Motivations for Reframing Large Language Model Benchmarking for Legal Applications

Riya Ranjan
Department of Computer Science
Stanford University
Palo Alto, CA 94305
rranjan1@stanford.edu

Megan Ma
Stanford Center for Legal Informatics
Stanford University
Palo Alto, CA 94305
meganma@law.stanford.edu

Abstract

Informative benchmarks of large language model performance in domain-specific areas are limited. Particularly, benchmarks of LLMs for legal applications are insufficient, as they are often confined to a narrow set of tasks that do not imitate true legal workflows, or are difficult to replicate, with a lack of transparency about how criteria are discerned and outputs scored. We propose a new framework for benchmarking legal LLMs based on tasks that accurately reflect real legal workflows: lawyer preference. We argue that benchmarking for preference can capture nuances in how legal practitioners evaluate their own work, and thus provides a more suitable metric for the quality of LLMs for legal work.

1 The Current Standard for Legal LLM Benchmarking

Numerous studies have published benchmark tasks to evaluate the performance of large language models (LLMs). Many of these benchmarks are used to assess the quality of LLMs by evaluating their performance on a subset of human tasks [8] [17] [2]. Further benchmarks attempt to measure performance of LLMs in domain specific areas, including medicine, law, and finance.

Emerging use of generative AI tools in the legal profession has made benchmarking LLMs for legal use increasingly important; this year, nearly 35 percent of law firms worked with a generative AI provider [11]. Numerous studies [6] [1] [9] have considered whether LLMs could perform legal tasks; others use hallucination rate, accuracy, and completeness of responses as evaluation criteria [15] [13].

While these benchmarks attempt to provide a concrete metric for comparing LLMs against one another, the actual usefulness of the metrics is uncertain. Few benchmarks today consider how models realistically measure up against human lawyer performance for true, holistic legal workflows. For example, in evaluating the hallucination rate of LLMs, a majority of published benchmark tasks prompt models to recall hyper-specific court opinions or niche statutes, neither of which capture how legal workflows operate in practice [13]. Thus, these metrics do not inform whether LLMs are valuable in the creation of legal work products – or whether LLMs are truly useful for lawyers.

Additionally, the notion of quality in these tasks is difficult to ascertain and reproduce. For example, Big Law Bench [15], recently released by Harvey, assigns point values to a set of criteria to assess the quality of LLM output on particular legal tasks, including memorandum drafting and legal contract analysis. However, these points appear arbitrarily assigned, as certain criteria (e.g., style, structure) account for a much smaller weight than whether individual facts are included/excluded in responses [5]. These metrics fail to consider more nuanced interpretations of quality such as creativity of argument or thoroughness of response. There is an additional lack of visibility as to how these criteria are scored for each output – whether by lawyers or a research team — and who, ultimately, determines whether a response is stylistically sufficient or contains "extraneous" information.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

Another key shortfall of existing benchmarks is that they fail to preserve and elevate lawyer heterogeneity. Ma et al. [12] found that even senior attorneys rarely converge in conclusions when performing legal analyses. This variation and creativity in legal thought exists even within law firms, contributing to the creation of more nuanced legal strategies and improved outcomes for clients. Attorney opinion remains a fundamental work product of the legal field [18], the value of which is diminished by encouraging blanket homogeneity in legal thinking. Existing LLM benchmarks that seek out accuracy and completeness on legal tasks as singular metrics for quality fail to evaluate how LLMs supplement legal thinking and opinions in other meaningful ways and/or how they provoke importantly different thought and reasoning from individual legal professionals.

We contend that normative metrics for benchmarking LLMs in the legal domain fall short in their usefulness for assessing how LLMs can replicate and augment legal workflow and legal thinking. We argue that preference rating LLM outputs to determine more holistic criteria for what lawyers expect, like, and dislike on a comprehensive set of real legal workflows will provide a more effective benchmark to determine how useful both enterprise LLMs ([7], [14], [19], [16], [10]) and foundation models are for legal practitioners.

2 Beyond Shallow Metrics: Why Lawyer Preference Ratings Matter

Preference rating has been used in tandem with reward models to align LLMs to human preference via reinforcement learning with human feedback (RLHF). Zheng et al. [20] found that models that are fine tuned in this manner are not properly evaluated by traditional LLM benchmarks [8], as the nuances of human preference may not be accurately captured by the benchmark tasks.

There is additional precedent for using preference rating for domain-specific benchmarking. For example, Fleming et al. [4] proposed MedAlign as a benchmark for clinician preference on LLM output, asking clinicians to evaluate LLM output on a variety of concrete clinical tasks in comparison to gold-standard, human-created output. Their results indicated that clinician preference was only moderately correlated with existing traditional benchmarks for clinical LLM applications, thereby confirming the importance of preference as an additional metric.

We propose a similar framework for evaluating LLM performance on real legal tasks: benchmarking lawyer preference. Our first step is creating a set of legal tasks that encapsulate a legal workflow in practice to determine how LLMs, against human counterparts, can impact both final legal work products and interposed legal work processes. Our second step is a blind evaluation of these outputs from senior legal practitioners to determine lawyer preference; lawyers will not only rate which outputs they prefer, but provide detailed justifications for why they are preferable [1]. Concretely capturing lawyer preference about LLM outputs versus human created outputs will additionally help create more precise and nuanced criteria for continued evaluation of legal LLMs.

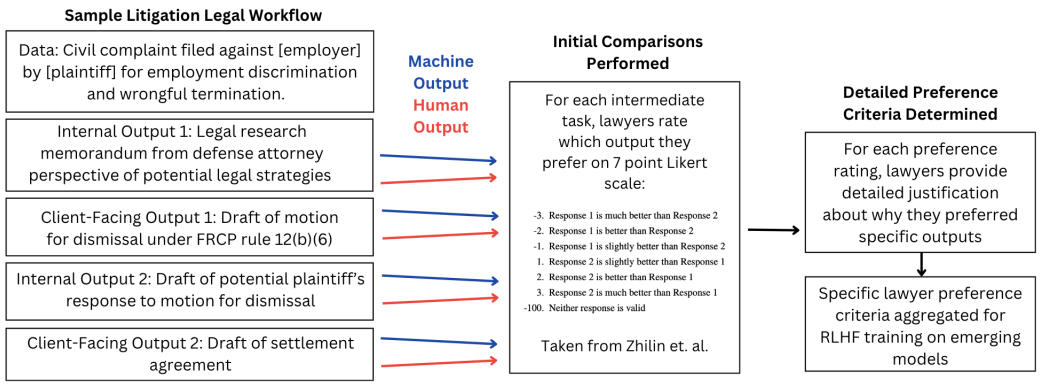

Figure 1: A sample outline of our framework. In implementation, multiple machine outputs will be compared to differentiate foundation and enterprise model capabilities.

3 Limitations and Social Impact

3.1 Limitations

We do acknowledge that there are limitations to our proposed framework. Namely, preference rating runs the risk of a lack of broader convergence on metrics around quality; preference rating from individual lawyers and specific law firms may yield results that are too heterogeneous to be exactly replicated. Nevertheless, we consider that observations from these investigations will be informative to research in domain-specific evaluation. Additionally, as models have probabilistic outputs, it may be difficult to replicate exact work products generated during experimentation. However, we plan to mitigate this in our experiments by disclosing prompts and explicit prompt engineering methodologies.

Additionally, we acknowledge that a limited set of tasks may not reflect legal workflows on an international scale, particularly in non-Western or non-common law contexts. This can be addressed by extending our framework to globally diverse workflows, with preference evaluations performed by practitioners outside of the United States. Such initiatives are already being led with the implementation of generative AI tools in Singapore [3]; further initiatives can be extended in the Global South. This may yield even more heterogeneous results; however, such work will provide important context about how LLM-based legal tools can support lawyers globally, and how such tools for non-Western communities may be crafted.

3.2 Broader Impact

This work offers value to a multitude of stakeholders across the legal ecosystem. Law firms will be able to determine which models more effectively meet legal workflow needs on relevant criteria, while vendors developing generative AI legal products will better understand how their tools could more effectively support legal practitioners. More broadly, prospective and current consumers of legal services will have more transparency around the parameters and requirements of quality legal work products.

References

[1] Jon Choi, Kristin E. Hickman, Amy B. Monahan, and Daniel Schwarcz. Chatgpt goes to law school. *Journal of Legal Education*, 71(3):387–400, 2022. ISSN 0022-2208. URL https://jle.aals.org/home/vol71/iss3/2/.

[2] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems, 2021. URL https://arxiv.org/abs/2110.14168.

[3] Ella Fincken. Singapore legal sector embraces ai, Sep 2024. URL https://www.globallegalinsights.com/news/singapore-legal-sector-embraces-ai/#:~:text=With%20a%20vision%20of%20greater,with%20the%20Legal%20Technology%20Platform.

[4] Scott L. Fleming, Alejandro Lozano, William J. Haberkorn, and et. al. Medalign: A clinician-generated dataset for instruction following with electronic medical records, 2023. URL https://arxiv.org/abs/2308.14089.

[5] Niko Grupen and Gabriel Pereyra. Big law bench, 2024. URL https://github.com/harveyai/biglaw-bench?tab=readme-ov-file#readme.

[6] Neel Guha, Julian Nyarko, Daniel E. Ho, Christopher Ré, Adam Chilton, Aditya Narayana, and Alex Chohlas-Wood et. al. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models, 2023. URL https://arxiv.org/abs/2308.11462.

[7] Harvey, 2024. URL https://www.harvey.ai/.

[8] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding, 2021. URL https://arxiv.org/abs/2009.03300.

[9] Daniel M. Katz, Michael J. Bommarito, Shang Gao, and Pablo Arredondo. Gpt-4 passes the bar exam. *Philosophical transactions. Series A, Mathematical, physical, and engineering sciences*, 2024. URL https://royalsocietypublishing.org/doi/10.1098/rsta.2023.0254.

[10] Bloomberg Law. Legal research software, 2024. URL https://pro.bloomberglaw.com/?utm_medium=paidsearch&utm_source=google&gclsrc=aw.ds&gclid=Cj0KCQjwrp-3BhDgARIsAEWJ6SyJWO3yRbnNT_siD2M2ErBBvN9ZMoexPN2aq5bg3_ICdCkCydaexb4aAoKlEALw_wcB&trackingcode=BLAW24111822.

[11] Steven Lerner. Where lawyers stand on generative ai tools, 2024. URL https://www.law360.com/pulse/articles/1827683/where-lawyers-stand-on-generative-ai-tools.

[12] Megan Ma, Brandon Waldon, and Julian Nyarko. Conceptual questions in developing expert-annotated data. *ICAIL*, 2023. URL https://dl.acm.org/doi/abs/10.1145/3594536.3595139.

[13] Varun Magesh, Faiz Surani, Matthew Dahl, Mirac Suzgun, Christopher D. Manning, and Daniel E. Ho. Hallucination-free? assessing the reliability of leading ai legal research tools, 2024. URL https://arxiv.org/abs/2405.20362.

[14] Lexis Nexis. Lexis+ ai: Ai legal assistant by lexisnexis, 2024. URL https://www.lexisnexis.com/en-us/products/lexis-plus-ai.page.

[15] Julio Pereyra, Elizabeth Lebens, Matthew Guillod, Laura Toulme, Cameron MacGregor, David Murdter, Karl de la Roche, Emilie McConnachie, Jeremy Pushkin, Rina Kim, and et al. Introducing biglaw bench, Aug 2024. URL https://www.harvey.ai/blog/introducing-biglaw-bench.

[16] Thomson Reuters, 2024. URL https://www.thomsonreuters.com/en/artificial-intelligence.html.

[17] Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, and Adrià Garriga-Alonso et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models, 2023. URL https://arxiv.org/abs/2206.04615.

[18] Wex Definitions Team, 2021. URL https://www.law.cornell.edu/wex/attorney_work_product.

[19] vLex, 2024. URL https://vlex.com/vincent-ai-home.

[20] 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. URL https://arxiv.org/abs/2306.05685.