Lawyers Should Not Trust AI

A call for an Open-source Legal Language Model

Samuel DAHAN[a,1], Rohan Bhambhoria[b] David Liang[b], Xiaodan Zhu[b]
[a] *Queen's Law, Cornell Law School, Conflict Analytics Lab*
[b] *Queen's Computer and Electrical Engineering*
ORCiD ID: https://orcid.org/0000-0002-6177-5729

Abstract. Generalized AI like ChatGPT cannot and should not be used for legal tasks. It presents significant risks for both the legal professions as well as litigants. However, domain-specific AI should not be ruled out. It has potential for legal research as well as access to justice. In this paper we call for the development of an open-source and distributed legal AI accessible to the entire legal community. We believe it has the potential to address some of limitations related to the use of general AI for legal problems and resolving disputes -- shortcomings that include legal misinformation or hallucinations, lack of transparency and precision, and inability to offer diverse and multiple narratives.

Keywords. Legal AI, Open source, Decentralized and Distributed Learning, Feedback, Legal profession.

1. Introduction

Recent evidence shows that AI is becoming less intelligent, and the reasons are unknown. Findings suggest that ChatGpt is "drifting"[1] – also known as wild fluctuations in the technology's ability to perform certain tasks. Over just a couple of months, the machine went from answering a simple math question 98% of the time to just 2%.

What does this mean for the use of AI in law? Not that much, considering that general AI systems have never performed well in law. The use of AI for law has been a hot subject in computer science for quite some time.[2] Very few people outside this world have been paying attention to language models. However, when OpenAI launched Chat-GPT – the fastest-growing consumer application in history[13] – many realised the capability of AI for productivity tasks such email drafting, but also for more advanced technical task such as contract drafting or medical diagnostics. But here is the problem: the technology is not ready to be used at this scale. AI models are "notorious bullshiters". [14] They are excellent at predicting the next word in a sentence, but they have no knowledge of what the sentence means.

But we all knew about this. This is probably why OpenAI's leader has reiterated that ChatGPT is a research project – although this has not stopped Microsoft from introducing it into Bing despite the inherent flaws of the technology.

Now that the "cat is out of the bag," tech giants are battling each other to develop the highest performing generative language model. And this is not limited to AI-powered search. They compete in computing services, productive software, and enterprise

[1] Corresponding Author: Samuel Dahan, Samuel.dahan@queensu.ca.

[2] Several research labs and private ventures have been exploring these issues for a while. There has been a growing interest since the 2010s. See for instance [2]–[12]

software. What's problematic is that we, the users are acting as "guinea pigs" by testing the technology for free. This war is now having "knock-off" effects in other areas like medicine and law. For instance, there is evidence of overreliance on general AI tools for legal advice and legal research. There have been several high-profile instances of the misuse of generative AI in courts. For example, a recent Forbes headlined "*Lawyer Used ChatGPT In Court—And Cited Fake Cases*."

While the issues of hallucination and citation are important, especially in the legal context, this paper will not be looking at AI flaws in depth. In fact, these have already been well documented.[17] Instead, this research is a non-technical doctrinal effort aimed to explore potential solutions for implementing dependable legal AI solutions that are accessible to the legal community as a whole. This project is part of a greater endeavour to develop an open-source legal AI system, *OpenJustice.ai*. It is also a modest attempt to address generative AI systems' shortcomings when it comes to solving legal problems -- shortcomings that include legal misinformation, lack of transparency and precision, and inability to perform contextual legal reasoning tasks.[3] The next section will review the main risks associated with the use of General AI tools in the legal context (2). The following section discusses the core features of reliable legal AI (3).

2. Why Lawyers Should not Trust General AI

While recent findings shows that generative AI can perform a wide range of legal task, and even pass the bar exam[18] the technology is not there yet. Generative AI architectures notoriously "hallucinate" incorrect answers with strong confidence, fabricating facts, citations, and details. Generalised LLMs like ChatGPT generate text by predicting the set of words that *should* follow, given an initial set of user-provided inputs. A "creative" element is introduced by randomly selecting sentence elements from a list of probable responses.

In other words, AI systems are mostly statistical and therefore do not understand much, let alone legal problems. Generative systems are unable to grasp the semantic nuances of legal terminology. The same word can have different meanings in different jurisdictions: for example, "layoff" means "suspension" in Canada, but "termination" in the United States. More concerning, generative AI systems may *appear* to grasp legal concepts, but be unable to perform counterfactual legal reasoning or classify modes of legal reasoning.[19]

This might not be an issue for statistical legal tasks, like retrieving a precedent or applying simple rules to facts. However, this could be a hurdle for many deep legal reasoning that requires a multidimensional approach that involves an in-depth comprehension of a legal issue. In addition to hallucinations, there are also concerns biased analyses, a well-documented issue with the use of predictive AI in law.[8], [20] In fact, ChatGPT exhibits biases commonly found in humans, such as conjunction bias, probability weighting, overconfidence, framing, anticipated regret, reference dependence, and confirmation bias.[21]

However, the concerns arising from the use of LLMs are not only confined to inaccuracies and the spread of legal misinformation. Even when generative AI provides reliable information the use of general LLMs may have adverse societal effects because

[3] OpenJustice aims to act as an education tool to enable the current generation of language models to learn factually-grounded information, whilst facilitating a professional learning environment for lawyers. Openjustice is created to provide supervision at a large scale through interactions with the current SOTA generation models.

of their inclination to reflect a mainstream worldview. Provided LLMs become an important source of information this could lead to feedback loops, whereby the texts generated by large language models will percolate back into the web and serve as training data for the next generation of text generators, thus creating "AI echo-chambers" that will further narrow our universe of thinkable thoughts.[17, p. 30] This could undermine cultural diversity, limit the multiplicity of narratives that build collective memory, narrow users' perceptions, or impede democratic dialogue.

Such influence is particularly problematic in areas such as law where there is no precise, mathematical "right answer," but rather a range of acceptable answers and room for discretion. The use of LLMs in such cases will inevitably act as information facilitation, which implies that the answers they generate are not neutral representations of information. An LLMs of this kind raises two main concerns: First, it assumes that most legal problems are algorithmic and always call for a straightforward answer. This is a false assumption as we know that lawyers and adjudicators often offer inconsistent solutions for the same legal case.[9] This is probably why lawyers' favorite answer is 'it depends'. Second, an LLMs of this kind that the facts of future cases will be unchanged from those in the past. Yet, this is almost always never the cases considering that social context is constantly bring forward new facts along with new legal problems.[8, p. 18], [22, p. 31] Thus, employing such LLMs would pose to the autonomy and evolution of the law as it would lead to the de-norming of law as well as its ossification.

Finally, generative AI systems are unexplainable as most LLMs are unable to cite their sources. As legal practice involves substantiating propositions based on relevant legal authorities, the lack of citations introduces significant risk, especially since generative AI systems tend to hallucinate case law or legislation. The process of citing the correct legal authorities is central to the job of lawyers - without citations, cases will have no legal standing. GPT4 have been demonstrated to cite the relevant statute and legal authority, however it continues to produce hallucinations and lacks the ability to connect relevant authorities together. The issue is further complicated by the fact that GPT4 is built on a snapshot of the internet at a set point in time. It is unable to account for any new developments in jurisprudence.

3. Towards a Dependable Legal AI: Distributed Domain Adaptation

Considering the evidence discussed in the previous section, a strong argument can be made against the use of legal applications that solely rely on a GPT. A sound alternative is to train a domain specific LLM for law. There are two ways to go about it. One option is to develop a purpose-built generative AI model from scratch for law. A similar initiative has been undertaken in finance. In fact, Bloomberg recently deployed BloombergGPT, a 50-billion parameter large language model build from scratch [23]. To the best of our knowledge, such an initiative has not been undertaken in law. A second (cheaper) option is to fine-tune open-source foundational language models to customize the model with domain-specific and proprietary data. Base language models may be trained on -- (i) unstructured data, of which there is an abundance in many areas of law. This includes case law, journals, and other legal resources. (ii) structured data, which is more costly as it includes annotated data. There are several layers of fine-tuning that can be performed with language models. See Figure 3.

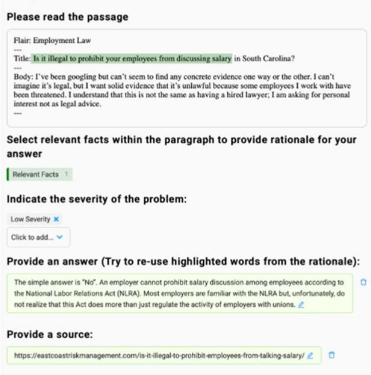
3.1. Raw Data Fine-tuning

Models trained on unstructured legal are tuned on a “masked language modelling task” [24] in which the model is essentially trained on a fill-in-the-blank task. As the “blanks” are already considered as present in the unstructured dataset by simply omitting a part of the data, this form of training on unstructured data is also known as “self-supervised training”. With this method, the model can learn the nuances of a legal language. Note that the model can be refined to a specific area of law. All that is needed is a corpus of raw legal documents.

3.2. Instruction Fine-tuning

Instruction response annotation or fine-tuning is a process that involves feeding the model structured data in the form of question-response pairs. The model is trained using these annotated examples. The model learns to recognize patterns and make predictions based on the given instructions and desired responses. This interdisciplinary approach ensures that AI systems become more intelligent, responsive, and capable of effectively assisting users in a conversational manner.

Figure 2 shows how fine-tuning works in the legal context.[4] Through a secured interface, law students – under the supervision of legal professionals –provide insights on real-world questions found in popular forums or online community pages such as Reddit and Law Stack Exchange.[5]



Please read the passage

Flair: Employment Law

Title: Is it illegal to prohibit your employees from discussing salary in South Carolina?

Body: I've been googling but can't seem to find any concrete evidence one way or the other. I can't imagine it's legal, but I want solid evidence that it's unlawful because some employees I work with have been threatened. I understand that this is not the same as having a hired lawyer; I am asking for personal interest not as legal advice.

Select relevant facts within the paragraph to provide rationale for your answer

Relevant Facts 1

Indicate the severity of the problem:

Low Severity ✕

Click to add...

Provide an answer (Try to re-use highlighted words from the rationale):

The simple answer is "No". An employer cannot prohibit salary discussion among employees according to the National Labor Relations Act (NLRA). Most employers are familiar with the NLRA but, unfortunately, do not realize that this Act does more than just regulate the activity of employers with unions.

Provide a source:

https://eastcoastriskmanagement.com/is-it-illegal-to-prohibit-employees-from-talking-salary/

Figure 1. Question-Response Fine-tuning.

3.3. Open-Source Feedback Fine-tuning

Reinforcement human learning feedback is a concept in which AI models are trained using feedback from human experts to improve their performance. This involves creating an interface that allows the user to test the model and provide feedback. For example, if the system provides incorrect information in response or citation to a query, a human

[4] This is drawn from the OpenJustice project (originally called Smart Legal Clinic).

[5] The process will contain two stages. In Stage 1, the development process consists of the following steps (Figure 1): (1) Highlight text to tag the legal domain along with the legal problem in non-legal terms as described by a user. (2) Reframe the problem in legal terms and associate the relevant legal source. (3) Apply facts to the relevant law and translate answers in non-legal terms in the form of a short test. In Stage 2, Llama2 and other open-access language models are used to generate additional legal scenarios and legal contracts, such as employment contracts. Law students and legal professionals can help further train models by annotating these additional samples identify legal problems and present solutions.

expert can correct or validate the results. In the legal context, we strongly recommend an open-source approach, that is: a non-proprietary version of the model should be openly accessible to the entire legal community; that is, both law schools and legal professionals (Figure 2). In fact, we think it is key to invest in truly open LLMs for law as one of the most immediate issues for the research and legal community is the lack of transparency in these systems. For instance, most of these systems are unable to provide the chain of reasoning and are unable to provide accurate citations. We think that one of the reasons is that the underlying training sets of ChatGPT and its others are not publicly available. This goes against the move towards open science and makes it hard to use LLMs for legal tasks considering the importance of transparency in legal reasoning. To counter this opacity, the development of open-source alternatives should be prioritized.[25] While non-commercial organizations lack the resources to compete with private ventures, some academic collaborations have emerged. BigScience has created a large LLM called Bloom, and the Conflict Analytics Lab has created an open-source legal AI called Openjustice.ai.

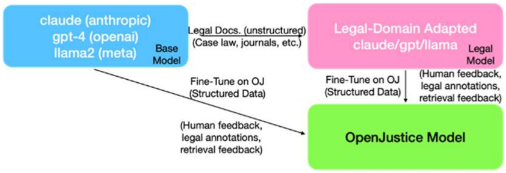


Figure 2. Legal Community Feedback

With this in mind, we do not think it should be open to the public – at least yet – for several reasons. First, AI systems (even domain-specific), hallucinate too often. It would be irresponsible to let such a system provide legal information for self-represented litigants. It needs further training before it is released to the public. Second, opening feedback opportunities to the public might jeopardize the integrity of the data as non-lawyers may be inclined to provide inaccurate and misleading feedback to the model. We believe that such systems will help to obtain clean performance for training language models in contrast models entirely open to the public, such as ChatGPT. In fact, the openness of GPT models might be one of the reasons that GPT4's performance is drifting [1]. Finally, even if there was a perfect language model, even untrained lawyers are likely to have the required prompting skills to extract the most useful information out of the model.

3.4. Decentralised Fine-tuning: Combining Open and Closed Systems

We suggest here a novel approach to reinforcement learning with a combination of open-source and closed datasets. This would create customized intelligence capabilities. As discussed earlier, the open-source dataset would rely on the legal community at large including law schools, legal clinics, industry partners, and research users who can contribute to the open model. Inputs are decentralized only by legal professionals to distil legal principles into the language models rather than misinformation from the general public.

As for the closed dataset, it would be drawn from industry partners' proprietary data and feedback. While proprietary data cannot be disclosed, the two systems will learn from each other and improve the underlying generalized legal model. Such a system would involve training the language model over remote devices or siloed data centers, such as mobile phones or law firms servers, while keeping data localized [26].

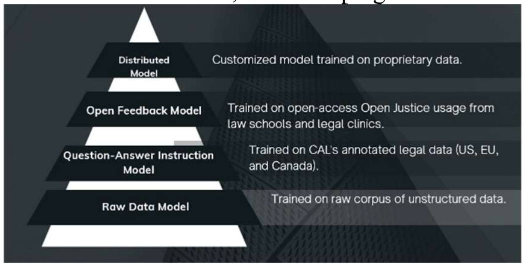

Distributed Model
Customized model trained on proprietary data.
Open Feedback Model
Trained on open-access Open Justice usage from law schools and legal clinics.
Question-Answer Instruction Model
Trained on CAL's annotated legal data (US, EU, and Canada).
Raw Data Model
Trained on raw corpus of unstructured data.

Figure 3. Multilayered Fine-tuning

Larger-scale initiatives such as OpenAI or Hugging Faces are, by contrast, taking different approaches, i.e., to develop an Artificial General Intelligence. However, while the underlying rationale is different, these models can serve as solid foundational models to be built upon for domain specific models. As for smaller scale projects, such as Lexis + AI or Harvey, they rely primarily on proprietary closed datasets. We believe the consequences of that will be embedded bias and inconsistent performance. Thus, open-source initiatives are in some ways more modest in the sense that we are limited to distilling principles established from law into language models. However, they might show better performance as well as societal benefits for the legal community and self-represented litigants. In fact, early evidence shows fine-tuning language models for law shows impressive results with fewer hallucinations and more explainable results with better citation retrieval [2], [3], [19], [27]. With that in mind, we note that legal citation retrieval techniques are still limited and cause LLMs to poorly answer questions on incorrectly sourced raw data. At the time of writing, this remains a question that deserves further research [28], [29].

4. Conclusion

We have demonstrated a strong argument that generalized AI such as ChatGPT cannot and should not be used for legal tasks. Its use presents significant risks for the legal professions as well as litigants. However, it should not be ruled out. It has potential for legal research as well as access to justice. This paper called for the development of an open-source and distributed legal AI accessible to the entire legal community. We believe it has the potential to address some limitations related to the use of general AI for legal problems and resolving disputes -- shortcomings that include legal misinformation or hallucinations, lack of transparency and precision, and inability to offer diverse and multiple narratives. Early evidence on fine-tuning shows impressive results with fewer hallucinations and more explainable results. With that in mind, many questions deserve further research. In particular, domain-specific LLMs research calls for empirical findings on performance. Having an industry specific measuring stick to gauge performance is essential. There must be clear metrics to assess how much an AI hallucinates or whether it can provide its chain of reasoning with diverse narratives and accurate citations. In addition, research must reflect on the human-AI collaboration component of LLMs. In the legal context, formulating effective prompts is challenging for non-experts, especially when it comes to legal inquiries. Future research should address these concerns related to open access legal generative AI by investigating the ability of non-lawyers to engage effectively in "end-user prompt engineering."

References

[1] L. Chen, M. Zaharia, and J. Zou, “How is ChatGPT’s behavior changing over time?,” *arXiv preprint arXiv:2307.09009*, 2023.

[2] R. Bhambhoria, S. Dahan, and X. Zhu, “Investigating the State-of-the-Art Performance and Explainability of Legal Judgment Prediction.,” in *Canadian Conference on AI*, 2021.

[3] C. F. Luo, R. Bhambhoria, S. Dahan, and X. Zhu, “Evaluating Explanation Correctness in Legal Decision Making,” in *Proceedings of the Canadian Conference on Artificial Intelligence (5 2022). https://doi. org/10.21428/594757db. 8718dc8b*, 2022.

[4] R. Bhambhoria, H. Liu, S. Dahan, and X. Zhu, “Interpretable low-resource legal decision making,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022, pp. 11819–11827.

[5] C. F. Luo, R. Bhambhoria, S. Dahan, and X. Zhu, “Prototype-Based Interpretability for Legal Citation Prediction.” arXiv, May 25, 2023. doi: 10.48550/arXiv.2305.16490.

[6] Samuel Dahan, Maxime Cohen, and Colin Rule, “Conflict Analytics: When Data Science Meets Dispute Resolution,” *Management Business Review*, forthcoming 2020.

[7] Behnam Manavi Tehrani, Maxime Cohen, Samuel Dahan, and Jonathan Touboul, “‘Independent contractor’ vs employee: worker classification with machine learning,” *Forthcoming*, 2020.

[8] M. C. Cohen, S. Dahan, W. Khern-Am-Nuai, H. Shimao, and J. Touboul, “The use of AI in legal systems: determining independent contractor vs. employee status,” *Artificial intelligence and law*, pp. 1–30, 2023.

[9] Samuel Dahan, Jonathan Touboul, Jason Lam, and Dan Sfedj, “Predicting Employment Notice Period with Machine Learning: Promises and Limitations,” *McGill Law Journal*, 2020.

[10] J. Lam, Y. Chen, F. Zulkernine, and S. Dahan, “Detection of Similar Legal Cases on Personal Injury,” in *2021 International Conference on Data Mining Workshops (ICDMW)*, IEEE, 2021, pp. 639–646.

[11] Samuel Dahan, Yifei Yin, and Farhana Zulkernine, “Determining Worker Type from Legal Text Data using Machine Learning,” *Pervasive Intelligence and Computing (IEEE PICom)*, 2020.

[12] J. T. Lam, D. Liang, S. Dahan, and F. H. Zulkernine, “The Gap between Deep Learning and Law: Predicting Employment Notice.,” in *NLLP@ KDD*, 2020, pp. 52–56.

[13] C. Gordon, “ChatGPT is the fastest growing app in the history of web applications,” *Forbes*, 2023.

[14] Melissa Heikkilä, “Why you shouldn’t trust AI search engines,” *MIT Technology Review*, Feb. 2013. Accessed: Aug. 20, 2023. [Online]. Available: https://www.technologyreview.com/2023/02/14/1068498/why-you-shouldnt-trust-ai-search-engines/

[15] J. Elias, “Google execs warn company’s reputation could suffer if it moves too fast on AI-chat technology,” *CNBC*, Dec. 13, 2022. Accessed: Aug. 20, 2023. [Online]. Available: https://www.cnbc.com/2022/12/13/google-execs-warn-of-reputational-risk-with-chatgbt-like-tool.html

[16] R. Maruf, “Google fires engineer who contended its AI technology was sentient | CNN Business,” *CNN*, Jul. 23, 2022. Accessed: Aug. 20, 2023. [Online]. Available: https://www.cnn.com/2022/07/23/business/google-ai-engineer-fired-sentient/index.html

[17] M. Shur-Ofry, “Multiplicity as an AI Governance Principle.” Rochester, NY, May 10, 2023. doi: 10.2139/ssrn.4444354.

[18] D. M. Katz, M. J. Bommarito, S. Gao, and P. Arredondo, “GPT-4 Passes the Bar Exam.” Rochester, NY, Mar. 15, 2023. doi: 10.2139/ssrn.4389233.

[19] Jed Stiglitz, “Modeling Legal Reasoning: LM Annotation at the Edge of Human Agreement,” *Working Paper*, 2023.

[20] J. Kleinberg, J. Ludwig, S. Mullainathan, and C. R. Sunstein, “Discrimination in the Age of Algorithms,” *Journal of Legal Analysis*, vol. 10, pp. 113–174, Dec. 2018, doi: 10.1093/jla/laz001.

[21] Y. Chen, M. Andiappan, T. Jenkin, and A. Ovchinnikov, “A Manager and an AI Walk into a Bar: Does ChatGPT Make Biased Decisions Like We Do?” Rochester, NY, May 20, 2023. doi: 10.2139/ssrn.4380365.

[22] C. Markou and S. Deakin, “Ex Machina Lex: Exploring the Limits of Legal Computability.” Rochester, NY, Jun. 21, 2019. doi: 10.2139/ssrn.3407856.

[23] S. Wu *et al.*, “BloombergGPT: A Large Language Model for Finance.” arXiv, May 09, 2023. doi: 10.48550/arXiv.2303.17564.

[24] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.” arXiv, May 24, 2019. doi: 10.48550/arXiv.1810.04805.

[25] E. A. Van Dis, J. Bollen, W. Zuidema, R. van Rooij, and C. L. Bockting, “ChatGPT: five priorities for research,” *Nature*, vol. 614, no. 7947, pp. 224–226, 2023.

[26] T. Li, A. K. Sahu, A. Talwalkar, and V. Smith, “Federated learning: Challenges, methods, and future directions,” *IEEE signal processing magazine*, vol. 37, no. 3, pp. 50–60, 2020.

[27] C. F. Luo, R. Bhambhoria, S. Dahan, and X. Zhu, “Prototype-Based Interpretability for Legal Citation Prediction.” arXiv, May 25, 2023. doi: 10.48550/arXiv.2305.16490.

[28] J. Hilton, R. Nakano, S. Balaji, and J. Schulman, “WebGPT: Improving the factual accuracy of language models through web browsing,” *OpenAI Blog, December*, vol. 16, 2021.

[29] R. Nakano *et al.*, “Webgpt: Browser-assisted question-answering with human feedback,” *arXiv preprint arXiv:2112.09332*, 2021.