LLM Evaluation: Democratizing Legal Assistance

Ruite (Harry) Guo, Alton Lee, Alyssa Simmons, Sarah Wilen
University of California, Los Angeles
{ruiteg,altonrun4ever,lsssmmns,swilen1}@g.ucla.edu

Abstract

Legal counsel tends to be prohibitively expensive. In 2019, the State Bar found that even when experiencing problems significantly impacting them, less than 1 in 3 Californians sought legal assistance. Large Language Models, such as Open AI's ChatGPT, may be able to democratize access to legal information. However, the reliability of LLMs in delivering legal advice remains relatively unexplored, with hallucinations and inaccuracies from chatbots being dangerous for individuals traversing the already confusing legal system. The goal of our study is to assess whether different LLMs can accurately answer typical legal questions relating to California litigation and also to determine if these answers are presented in an interpretable manner for individuals who do not have a legal background.

The evaluation framework used two datasets: (1) a basic Q&A dataset, which contains 66 real-world legal questions asked by the public and answers from lawyers, and (2) an augmented dataset, which highlights the specific California codes guiding the lawyer's answers, which are verified through a combo of LLM and manual human review. The performance of ChatGPT-3.5 and Claude-3 was evaluated based on automated and manual accuracy scores and interpretability scores.

We found that both ChatGPT and Claude performed similarly in terms of semantic similarity and entailment. However, Claude slightly outperformed ChatGPT in human-annotated accuracy for identifying relevant California statutes and demonstrated better interpretability.

This project motivates the potential of LLMs to enhance access to legal information. By evaluating the performance of LLMs in providing legal advice, we aim to understand LLM's role in increasing access to legal counsel for the public.

Disclaimer

LLMs may have been used in creating this report for clarity and grammatical correctness purposes only. All written ideas are original thoughts.

1 Introduction

Legal counsel tends to be prohibitively expensive. For example, typical civil harassment restraining order attorneys in California charge an average of $750/hour, making counsel inaccessible to many. Moreover, self-representation may be daunting because mistakes in legal matters can lead to dire consequences such as large fines or the dismissal of cases. In light of these challenges, there is a pressing need for affordable and accessible legal assistance. Large Language Models (LLMs), such as the Open AI's ChatGPT, may be able to bridge the gap between laypeople and affordable legal assistance by serving as a resource in providing accurate and interpretable legal advice. These models can democratize access to legal information and empower individuals to navigate legal challenges without incurring significant costs. However, the reliability of LLMs in delivering legal advice remains relatively unexplored.

Our goal is to assess whether different LLMs, specifically Open AI's ChatGPT and Anthropic's Claude, can accurately answer typical legal questions relating to California litigation and also to determine if these answers are presented in an interpretable manner for individuals who do not have a legal background. By evaluating the performance of LLMs in providing legal advice, we aim to understand their potential role in increasing access to legal counsel for laypeople.

To tackle this problem, we designed an evaluation framework using two distinct datasets: a basic Q&A dataset and an augmented Q&A dataset. The basic dataset comprises 66 real-world legal questions and answers sourced from the website

JUSTIA, covering four key legal areas: Civil, Employment, Intellectual Property, and Criminal law. We additionally developed an augmented dataset that includes specific California codes or statutes guiding the lawyer's answers, which was verified through a combination of LLM and manual human review.

Our primary contribution is the systematic evaluation of LLMs in the context of California legal Q&A. Through our experiments, we found that both ChatGPT-3.5 and Claude-3 perform similarly in terms of semantic similarity (BERTScore) and entailment (BLEURT), with Claude slightly outperforming ChatGPT in human-annotated accuracy for identifying relevant California statutes. Furthermore, Claude demonstrated better performance in terms of interpretability, achieving higher Flesch Reading Ease and lower Flesch-Kincaid Grade Level scores, which indicates more readable and accessible generated answers for lay users.

Overall, our research highlights the potential of LLMs to enhance access to legal information while also identifying areas for improvement, such as the need for better handling of complex and context-rich queries. This work lays the foundation for using Large Language Models to make legal knowledge more accessible and reliable for the general public!

2 Related Work

2.1 LLMs in Law

ChatGPT-3.5 ChatGPT has been analyzed as a possible legal resource in many different papers. Its ability to pass the United States Bar Exam (both providing correct answers and explanations) was explored in a 2022 study [Bommarito II and Katz, 2022] which found that it could pass in two of the six areas of the test and their 2nd and 3rd most probable answers were 71% and 88% correct. This shows that ChatGPT should be advanced enough to not only understand legal questions but also reply with logical explanations, which is one of the reasons we chose it as one of the LLMs to analyze. Another paper analyzed ChatGPT's ability to fill in instruction-following responses concerning their own LexGLUE Dataset using SentenceBERT to consider the similarity between the LLM's response and the expected output [Chalkidis, 2023]. Similarly, we have chosen BERTScore (another entailment method) to analyze the LLM's response to our unique dataset.

ChatGPT can undertake legal research, as it has the facility to retrieve and analyze legal information (including legal cases and legislation) and summarize research. In February 2023, Mishcon de Reya advertised for a "GPT Legal Prompt Engineer" to support the firm in incorporating natural language models into its practice [Reya, 2023]. There are also models of AI designed specifically for use in individual law firms. Allen & Overy announced in February 2023 that it was integrating Harvey into its practice. Harvey uses GPT AI technology developed by OpenAI through a collaboration with lawyers, technologists, and entrepreneurs [Wakeling, 2023]. The platform can automate and assist with elements of legal work including contract analysis, insights, research, and creation of legal documents. Harvey is working with other law firms so we should expect to see further announcements on its adoption in other practices soon [Grady and Curnin, 2023]. It is expected that AI systems comparable to ChatGPT will be integrated into legal practice within a relatively short period of time [Ajevski et al., 2023].

Claude-3 Claude is not as highly researched as ChatGPT given that Anthropic released Claude-1 in March of 2023. However, [Sargeant et al., 2024] utilizes Claude to classify summary judgment cases by topic, building on previous papers that use various topic model methodologies such as Latent Dirichlet Allocation (LDA) for the analysis of legal texts across various jurisdictions.

2.2 Evaluation Methods

2.2.1 Accuracy

BERTScore BERTScore [Zhang et al., 2019, Özbolat] is an automatic evaluation metric for text generation. Unlike other metrics using the n-gram matching approach, they proposed utilizing BERT contextual embedding to calculate the cosine similarity between two texts.

BLEURT Entailment Score BLEURT [Sellam et al., 2020] is also a learned metric on synthetic samples based on BERT. The model would compute the probability of three labels: Entail, Contradict, and Neutral on the textual entailment from reference text to generated text.

GPTScore GPTScore [Fu et al., 2023] is

a metric trained by a neural network based on 33 datasets and 22 evaluation aspects. They proposed utilizing the emergent abilities of nineteen 19 generated pre-trained models on the datasets in different evaluation aspects and it aimed to evaluate the comprehensive quality of the generated text.

2.2.2 Interpretability

Flesch Reading Ease and Flesch-Kincaid Grade Level The Flesch Reading Ease score [Flesch, 1948] is a statistical formula developed to be used as a tool in the selection of reading materials in adult education to fit the reading capacities of different audiences. It has since been used in a multitude of applications such as evaluating training materials for the Navy [Kincaid et al., 1975] where these scores were then correlated to education level. It measures the semantic and syntactic difficulty of a piece of literature. As such this is useful in analyzing the readability of LLM output concerning our goal of analyzing whether LLMs are a practical tool in increasing accessibility to legal counsel.

Lexical diversity The type/token ratio [Johnson, 1939] was built off of the idea of type (unique form; e.g. "the") and token (single event; e.g. the "the" appearing as the underlined word in this sentence being separate from *the* italicized "the") which was first introduced by Charles Sanders Peirce in 1906 [Peirce, 1906], and has been used to measure lexical diversity since its introduction in 1939 by Wendell Johnson. It measures lexical diversity by dividing the number of unique tokens (types) by the total number of tokens. This method has been used to analyze LLMs' ability to mimic the complexity of human speech [Muñoz-Ortiz et al., 2023, Herbold et al., 2023, Reviriego et al., 2023]. Likewise, our paper will utilize the type/token ratio to analyze the complexity of the legal responses provided by the LLMs to ensure the answers provided include enough detail to be useful.

Topic Coherence

3 Methodology

We will perform LLM evaluation on two datasets: (1) the basic Q&A dataset and (2) the augmented Q&A dataset.

3.1 Basic Q&A Dataset

Our goal is to evaluate how LLMs perform on legal questions asked by people who do not have legal backgrounds. Therefore, our reference dataset is made up of 66 samples of questions asked by laypeople and answers by professional lawyers. The questions and answers came from the website JUSTIA where anyone can ask a legal question and lawyers may answer them. We pulled these real-world questions from four legal areas: (1) Civil, (2) Employment, (3) Intellectual Property, and (4) Criminal. We decided to pull questions from different topics as the LLM could perform better on one topic than another. For example, there is a lot of nuance in Civil and Employment, whereas, Criminal law may be more straightforward and Intellectual Property law may be a combination of the two.

We describe our rubric for selecting questions in Appendix A.5.

After consolidating our dataset, we noticed that some of the lawyer answers seemed to be AI-generated. In fact, after inputting some of the responses into GPT-0, we noticed that a majority of the answers from some of the more responsive lawyers were AI-generated. For example, there were lawyers whose answers were 100% AI-generated. Therefore, to properly screen out AI-generated answers from the human-generated answers, we put all the answers into GPT-0 and removed the AI-generated answers.

To generate responses from the LLMs for the legal questions, we input verbatim the questions as they were originally asked. Some questions had grammar issues and may have been difficult to understand. We deliberately chose to include those because they sample the distribution of true questions that lay people who have a range of understanding of the law.

We then evaluated the responses to quantify each LLM's accuracy and interpretability such that we could discuss their potential as a legal aid.

3.2 Augmented Q&A Dataset

To supplement our basic Q&A dataset and aid in manual human review of the generated answers, we created an augmented Q&A dataset. We chose five Q&A samples from each of the four topics to "augment". Our augmented Q&A dataset includes the specific California code/statute that the lawyer's answer is guided by. The purpose of this augmented dataset is to use codes/statutes as an

objective metric to evaluate accuracy.

We obtained the reference CA code/statute by prompting ChatGPT with the questions and answers using a prompt template, which we have attached in Appendix A.6.

ChatGPT's response was then manually verified by a human reviewer to have output the correct code/statute. See an example of augmented Q&A data generation in Appendix A.7.

3.3 LLMs to Evaluate

ChatGPT-3.5 We chose ChatGPT-3.5 as one of the LLMs under test since our goal is to analyze LLMs' ability to increase accessibility to legal counsel in California, given that it is one of the most widely available LLMs in the United States at the moment. Further ChatGPT-3.5's free and easy-to-use prompt-based UI makes it the likely LLM choice for the average person thereby allowing us to simulate the general population's experience with LLMs for our test environment.

Claude-3 We also chose Anthropic's Claude-3 as the second LLM. Anthropic's differentiator is AI safety. ChatGPT is trained on human preferences, but Anthropic specifically defined a set of constitutional AI principles to refine their model. Some examples of principles in this constitution include opposition to inhuman treatment and privacy.

3.4 Evaluation Methods

To increase the scalability of our experiment, we have selected automated evaluation metrics where possible, but still include some human judgment-based evaluation metrics as the most important element of legal Q&A is whether it is interpretable by humans. We chose evaluation methods that we believe are very similar to human judgment. Thereby, the results of the evaluations can directly lead to LLM performance improvements for its downstream purpose of responding to legal questions.

We have selected two evaluation aspects "Accuracy" and "Interpretability", each of which contains three automated metrics and one human-reviewed metric for a total of eight evaluation metrics. We chose accuracy to be how closely the generated answer is to the lawyer's answer. We chose interpretability to be how readable the generated answer is. Our intended users are lay people who do not have legal backgrounds, so the generated output should be interpretable for a general audience.

3.4.1 Accuracy

BERTScore BERTScore [Zhang et al., 2019] uses BERT contextual embeddings to measure the semantic similarity between a reference text and generated text. The value of BERTScore falls between -1 and 1. The higher the BERTScore, the higher the similarity between the two texts.

BERT Score is a combination of precision and recall to compute the F1 measure. See below, the basic principle equations for BERTScore taken from [Zhang et al., 2019]:

$$R_{BERT} = \frac{1}{|x|} \sum_{x_i \in x} \max_{\hat{x}_j \in \hat{x}} x_i^\top \hat{x}_j \tag{1}$$

$$P_{BERT} = \frac{1}{|\hat{x}|} \sum_{\hat{x}_j \in \hat{x}} \max_{x_i \in x} x_i^\top \hat{x}_j \tag{2}$$

$$F_{BERT} = 2\frac{P_{BERT} \cdot R_{BERT}}{P_{BERT} + R_{BERT}} \tag{3}$$

[Zhang et al., 2019] showed that it could correlate the two sentences better than human judgments. Hence, we will use BERTScore to measure the similarity between the LLM and the lawyer's answer. The higher accuracy of the LLM's answers is reflected by higher BERTScore.

BLEURT Entailment Score Entailment Score [Sellam et al., 2020] leverages natural language inference to determine where a generated text is Entailment/Positive, Contradiction/Negative, or Neutral. The value of the Entailment Score falls between -1 and 1. The higher the Entailment score, the more strongly the generated text is entailed by the reference text. We will use the Entailment Score to determine if the answer from LLM entails the lawyer's answer (legal question). Higher accuracy of the LLM's answers comes along with a higher Entailment Score.

GPTScore GPTScore [Fu et al., 2023] is a training metric by a neural network that measures the quality and fluency of the generated text. The value of GPTScore falls between 0 and 1. Generally, scores above 0.8 represent high-quality generation and 0.5 or below represent low quality

Human Annotation and F1 Score Using the augmented dataset, which has the ground truth being the California codes and statutes that mandate the lawyer's response, we will perform a human evaluation of the accuracy of the ground truth codes and the codes generated by the LLMs. The measure of accuracy will be quantitative and evaluated by the F1 score.

3.4.2 Interpretability

Flesch Reading Ease The Flesch reading ease score will be used to judge whether an output is suitable for the intended audience. Our intended audience does not have a legal background, so the generated result should not be difficult to understand. The Flesch score gives a text a score between 1 and 100. The score takes into account the average sentence length and average word length in syllables to compute the overall readability score. Appendix A.8 describes how to interpret Flesch reading scores.

This score is easily calculated using Equation 4, where RES is Flesch Reading Ease, SL is average sentence length, and SW is average word length in syllables.

$$RE = 206.835 - (1.015 \cdot SL) - (84.6 \cdot SW) \quad (4)$$

Our goal is to have the LLMs generate a score around 60-70 which is described as a standard level of difficulty.

Flesch-Kincaid Grade Level Similar to the Flesch reading ease score, the Flesch-Kincaid grade level will be used to judge whether an output is suitable for the intended audience. Our intended audience does not have a legal background, so the generated result should not be difficult to understand for the average person. Statistics derived from [U.S. Census Bureau, 2022] data describe the highest education level attained by percentage of the U.S. population 18 years and older in Appendix A.9. From this data we find 90.67% of the U.S. adult population has at minimum a high school degree, so the LLMs will be considered interpretable by the intended audience if it can achieve a Flesch-Kincaid grade level of 12 or less.

Lexical diversity The next score is Lexical Diversity. A high lexical diversity score indicates that the model is not simply rephrasing the same idea multiple times. With grammatically rich structure, we can quantify the model's creativity and originality in its response. The lexical diversity formula is simple and is just the total number of tokens divided by the number of different words.

Topic Coherence Topic coherence will be implemented by Latent Dirichlet Allocation. Our goal is to see if the model generates an answer focused on a specific idea throughout the output. The topic coherence score is evaluated on a scale between [0, 1], with scores closer to 1 suggesting that the topic is supported by the text.

4 Results

4.1 Accuracy

Our interpretability result data can be found in appendix A.10.

BERTScore There was no significant difference between ChatGPT and Claude in terms of BERTScore as their overall average difference is less than 0.01.

However, we have spotted an interesting finding that they coincidently worked better in Civil compared with Criminal. It could be due to the answering style for different categories as Civil Law cases are usually about contracts in which LLMs could better answer with relatively general answers. However, for Criminal Law most cases are unique, and hence more difficult to have a more answers closer to the reference answers by lawyers.

BLEURT Entailment Score ChatGPT and Claude showed similar performance in BLEURT (Entailment Score) as well, but again Claude has a less than 0.01 score higher than ChatGPT.

Compared with the BERT Score, we could see the scores are much lower. A higher BLEURT means the reference answer by lawyers implies the answers by LLM, but in our case, it should not be true. Hence, we realized that this is not the best score to evaluate the application we selected but still, we would like to keep this in the reference and look for improvement next time.

GPTScore In the pairwise comparison approach, the LLM judge essentially decides the better response or whether they are equivalent. The graph indicates roughly identical trends and

the statistics show Claude slightly outperforms ChatGPT as the mean is 2.5% higher.

The single answer grading method on the other hand shows Claude outperforms ChatGPT by an even higher margin: 3.7% better on average.

Finally, for reference-guided grading methods where a reference solution was provided for comparison, the models performed nearly identically. This might be due to the fact that both LLMs provided legal answers containing the same set of legal codes which apply to the original questions and are pertinent to the reference answer.

As we are using ChatGPT itself as a LLM judge, it might be biased towards the response generated by another ChatGPT answer. This is also explained by the concept of positional biases [Zheng et al., 2023]. Position bias describes the phenomenon that LLM tends to favor certain positions over others, a bias common in the NLP domain, as well as in human-decision making. Since one of the LLM response generators and the judge are sharing identical model architecture, the favoritism might come into play when rating pairwise scores on things such as creativity, details, level of depth etc.

In addition to the concept of positional bias, self-enhancement bias can also play a role in the result here. Self-enhancement bias is a statistically examined concept that indicates LLMs such as GPT-4 favor their own responses with a 10% higher win rate [Zheng et al., 2023]. Despite all the biases in play here, we can still see Claude performs slightly better. Therefore, we can infer the actual difference should be larger than observed statistics, indicating that Claude can be more advantageous in acting as an AI lawyer. This result is consistent with our observations from above.

Human Annotation and F1 Score Overall, ChatGPT and Claude performed very similarly, with F1 scores ranging just below 0.5.

Interestingly, ChatGPT did not understand the question to select specific California codes as well as Claude did, with ChatGPT frequently outputting an answer that did not call out a specific code, but talked about generic legal principles. Overall, qualitatively, our team agreed that Claude's performance was much better than ChatGPT's in understanding the question and identifying the specific California codes that the case question and response was guided by.

There were a few drawbacks to our evaluation method. The first being that we performed a single human evaluation on each output. Since we all have varying expectations and backgrounds, we may have scored the code outputs differently. To better improve the results in the future, at least two human evaluators should evaluate the same output and their scores should be averaged. That being said, this adds increased cost and time to the evaluation.

Another drawback is that the scoring system penalizes when the model would output more codes than the reference equally to when the model outputs wrong codes or does not include them at all. In this way, overall Claude could have performed much better than ChatGPT, but was scored lower because there were many cases where Claude outputted more codes than the reference text. Although these scores were correct, for uniform scoring, this was still penalized. This is an artifact of the F1 score, which is a harmonic mean between precision and recall. To improve this in the future, rather than creating the augmented code set by prompting ChatGPT for codes and statutes to serve as a reference set, we should ask real lawyers to evaluate the outputs.

The reason why Claude performed better than ChatGPT may be rooted in the Claude architecture. Claude has a larger context window compared to ChatGPT. While Claude has a context window of 1,000,000 tokens, ChatGPT only has a context window of 32,000 tokens and may be even smaller for the free version ChatGPT-3.5 that we used to evaluate. With the larger context window, Claude can process a greater amount of text before generating its response, which allows it to understand the bigger picture. Our prompts were very lengthy because they included a large amount of background context and in addition to the question. Claude demonstrated full understanding and comprehension of the large context and the ultimate question, whereas, ChatGPT seemed to "lose memory" of past context and occasionally provided irrelevant information to fill the gaps.

4.2 Interpretability

Our interpretability result data can be found in appendix A.11.

Flesch Reading Ease Both models performed similarly in terms of Flesch Reading Ease with Claude having a slight advantage over

ChatGPT. Per Flesch's measurement, ChatGPT's responses for all categories of legal questions would be considered difficult to read, while Claude's responses would be considered fairly difficult (aside from the intellectual property responses which fell into the difficult category by just under two points). In either case, neither LLM was able to achieve standard score (between 60 and 70) implying that both are using more complex vocabulary than would be found in the average person's reading material.

It should be noted however, that when the Flesch Reading Ease Score was developed by Rudolf Flesch in the 1940s only 24.5% of people aged 25 and older had at minimum a highschool education [U.S. Census Bureau, 2015] compared to 91.2% as of 2022 [U.S. Census Bureau, 2022] . So it is likely the general populations' literacy has improved since the score's creation. Further the first two paragraphs in the popular children's novel Alice's Adventures in Wonderland received a Flesch Reading Ease score of 40.02, so it is not a perfect score of literary complexity.

Flesch-Kincaid Grade Level With respect to the Flesch-Kincaid Grade Level, Claude slightly outperformed ChatGPT with responses for all categories of legal questions falling within the 10th-12th and 12th-14th grade range respectively. Given that 28.98% of adult Americans' highest level of education is a high school graduate this could put Claude at a considerable advantage in terms of general usability as a source for legal advice so as not to exclude over a quarter of the adult population.

Once again it is important to note the Flesch-Kincaid Grade Level is more a suggestion than a rule. In reality it only shows that Claude on average uses shorter words and sentences than ChatGPT, which is not necessarily a perfect translation into a passage's overall linguistic complexity. Once again using the Alice's Adventures in Wonderland example: the first two paragraphs would give a Flesch-Kincaid Grade Level of 21.6, suggesting the target audience would need a PhD, when in reality Lewis Carroll writes long sentences with many commas/colons and uses words with many syllables such as "conversations" and "considering".

Topic Coherence Both LLMs produced low topic coherence scores with negligible differences. Interestingly responses to legal questions falling under the Civil category had significantly lower coherence scores in both LLMs, this may be due to the larger variation in types of civil disputes as opposed to more specific categories of law such as intellectual property and employment. If this is the case, however, then it is surprising that the Criminal category was not noticeably affected by this as well. In either case the topic coherence may have been marginally improved by increasing the number of topics within the LDA model, but for our purposes we chose the topic number to be equal to the number of questions in each dataset.

Lexical diversity For lexical diversity, both models performed similarly. Additionally, there was not a big difference in score across different categories of legal questions. Overall, the lexical diversity score was a little less than 0.5, which indicates a moderate level of vocabulary variation. The LLM outputs are balanced between repeated and unique words.

In academic writings, higher lexical diversity scores are desired, but more conversational text or instructional texts prioritize lower lexical diversity scores because clarity can come from repetition for better conveyance of the ideas. It makes sense for our specific audience of lay people that the lexical diversity score is on the lower side in order to prioritize clarity, however, it is not so low that it would be seen as text in a children's book (for example a score between 0-0.3).

With this said, it's important to note that lexical diversity is a purely quantitative measure that does not account for semantic or synaptic coherence or complexity. The measure used for the lexical diversity calculation in this paper was purely the basic type-token ratio, which can be heavily skewed by text length. For example, in longer texts, it is more likely to have lower lexical diversity score because there is more opportunity for repetition. The LLM generated outputs in this paper are typically skewed on the longer side because there was a lot of information to convey, which demonstrates why the scores of lexical diversity are also on the lower side. For an improved score, in the future, a lexical diversity score calculated with a moving-average may mitigate the length dependence.

5 Conclusion

Our evaluation has covered accuracy and interpretability, also from automated to human evaluation, and we could observe from the result that Claude-3 works slightly better than ChatGPT-3.5 in the role of legal assistant for California Laws. Although there are some limitations and future work to be done, this research could still be a good basis for any subsequent evaluation on "LLMs as Legal Assistant".

5.1 Limitations

Despite promising results shown by ChatGPT and Claude in addressing legal queries. Above data revealed that lLms could produce responses only partially relevant to the legal information. Specifically the legal codes provided sometimes are not directly relevant to the questions posed. This indicates a lack of contextual understanding compared to trained lawyers.

Since real world lawyer responses were used as ground truth, this study did not do a further validation on lawyer responses. This inherently introduces biases as the answers from online lawyering are sometimes incomplete because the goal of these Q and A is directed towards actual offline attorney-client relationships and paid consultations. Therefore there might be cases where LLMs were providing more details than actual lawyer responses.

Furthermore, the automated metrics like BERTScore and BLEURT are still essentially mechanics and difficult to capture the correctness of legal advice. Using sometimes incomplete answers from real lawyers may lower the score of more detailed LLM responses.

Research could address these limitations by leveraging questions and answers that already have standardized answers from legal education materials to improve dataset quality, involving expert legal review into the evaluation to verify the quality of data on both ground truth and collected LLM answers to improve evaluation processes. These approaches can help align the specific legal context of queries and enhance relevant and reliability of evaluation processes.

5.2 Future Works

While there are LLMs that focus specifically on laws (LaWGPT, Lawyer LLaMA, LexiLaw), there have not been any created specifically to answer questions relating to California legislation. In the future, should a California legislation specific LLM be implemented, it would add to this topic to compare its effectiveness as a legal counselor with that of the generic LLMs analyzed in this paper. This paper's analysis could even be used as a baseline for success criteria for the team designing such a California legislation specific LLM as it would be reasonable to assume that this specific LLM should perform this function with greater efficiency and rigor than their generic counterparts.

Despite the choices of LLMs, there are some future improvements on the evaluation methods with some legal experts joining in. First, we could accurately evaluate the F1 score on comparing the correct legal code/statues from the legal questions and LLM answers. Even without legal experts' support, we could arrange 2 or more teammates to work on the human evaluation to raise the accuracy. Secondly, we could verify if the reference answer from an online lawyer is correct by legal experts, as all our evaluations are with the assumption that the reference answer is ground truth.

In most LLM applications, prompt engineering is one of the topics to be visited. Our team has discussed this but the purpose of the research is to evaluate if LLM can be a resource in providing accurate and interpretable legal advice. Hence, we would not expect an end-user who is looking for assistance to always have the standardized questions that could fit into an augmented prompt, therefore we decided to leave it as open-ended questions from an online forum. But in the future if any developers plan to build a legal question focused LLM application, some guidelines on prompts would help increasing the accuracy as well.

6 Work Distribution

Please, refer to Appendix A.4.

References

Marjan Ajevski, Kim Barker, Andrew Gilbert, Liz Hardie, and Francine Ryan. Chatgpt and the future of legal education and practice. *The Law Teacher*, 57(3):352–364, 2023. doi: 10.1080/03069400.2023.2207426.

Michael Bommarito II and Daniel Martin Katz. Gpt takes the bar exam. 2022. doi: https://dx.doi.org/10.2139/ssrn.4314839.

Ilias Chalkidis. Chatgpt may pass the bar exam soon,

but has a long way to go for the lexglue benchmark. 2023. doi: http://dx.doi.org/10.2139/ssrn.4385460.

Rudolf Flesch. A new readability yardstick. *Journal of Applied Psychology*, 32(3):221–233, 1948. doi: 10.1037/h0057532.

Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. Gptscore: Evaluate as you desire, 2023.

Pat Grady and Charlie Curnin. Partnering with harvey: Putting llms to work, 2023. URL https://www.sequoiacap.com/article/partnering-with-harvey-putting-llms-to-work/.

Steffen Herbold, Annette Hautli-Janisz, Ute Heuer, Zlata Kikteva, and Alexander Trautsch. A large-scale comparison of human-written versus chatgpt-generated essays. *Scientific Reports*, 13(1):18617, 2023. doi: 10.1038/s41598-023-45644-9. URL https://doi.org/10.1038/s41598-023-45644-9.

Wendell Johnson. *Language and speech hygiene: An application of general semantics; Outline of a course*. 1939.

J. Peter Kincaid, Richard L. Rogers, Brad S. Chissom, and Robert P. Jr. Fishburne. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical report, Naval Technical Training Command, 1975. URL https://apps.dtic.mil/sti/pdfs/ADA006655.pdf.

R. D. Mullen. From standard magazines to pulps and big slicks: A note on the history of us general and fiction magazines. *Science Fiction Studies*, 22(1):144–156, 1995. URL http://www.jstor.org/stable/4240420.

Alberto Muñoz-Ortiz, Carlos Gómez-Rodríguez, and David Vilares. Contrasting linguistic patterns in human and llm-generated text, 2023.

Charles Santiago Sanders Peirce. Prolegomena to an apology for pragmaticism. *The Monist*, 16(4):492–546, 1906. ISSN 00269662. URL http://www.jstor.org/stable/27899680.

Pedro Reviriego, Javier Conde, Elena Merino-Gómez, Gonzalo Martínez, and José Alberto Hernández. Playing with words: Comparing the vocabulary and lexical richness of chatgpt and humans, 2023.

Mishcon De Reya. Mishcon de reya's exploration of ai technologies featured in the media, 2023. URL https://www.mishcon.com/news/mishcon-de-reyas-exploration-of-ai-technologies-featured-in-the-media.

Holli Sargeant, Ahmed Izzidien, and Felix Steffek. Topic modelling case law using a large language model and a new taxonomy for uk law: Ai insights into summary judgment. (21/2024), May 2024. Available at SSRN: https://ssrn.com/abstract=4836558 or http://dx.doi.org/10.2139/ssrn.4836558.

Thibault Sellam, Dipanjan Das, and Ankur P. Parikh. Bleurt: Learning robust metrics for text generation. 2020. doi: https://doi.org/10.48550/arXiv.2004.04696.

U.S. Census Bureau. A half-century of learning: Historical census statistics on educational attainment in the united states, 1940 to 2000. U.S. Department of Education, 2015.

U.S. Census Bureau. Educational attainment of the population 18 years and over, by age, sex, race, and hispanic origin: 2022. U.S. Department of Education, February 2022.

David Wakeling. Ao announces exclusive launch partnership with harvey, 2023. URL https://www.aoshearman.com/en/News/ao-announces-exclusive-launch-partnership-with-harvey.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. 2019. doi: https://doi.org/10.48550/arXiv.1904.09675.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena, 2023.

Hatice Özbolat. Text summarization: How to calculate bertscore. URL https://haticeozbolat17.medium.com/text-summarization-how-to-calculate-bertscore-771a51022~:text=BertScore%20is%20a%20method%20used,gram%2Dbased%20metrics%20often%20encounter.

A Appendix

A.1 Code Repository

We compiled our code repository in GitHub:
Code Repo.

A.2 Q&A Dataset

We compiled all datasets and scores generated into a Google Sheets document:
Q&A Dataset and Scores.

A.3 Paper presentation

We recorded our presentation and posted it to YouTube:
Paper presentation

A.4 Work Distribution

Table 1 outlines our work distribution.

Task	Description
Kick-off	1) Discuss what topic for research 2) Discuss the rubrics for the dataset (criteria for what Q&A will be in the dataset)
Brainstorm Evaluation Metrics	Alton and Sarah: 1) Accuracy 2) Interpretability
Brainstorm LLM To Use	Alyssa and Ruite: 1) GPT-3 2) Claude
Dataset Design Preparation	**Legal Questions Preparation** Each of the teammates to prepare 15 questions in 4 topics from online lawyer Q&A websites: **LLM Answers Generation** 1) Prompt: Question 2) Ground truth: Lawyer response 3) Augmented Question 4) Code/Statues Verification & F1 Score **Task Distribution** 1) Employment Law: Alton 2) Civil litigation: Sarah 3) Criminal: Ruite 4) Intellectual Property: Alyssa
Evaluate LLMs	**Accuracy** 1) BERT Score: Alton 2) BLEURT Score: Alton 3) LLM-as-a-Judge Score: Ruite 4) Augmented dataset + Human Score: Sarah **Interpretability** 1) Flesch Reading Ease Score: Alyssa 2) Flesch-Kincaid Grade Level: Alyssa 3) Lexical Diversity: Sarah 4) Topic Coherence: Alyssa
Project Reports	Each of the teammates worked on 3-4 sub-sections in both midterm report and final report
Presentation	Each of the teammates worked on the slides of their presented parts

Table 1: Distribution of work by author and task.

A.5 Question Selection Rubric

RULE 1: Don't include questions that call out specific companies by name without further context.
REASONING: LLMs like chatGPT-3.5 do not have access to the internet and were only trained with information up to April 2023.

RULE 2: Choose questions asked about California law.
REASONING: Different states follow different codes and statutes. For ease of automatic and manual evaluation, we restricted our scope to CA-based questions only.

RULE 3: If there are multiple responses from lawyers (add up to three), prioritize the answer with the most "votes".
REASONING: If there are more responses from lawyers, we can compare the LLM-generated answer with multiple interpretations of the question from different lawyers.

A.6 Prompt augmented Q&A data generation

Table 2 shows the prompt that we use to query the generation of CA Codes and statutes for the Augmented Q&A task.

Prompt Template
"I'm going to give you a question and an answer from a lawyer. Please tell me what California code or statute guides this answer. Question: [QUESTION] Answer: [ANSWER]"

Table 2: Prompt template for augmented Q&A data generation.

A.7 Sample augmented Q&A data generation

Table 3 demonstrates the steps we use to query the generation of CA Codes and statutes for the Augmented Q&A task.

A.8 Flesch Reading Ease Interpretation

Table 4 is a guideline for interpreting Flesch reading scores from Flesch [1948].

[1]Slick fiction was a term for middle-class mass produced fiction magazines, named so due to the better quality, shinier paper on which they were printed. [Mullen, 1995]

[2]Pulp fiction was a term for low grade mass produced fiction magazines, named so due to the the technique of producing cheap paper from wood-pulp [Mullen, 1995].

Question from person
Can I record someone without their consent over the phone?
Answer from lawyer
No, you may not because of California's two-party consent law.
Ask LLM
"I'm going to give you a question and an answer from a lawyer. Please tell me what California code or statute guides this answer. Question: Can I record someone without their consent over the phone? Answer: No, you may not because of California's two-party consent law."
Code/statute answered by ChatGPT and verified by human manual review
Cal. Penal Code § 632.

Table 3: Example of augmented Q&A data generation.

A.9 U.S. Educational Attainment

Table 5 describe the highest education level attained by percentage of the U.S. population 18 years and older as derived from U.S. Census Bureau [2022].

A.10 Accuracy Data

Table 6 and Table 7 shows the result data of our interpretability evaluations - BERT, BLEURT, and F1 Score on the Human Annotated augmented dataset - for ChatGPT-3.5 and Claude-3, respectively. Figure 1 graphically shows the comparison of ChatGPT-3.5 and Claude-3 performance with respect to the three GPTScore (LLM-as-Judge) - Pairwise Comparison, Single Answer Grading, and Reference Guided Grading - results as tested on all 66 test prompts (datapoints from the Q&A Dataset). Table 8 shows minimum, maximum, and average ChatGPT-3.5 and Claude-3 performance with respect to the three GPTScore (LLM-as-Judge) - Pairwise Comparison, Single Answer Grading, and

Reading Score	Description of Style	Typical Magazine	Average Word Length in Syllables	Average Sentence Length in Words
0 - 30	Very Difficult	Scientific	1.92 or more	29 or more
30 - 50	Difficult	Academic	1.67	25
50 - 60	Fairly Difficult	Quality	1.55	21
60 - 70	Standard	Digests	1.47	17
70 - 80	Fairly Easy	Slick fiction [1]	1.39	14
80 - 90	Easy	Pulp fiction [2]	1.31	11
90 - 100	Very Easy	Comics	1.00 - 1.23	1 - 8

Table 4: Flesch reading scores interpretation as compiled in Flesch [1948]

Grade Level	Population 18+ (%)
None	0.31
1st - 4th grade	0.61
5th - 6th grade	1.13
7th - 8th grade	1.28
9th grade	1.16
10th grade	1.30
11th grade	3.54
High school graduate	28.98
Some college, no degree	15.90
Associate's degree, occupational	4.26
Associate's degree, academic	5.85
Bachelor's degree	22.50
Master's degree	9.80
Professional degree	1.41
Doctoral degree	1.98

Table 5: Highest education level attained by percentage of the U.S. population 18 years and older as derived from U.S. Census Bureau [2022]

Reference Guided Grading - results.[3]

A.11 Interpretability Data

Table 9 and Table 10 show the result data of our interpretability evaluation for ChatGPT-3.5 and Claude-3, respectively.

[3]Data point 53 (shows a score of 0) did not have either a GPT or Claude response in our dataset and should have been removed prior to testing.

Category of Question	BERT	BLEURT	F1 Scores - Human Annotated
Civil	0.904	0.472	0.444
Employment	0.886	0.480	0.333
Intellectual Property	0.855	0.481	0.367
Criminal	0.830	0.459	0.687
Overall average	0.886	0.480	0.458

Table 6: Summary of ChatGPT-3.5 Accuracy Scores as measured by Evaluation Methods: BERT, BLEURT, and F1 Scores - Human Annotated

Category of Question	BERT	BLEURT	F1 Scores - Human Annotated
Civil	0.915	0.474	0.361
Employment	0.889	0.499	0.453
Intellectual Property	0.861	0.500	0.519
Criminal	0.830	0.452	0.557
Overall average	0.892	0.488	0.473

Table 7: Summary of Claude-3 Accuracy Scores as measured by Evaluation Methods: BERT, BLEURT, and F1 Scores - Human Annotated

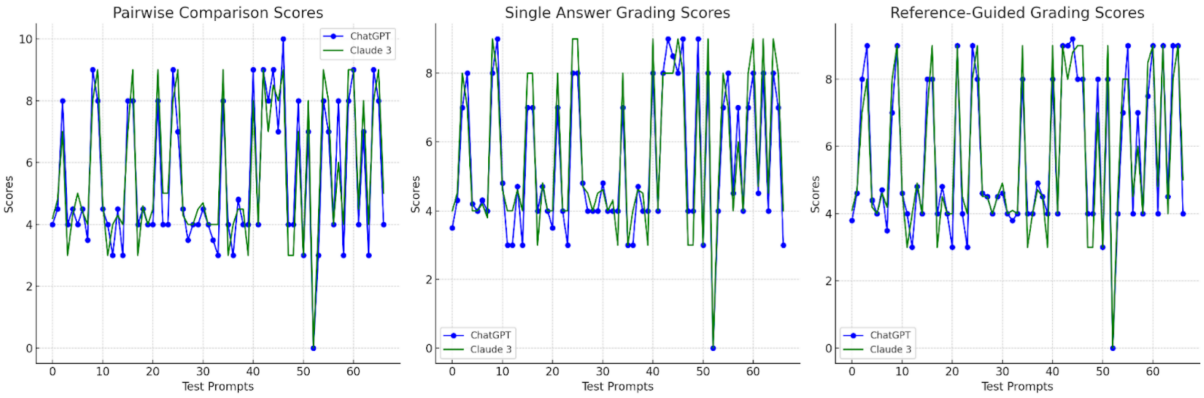

Figure 1: Comparison of ChatGPT-3.5 and Claude-3 performance with respect to the three GPTScore (LLM-as-Judge) - Pairwise Comparison (left), Single Answer Grading (middle), and Reference Guided Grading (right) - results as tested on all 66 test prompts

Test	count	mean	std	min	25%	50%	75%	max	variance
$\text{Pairwise}_{ChatGPT}$	67.0	5.422	2.269	0.0	4.0	4.5	8.0	10.0	5.149
$\text{Single}_{ChatGPT}$	67.0	5.385	2.108	0.0	4.0	4.5	7.5	9.0	4.444
$\text{Reference}_{ChatGPT}$	67.0	5.669	2.271	0.0	4.0	4.6	8.0	9.2	5.155
$\text{Pairwise}_{Claude3}$	67.0	5.557	2.245	0.0	4.0	4.5	8.0	9.0	5.039
$\text{Single}_{Claude3}$	67.0	5.581	2.241	0.0	4.0	4.5	8.0	9.0	5.020
$\text{Reference}_{Claude3}$	67.0	5.690	2.304	0.0	4.0	4.6	8.0	9.0	5.306

Table 8: ChatGPT-3.5 and Claude-3 comparative performance data with respect to the three GPTScore (LLM-as-Judge) - Pairwise Comparison, Single Answer Grading, and Reference Guided Grading

Category of Question	Flesch Reading Ease	Flesch-Kincaid Grade Level	Topic Coherence Score	Lexical Diversity Score
Civil	44.077	13.159	0.298	0.414
Employment	45.665	12.647	0.449	0.443
Intellectual Property	44.015	13.020	0.413	0.384
Criminal	48.037	12.743	0.543	0.496
Overall average	**45.264**	**12.923**	**0.426**	**0.431**

Table 9: Summary of ChatGPT-3.5 Interpretability Scores as measured by Evaluation Methods: Flesch Reading Ease, Flesch-Kincaid Grade Level, Topic Coherence Score, Lexical Diversity Score

Category of Question	Flesch Reading Ease	Flesch-Kincaid Grade Level	Topic Coherence Score	Lexical Diversity Score
Civil	51.572	11.705	0.314	0.445
Employment	55.816	10.833	0.425	0.440
Intellectual Property	48.423	11.873	0.545	0.456
Criminal	51.230	11.800	0.403	0.449
Overall average	**51.748**	**11.565**	**0.422**	**0.435**

Table 10: Summary of Claude-3 Interpretability Scores as measured by Evaluation Methods: Flesch Reading Ease, Flesch-Kincaid Grade Level, Topic Coherence Score, Lexical Diversity Score