models could accurately replicate legal reasoning with minimal data.

3.4.1 Chain of Thought System Prompt

We crafted a Chain of Thought (CoT) system prompt to structure the model's reasoning during fine-tuning. It's intentionally rather general to encourage the model to run through a step-by-step analysis of legal facts and the principles of proximate cause, while not restricting it on-high to a proximate cause framework. Ideally, the model learns from the training cases *how* to reason about proximate cause, as opposed to directly telling it. In this way, it may be able to capture more general principles than what a rigid CoT prompt would've permitted. See section 5 for discussion on potential creation of a more ideal prompt. The following prompt was used:

> *You are an expert in law with a focus on reasoning about proximate cause. Your goal is to carefully analyze the background of legal cases and determine the correct verdict using step-by-step legal reasoning.*
>
> *To reach a conclusion, follow these steps:*
> *1. Identify all relevant facts from the background.*
> *2. Analyze these facts in the context of proximate cause, highlighting key points of causation and foreseeability.*
> *3. Consider whether the defendant's actions could have reasonably led to the plaintiff's harm.*

4. Discuss any potential intervening causes, ambiguities, or contributing factors.
5. Synthesize these considerations into a coherent argument, and clearly state the final verdict in the last line of your reasoning. The verdict must be presented without additional quotes or formatting.

3.4.2 Fine-tuning Setup

In our setup, the case backgrounds were presented to the model as the user input, while the verdict was framed as the assistant response. This structuring choice reflects the natural division between the legal facts (input) and the legal decision (output), closely mirroring real-world legal reasoning workflows. By using this approach, the model learns to process background information, perform legal analysis, and predict a verdict, effectively simulating the role of a legal expert. This approach, combined with the CoT prompt, ensured the model was exposed to a clear, step-by-step legal reasoning process during training. We trained for 3 epochs.

4 Results and Evaluation

4.1 Evaluation Setup

Our evaluation was designed to assess two core objectives: (1) the accuracy of the fine-tuned model's verdict predictions in proximate cause cases, and (2) the quality of its legal reasoning compared to the base model, GPT-4o. We benchmarked the fine-tuned model against GPT-4o using a set of 22 evaluation cases, which were randomly selected from our dataset. These cases were chosen to test whether the fine-tuned model could effectively generalize legal reasoning to unseen cases.

To evaluate both objectives, we implemented the following approaches:

1. **Verdict Prediction:** We compared the verdicts predicted by GPT-4o and the fine-tuned model to the ground truth verdicts, which were based on real-life case outcomes.
2. **Reasoning Quality:** We used GPT-4o to evaluate the reasoning provided by both models, determining which model's reasoning best aligned with the legal standards of proximate cause.

While comprehensive legal evaluation requires input from trained legal professionals, this proof of concept aims to demonstrate that accurate legal reasoning can be learned. With an intentionally limited evaluation set of 22 cases, we focused on whether the model could capture the core principles of proximate cause reasoning, despite the absence of expert reviewers.

4.2 Evaluation Process

We automated the evaluation process by comparing both models' verdicts to the ground truth. First, we assessed whether each model's predicted verdict agreed with the ground truth. For this, we utilized GPT-4o to determine if the predicted and actual verdicts were in agreement. Next, we compared the reasoning quality between the models. GPT-4o was used to evaluate three verdicts per case: the ground truth, GPT-4o's verdict, and the fine-tuned model's verdict. The model then judged which of the two predictions (GPT-4o or fine-tuned) best aligned with the reasoning in the ground truth verdict.

This process generated a CSV file containing detailed results, including verdict agreement and reasoning quality. This file serves as the primary deliverable for comparing the models' performance across all evaluation cases.

Subtle adjustments were made to ensure that the base model's system prompt provided minimal guidance, whereas the fine-tuned model's prompt encouraged a structured, step-by-step analysis. This allowed us to maintain consistency while still evaluating the effects of fine-tuning on legal reasoning.

4.3 Results

In our evaluation of 22 cases, we observed a 20% increase in verdict accuracy with the fine-tuned model compared to the base GPT-4o model. Specifically, the fine-tuned model correctly predicted verdicts in 18 cases, while the base model succeeded in 14. Though the dataset is small, the results

indicate that fine-tuning increases the likelihood of reaching the correct verdict in real-world legal cases.

More importantly, the fine-tuned model demonstrated a significantly improved ability to reason through the complexities of proximate cause. In head-to-head comparisons of legal reasoning, our model aligned with real-life court reasoning in 15 of the cases, while GPT-4o matched in 7. This difference highlights the fine-tuned model's enhanced capacity to handle complex legal concepts and doctrines. The fine-tuned model can be accessed via the OpenAi API as model:

```
ft:gpt-4o-2024-08-06:yale::AAYnVGq6
```

4.4 Case Study: Verdict Comparison

To illustrate the effectiveness of our fine-tuned model, we compare the verdicts generated by the base GPT-4 model and our fine-tuned model against the ground truth verdict of a legal case. Table **??** presents selected quotations from each verdict, highlighting the similarities between the ground truth and the fine-tuned model.

Table 1: Comparison of Verdicts (Eval Case 20)

Key Point	**Ground Truth Verdict**	**Base Model Verdict**	**Fine-Tuned Model Verdict**
Applicability of Chapter 95	Weekley did not conclusively show that the Chapter 95 protections were applicable.	Unless appellants can produce compelling evidence, the initial motion for summary judgment by Weekley Homes might stand.	The appellate court will reverse the trial court's judgment and remand the case based on the existence of genuine issues of material fact regarding both prongs of Chapter 95.
Error in Granting Summary Judgment	The trial court's granting of the traditional summary judgment based on Chapter 95 was erroneous.	The trial court likely granted summary judgment assuming lack of substantial evidence.	The decision will be based on the existence of genuine issues of material fact.
Negligence and Premises Liability Claims	Appellants presented more than a scintilla of evidence demonstrating a duty owed and the possibility of a nexus between their injuries and the control of their working conditions by Weekley.	Without clear evidence showing Weekley's direct control, appellants may struggle to prove control.	Evidence could show that Weekley had control over the timing and conditions of appellants' work.
Denial of Motion for Reconsideration	Denial of appellants' motion for reconsideration and new trial was an abuse of discretion by the trial court.	Unless appellants can produce compelling evidence, the trial court's summary judgment might stand.	The trial court's denial of the motion for reconsideration was an erroneous decision warranting reversal.
Gross Negligence Claim	Appellants failed to provide substantive evidence regarding Weekley's actual, subjective awareness of risk.	Appellants would need robust evidence proving Weekley's actual knowledge.	The court will likely affirm the trial court's verdict on the gross negligence claim.
Final Judgment	The appellate court affirmed the trial court's verdict on the gross negligence claim, reversed the verdict on the negligence and premises liability claims, and remanded those parts back to the trial court.	Thus, if we assume appellants could not conclusively show these elements, the summary judgment would be upheld.	The appellate court will reverse the trial court's judgment and remand the case for further proceedings.

The fine-tuned model's verdict closely mirrors the ground truth in both outcome and legal reasoning, demonstrating its enhanced capability to understand and apply complex legal principles. Both the ground truth and the fine-tuned verdict recognize that Weekley didn't conclusively establish the applicability of Chapter 95 protections due to genuine issues of material fact concerning control and knowledge of the dangerous condition. Also, fine-tuned model accurately identifies the trial court's error in granting summary judgment on the negligence and premises liability claims, therefore the appellants presented sufficient evidence to raise genuine issues for trial. It also concurs with the ground truth in affirming the trial court's verdict on the gross negligence claim: lack of substantive evidence regarding Weekley's actual awareness of risk.

4.5 Case Study 2: Verdict Comparison

To further demonstrate the effectiveness of our fine-tuned model, we present a second case study comparing the verdicts generated by the base GPT-4o model and our fine-tuned model against the ground truth verdict. Table 2 similarly showcases selected quotations from each verdict.

Table 2: Comparison of Verdicts (Eval Case 14)

Key Point	**Ground Truth Verdict**	**Base Model Verdict**	**Fine-Tuned Model Verdict**
Affirmation of Judgment	The trial court's judgment was upheld.	The appellate court should consider remanding for a new trial.	The appeals court would likely affirm the trial court's rulings.
Limitation of TDHS Records	Limitation of TDHS records to those concerning Watson was within the court's discretion.	The trial court excluded these records, likely to prevent prejudice.	The court will opine that any error in excluding other staff supplements was harmless.
Exclusion of Dr. Cox's Testimony	Restrictions on the admissibility of Dr. James Cox's testimony were deemed non-arbitrary as his conclusions were speculative.	The restriction on Dr. James Cox's testimony could imply the trial court unnecessarily limited crucial evidence.	The court will conclude that the exclusion of certain expert testimonies, including those involving Dr. Cox, did not likely cause harm to appellants.
Confidentiality of TDHS Photographs	The TDHS photographs were considered confidential and not subject to disclosure.	If the photographs were highly relevant and non-prejudicial, their exclusion might constitute an error.	The court will find that appellants did not preserve their complaint concerning TDHS photographs for appeal.
Negligence Per Se Claims	Negligence per se claims were correctly struck as the cited laws did not provide for tort liability.	If the trial court erred in this determination, it may affect the basis upon which the jury considered negligence.	The court will argue that appellants lack standing because the NHRA does not confer an independent private right of action.
Overruling Appellants' Issues	The appellants' issues were overruled in their entirety, and the take nothing judgment against them was affirmed.	The appellate court should consider remanding for a new trial.	Predicting the outcome, the court will express that all points of error by appellants are overruled and the trial court's judgment is affirmed.

Once again, the fine-tuned model's verdict demonstrates a remarkable alignment with the ground truth, both in the final judgment and the underlying legal reasoning. It accurately reflects the appellate court's decision to affirm the trial court's rulings, where by capturing the nuances of the court's discretion and the application of legal doctrines. The fine-tuned model recognizes that any error in excluding additional TDHS records was deemed harmless, which mirrors the ground truth's emphasis on the court's discretionary power and the lack of substantial prejudice resulting from such exclusions.

Also, the fine-tuned model correctly identifies the appellants lack of standing to bring negligence per se claims, aligning with the ground truth's reasoning that the cited statutes did not confer an independent private right of action. It also acknowledges procedural intricacies, such as the appellants' failure to preserve certain complaints for appeal and the insufficiency of objections during the trial, which the ground truth verdict highlights as critical factors in the appellate court's decision.

4.6 Results Conclusion

The fine-tuned model shows significant progress in handling complex legal reasoning tasks, particularly in predicting verdicts for proximate cause cases. Its ability to produce verdicts that closely align with the ground truth demonstrates a promising understanding of legal principles. For instance, the model's handling of intervening causes like "acts of God" and foreseeability mirrors the nuanced reasoning observed in actual court decisions. In one case, the model's verdict on the applicability of Chapter 95 protections accurately captured the key issue of control, while in another, it correctly identified the trial court's error in granting summary judgment on negligence claims—both examples highlighting the model's deepened understanding of proximate cause.

Our approach combining a broad Chain of Thought (CoT) prompt with targeted fine-tuning has proven effective in fostering structured, step-by-step legal reasoning. Most compellingly, the fine-tuning process allowed the model to internalize legal primitives such as causation and foreseeability, which are fundamental to proximate cause cases. The success of these techniques suggests that the model is not just replicating outcomes, but genuinely learning to apply core legal concepts. This opens up exciting possibilities for future work in expanding its capabilities across different legal domains.

5 Discussion

Our results suggest that the fine-tuned model has taken significant steps toward successful legal reasoning in proximate cause cases. The fine-tuning process, coupled with a broad Chain of Thought (CoT) system prompt, allowed the model to learn key reasoning primitives, such as causation and foreseeability. This facilitated the model's ability to predict verdicts with reasoning that aligns closely with the ground truth in many instances. However, the model's verdict success rate, while improved over the baseline GPT-4o, remains modest, and its performance could be bolstered with further refinements and more data.

One of the key limitations of this work is the relatively small dataset of 450 proximate cause cases. Although this was sufficient for the model to internalize fundamental legal reasoning structures, it likely missed more nuanced or esoteric legal principles that emerge in particularly intricate cases. Additionally, while the fine-tuned model outperformed the baseline in reasoning tasks, its overall verdict accuracy was only slightly better, demonstrating that there is still room for substantial improvement. Furthermore, GPT-4o is already a very strong reasoner, so we are starting from a high baseline. Future efforts should aim to push the model to a level where it can more consistently predict correct outcomes across a wider variety of cases.

In terms of future work, while our broad CoT approach successfully encouraged structured reasoning without explicitly defining proximate cause rules, there is a case to be made for including more domain-specific legal reasoning rules directly in the CoT system prompt. A well-structured prompt based on actual legal reasoning frameworks might yield short-term improvements in verdict prediction, particularly if the model had more cases to train on. We intentionally avoided this due to the limited scope of our evaluation cases and the desire to let the model learn these rules autonomously. However, with a larger dataset, incorporating a more structured legal framework might be an avenue worth exploring.

Additionally, one of the largest areas for improvement lies in the "Case Optimization" process. Our data was optimized using base GPT-4, rather than more advanced models like GPT-4o or newer reasoning-optimized models from OpenAI. With stronger models, it's likely that the case summaries and reasoning processes could be much more robust, leading to better downstream performance in legal reasoning tasks. We recommend exploring new model architectures to generate more detailed reasoning tokens, which could then be summarized by another model. This iterative reasoning and summarization approach might unlock more sophisticated legal analysis. Unfortunately, we were

constrained by the release timeline of these newer models, but refining this data optimization step should be the focus of future improvements.

From a broader perspective, this work demonstrates the potential for AI systems to engage in serious legal reasoning. While current state-of-the-art (SOTA) models have not been explicitly trained for this purpose, our fine-tuning approach shows that it is possible to tailor models for legal tasks like proximate cause reasoning. In the future, such systems could significantly enhance the efficiency of legal professionals by generating structured analyses as a first pass, freeing up lawyers to focus on higher-order decision-making rather than tedious case review.

Finally, I would like to extend my sincere thanks to my advisors and mentors: Professors Ruzica Piskac, Scott Shapiro, and Timos Antonopoulos. Their unrelenting patience and guidance throughout this process were invaluable, especially as my health issues caused significant delays in the completion of this project. I am grateful for the opportunity to be involved in their research, and for their support in shaping this idea. Additionally, I would like to thank my Yale College Dean, Ployd, for his encouragement and patience throughout this journey.

References

[1] Chain of Logic: Rule-Based Reasoning with Large Language Models. Accessed: 2024-09-23.

[2] DISC-LawLLM: Fine-Tuning Large Language Models for Intelligent Legal Services. Accessed: 2024-09-23.

[3] Enabling Discriminative Reasoning in LLMs for Legal Judgment Prediction. Accessed: 2024-09-23.

[4] LawLLM: Law Large Language Model for the US Legal System. Accessed: 2024-09-23.

[5] LegalBench: A Collaboratively Built Benchmark for Measuring Legal Reasoning in Large Language Models. Accessed: 2023-10-16.

[6] LegalGPT: Legal Chain of Thought for the Legal Large Language Model Multi-agent Framework. Accessed: 2024-09-23.

[7] SaulLM-7B: A Pioneering Large Language Model for Law. Accessed: 2024-09-23.