JOURNAL OF
CROSS-DISCIPLINARY RESEARCH
IN COMPUTATIONAL LAW [CRCL]

Promises and pitfalls of artificial intelligence for legal applications

Sayash Kapoor*, Peter Henderson**, Arvind Narayanan***

Abstract

Is AI set to redefine the legal profession? We argue that this claim is not supported by the current evidence. We dive into AI's increasingly prevalent roles in three types of legal tasks: information processing, tasks involving creativity, reasoning, or judgment, and predictions about the future. We find that the ease of evaluating legal applications varies greatly across legal tasks based on the ease of identifying correct answers and the observability of information relevant to the task at hand. Tasks that would lead to the most significant changes to the legal professional are not only harder to evaluate; they are also most prone to overoptimism about AI capabilities. We make recommendations for better evaluation and deployment of AI in legal contexts.

Keywords: Generative AI, Predictive AI, Evaluation

Journal of Cross-disciplinary Research in Computational Law

DOI: pending

www.journalcrcl.org

* Computer science Ph.D. candidate at the Center for Information Technology Policy, Princeton University. sayashk@princeton.edu
** Assistant Professor at Princeton University with appointments in the Department of Computer Science and School of Public and International Affairs. peter.henderson@princeton.edu
*** Professor of Computer Science at Princeton University. arvindn@cs.princeton.edu

Introduction

DoNotPay, a U.S.-based AI startup, claimed to sell the services of a 'robot lawyer' to help customers prepare legal documents, contest parking tickets and cancel subscriptions [DoNotPay 2023a]. On January 8, 2023, CEO Joshua Browder claimed that the company would pay USD 1 million to any lawyer who used DoNotPay's robot lawyer to argue a U.S. Supreme Court case by using an earpiece to repeat the arguments made by the company's software [Browder 2023a]. Even setting aside the fact that the Supreme Court prohibits electronics in the courtroom, the U.S. has several laws prohibiting the unauthorised practice of law by individuals who are not licensed attorneys. Soon after the announcement, the CEO backed down [Browder 2023b], and the term 'robot lawyer' was changed to 'AI consumer champion' on the company's website [DoNotPay 2023b]. Still, the company is facing multiple class-action lawsuits [Pacheco 2023].

This was far from the first time when technology was claimed to replace a lawyer, and it will not be the last. After all, the company had been claiming to sell the services of a robot lawyer for more than four years. Claims about lawyers being replaced by digital technology predate the company. A 2011 New York Times headline read: 'Armies of Expensive Lawyers, Replaced by Cheaper Software.' [Markoff 2011] Since the article was published, the number of lawyers in the U.S. has actually *increased* by eight percent [Statista Research Department 2023]. How do we separate true advances from hype?

In this position paper, we argue that the kinds of legal applications we can legitimately use AI for should be determined by the evaluations that reflect these uses of AI in the real world [Hagan 2023; Linna Jr 2021a,b]. It is easy to get caught up in the hype, particularly for impressive demonstrations of generative AI that can be used to create text, images, or other forms of media. Many recent instances of AI that have received widespread attention are examples of generative AI [Anthropic 2023; Meta 2023; OpenAI 2023a]. In the law, some of this attention has focused on claims of improvements in the legal reasoning ability of text-based language models, including OpenAI's claims that GPT-4 can pass the bar exam. Yet, this is not evidence that GPT-4 is becoming as capable as lawyers: after all, it is not a lawyer's job to answer bar exam questions all day. While generative AI is our main focus, in this paper, we also consider AI used to predict the outcomes of court cases and make decisions about people (such as AI used for predicting a defendant's risk of recidivism).

The types of legal AI we analyze roughly correspond to the types of AI outlined in Diver et al.'s typology of legal applications [Diver et al. 2022], though at a coarser level of granularity. We analyze three broad uses of AI in the legal domain: (i) tasks involving information processing, such as summarization or legal information retrieval; (ii) tasks involving creativity, reasoning, or judgment, such as preparing legal filings; and (iii) tasks involving predictions about the future, such as criminal risk prediction as well as predicting the outcomes of court decisions. Of course, the lines separating these applications are blurry, but the high-level categories can offer useful insights about how AI applications should be evaluated and how useful they can be in the real world.

These applications vary in how difficult they are to evaluate [Hagan 2023]. For some, evaluation is relatively easy. For example, a tool that categorises a request for legal advice into particular areas of law (an example of an information-processing task) can be evaluated by comparing against corresponding labels from lawyers performing the same task. [Stanford Legal Design Lab and Suffolk LIT Lab 2018] In contrast, there is no clear 'correct' answer for other types of AI. For instance, if generative AI is used to prepare a legal filing (an example of a task involving creativity, reasoning, or judgment), there is no single correct answer on how the document should be written– reasonable people can disagree on what strategies to take. Tasks that are harder to evaluate also tend to be those that would lead to the most significant changes in the legal profession. If AI could be useful for consequential legal tasks like preparing legal filings, that would have much broader implications for the future of legal professionals compared to labelling text for different areas of law.

In our analysis, we examine the challenges that arise in meaningful AI evaluations in legal settings and offer recommendations for overcoming them. We argue that evaluations should be used to identify how well AI performs on a given task and which types of tasks it can be useful for.

Information processing

Many legal tasks involve processing information. Examples include summarizing court cases or long legal documents, translating text from one language to another, redacting sensitive information from documents before broader release, e-discovery to find relevant documents for litigation and legal information retrieval.

With the widespread adoption of generative AI, there have been many claims that it will revolutionise legal information processing. Compared to the other types of legal tasks we consider in the next two sections, evaluating information processing tasks is more straightforward. This is because:

- there is generally a *clear correct answer*: given information about the features used in the model and the model's output, it is easy to determine if the model's output is correct, such as in the task of categorizing legal requests by area of law [Stanford Legal Design Lab and Suffolk LIT Lab 2018][1] and
- there is *high observability* of the features relevant for decision making: The features relevant for using AI for information-processing tasks are available as inputs to the AI system.

These factors make it easier to develop valid evaluations for AI used for information processing. As a result, generative AI for information-processing tasks can be deployed based on evidence and robust evaluations. Still, claims about generative AI being a revolution might be overstated, and several nuances make a blanket assessment of generative AI for information processing hard.

For legal experts, generative AI for information processing is an evolution, not a revolution. A major reason why chatbots are exciting to the general public is that they can be instructed in natural language to perform tasks for which software may not have previously existed. But for those tasks where natural-language processing software already existed, the advent of large language models has generally led to an evolutionary improvement in accuracy. In law, software for information-processing tasks is not new. Automated tools for legal summarization have existed for over a decade [Markoff 2011]. The same goes for many other information processing tasks like legal document search, with entire companies built on the promise of automating information processing dating back decades. Recent instruction-tuned language models (chatbots) cannot necessarily outperform models fine-tuned on law-specific datasets [Chalkidis 2023]. Further, many information-processing tasks can also be carried out by professionals without a law degree. For these reasons, while large language models offer improvements over existing tools — possibly in terms of accuracy but especially in terms of cost, by decreasing the amount of task-specific software development required — they do not drastically change legal information processing for experts.

We need to better understand how generative AI impacts laypeople. The ability of chatbots to follow natural language directions means laypeople can use them to perform information-processing tasks, such as translation or getting pointers to relevant legal rules. Everyday users have increasingly turned to technology for legal advice in the past—for instance, a 2019 survey in the U.S. found that while 31% of the people used the internet (31%) for legal advice, only 29% relied on lawyers (29%), and that 63% of the people surveyed used information they found on the internet as a factor to resolve their legal problems [Gramatikov et al. 2021].

Yet, there is a paucity of evidence about how chatbots affect users who turn to them for information-processing tasks such as legal information retrieval or translation. Understanding how well they work is hard without naturalistic evaluations of everyday users who use chatbots. Errors in the outputs of chatbots on such tasks can be catastrophic. For example, asylum applications for refugees can be rejected if machine translation introduces errors because they cannot accurately infer context [Deck 2023]. It is unclear how people are using generative AI for such tasks. Research should help inform best practices for the use of AI by laypeople.

[1] There are, of course, some exceptions where evaluation is more ambiguous even within information processing, but the majority of cases in this category will be more straightforward to evaluate.

Unresolved limitations make the adoption of language models challenging. Language models for information processing suffer several unresolved issues that may pose challenges in shifting from existing solutions to language-model-based ones. A key limitation is their propensity to output incorrect information, often known as hallucinations [Lee et al. 2019; Zhao et al. 2020]. This is a significant hurdle in their adoption in consequential legal settings. While there are many ongoing efforts to improve factual accuracy [Shuster et al. 2021], it is as yet an unsolved research problem. As a result, the outputs of language models must be closely verified before they can be used in consequential settings.

Some information-processing tasks are harder to evaluate than others. Even within information-processing tasks, ease of evaluation is a spectrum. For categorizing cases by area of law, legal experts can label the correct answer [Stanford Legal Design Lab and Suffolk LIT Lab 2018], but in cases where there might be multiple areas of law implicated, experts might have higher rates of disagreement. Similarly, for tasks involving transcription or redaction, it is sometimes easy to create a clear source of ground truth based on past data. Yet, in adversarial settings, lawyers might disagree on how much context to redact and litigate over the issues. In *Kaiser Aluminum Warrick, LLC v. US Magnesium LLC*,[2] for example, parties disputed how much information should be redacted in documents produced during discovery and ultimately the court ordered the producing party to unredact information that was relevant to the case. For translation, evaluations must account for inherent ambiguity—such as when a source language uses gendered terms and a target language does not, when there is a lack of context to disambiguate a term, or when an idiom does not have a clear, direct translation. And parties dispute how e-discovery systems are evaluated, with requesting parties generally seeking to discover more information and producing parties wanting to reveal less[Guha, Henderson, et al. 2022]. Nonetheless, such disagreements are generally over atypical cases, and the bulk of information processing tasks will have consensus answers when polling a larger pool of annotators.

[2] WL 2482933 (S.D.N.Y. Feb. 27, 2023)

Creativity, reasoning, or judgment

Several legal tasks involve creativity, reasoning, or judgment. They range from tasks involving writing, such as preparing drafts of legal filings, to tasks involving judgment, such as automated mediation and dispute resolution. These tasks typically involve significant expertise and labour to get right. In contrast to information processing, if AI could indeed automate such tasks, the impact on the legal profession might be huge. When OpenAI announced its GPT-4 language model, it claimed the model could pass a 'simulated bar exam with a score around the top 10% of test takers' [Martínez 2023]. This led to much speculation about whether AI would soon replace lawyers, presumably because the tool could perform tasks requiring expertise and creativity.

But what does a high score on the bar exam mean—and more generally, how much can we trust benchmark evaluations? Here, we outline several concerns underlying evaluations of language models in legal settings that make it hard to trust their applicability to real-world legal tasks. We then provide recommendations for improving evaluations and outline tasks for which AI can be evaluated well and is arguably underutilised.

Hurdles in evaluating language models

Contamination

Contamination refers to including the same data in the training and evaluation data sets for a model [Brown et al. 2020; Magar and Schwartz 2022]. This can lead to overoptimistic estimates of model performance since a model can simply memorise solutions in its training set instead of being able to answer new questions. It is possible that evaluations such as OpenAI's claims about bar exam performance are overoptimistic due to contamination, but it is hard to know for sure due to the training and fine tuning data being proprietary.

However, as an illustration of the plausibility and seriousness of contamination, consider a different benchmark that OpenAI evaluated GPT-4 on. To benchmark its coding ability, OpenAI evaluated it on problems from Codeforces, a website that hosts coding competitions. The training data cutoff for the original GPT-4 model was September 2021. The model could correctly answer most Codeforces questions from before its training date cutoff, but could not answer questions after its training date cutoff correctly [He 2023]. This strongly suggests that the model memorised solutions from its training set—or at least partly memorised them, enough to fill in what it couldn't recall. That is, instead of developing the capability to answer *new* coding questions, it could only answer questions it had already been trained on. (The Codeforces results in OpenAI's technical report on GPT-4 were not affected by this, as OpenAI used problems from recent Codeforces competitions, resulting in the model being evaluated on fresh problems not in the training set. Sure enough, GPT-4 performed very poorly [OpenAI 2023b].)

To be clear, a temporal discontinuity in benchmark performance, such as in the case of GPT-4's performance on Codeforces, strongly implies contamination, but the lack of such a discontinuity does not imply the opposite. Without access to the data used to train and fine tune a model, researchers can only make informed guesses about the absence of contamination since there is no guarantee that a model is not already trained on later versions of a benchmark. For example, OpenAI could fine tune GPT-4 on more recent versions of the bar exam (even inadvertently) if a user inputs exam questions into ChatGPT.

Lack of construct validity

Construct validity refers to the extent to which an evaluation accurately represents and measures the construct it is designed to assess. For the bar exam, the construct might be the extent to which a lawyer has the necessary preparation to serve clients effectively. The assumption is that humans taking exams generalise the skills tested by the exam to a wider range of relevant tasks.

Unfortunately, it is well known that bar exam questions are not representative of the tasks professionals do in the real world—something that critics of the bar exam regularly lament, resulting in the recent restructuring of the bar exam [Sloan 2023] and proposals for alternative pathways to certification based on real-world training [Ching and Hershkowitz 2023]. Specifically, the bar exam overemphasises subject-matter knowledge and underemphasises real-world skills, which are far harder to measure in a standardised, computer-administered way. In other words, not only does it emphasise the wrong thing, it overemphasises precisely the thing that language models are good at.

Memorisation is a spectrum. Even if a language model has no exposure to an exact problem in a training set, it has inevitably seen examples that are pretty close, simply because of the size of the training corpus. That means it can get away with a much shallower level of reasoning. This issue is also referred to as *task contamination* [Li and Flanigan 2023]. As a result, legal benchmarks don't necessarily give us evidence that language models are acquiring the kind of in-depth reasoning skills that human test-takers might have. While inferring legal reasoning skills from standardised exams might already be somewhat dubious for humans, it is unfounded for language models that might take all sorts of shortcuts [Geirhos et al. 2020] and memorise key information to come to the right answer without generalising in any way.

In some real-world tasks, shallow reasoning may be sufficient—for example, it could be enough to build a chatbot to help applicants prepare for the bar exam where similar scenarios have played out thousands of times in textbooks and court cases. But the world is constantly changing, so if a bot is asked to analyze the legal consequences of a new fact pattern in the context of new judicial decisions, it does not have much to draw upon. In short, tests designed for humans lack construct validity when applied to bots.

Benchmarks are already wildly overused in AI for comparing different models [Raji et al. 2021]. They have been heavily criticised for collapsing a multidimensional evaluation into a single number [Thomas and Uminsky 2022]. As we've discussed, using benchmarks to compare humans to AI introduces a further set of problems. If an AI developer's goal is to predict how well it will do on real-world legal tasks, measuring bar exam performance is not a suitable approach.

Prompt sensitivity

Another issue with evaluating language models is their sensitivity to the user's prompts. Small changes to the prompt can significantly impact the model's outputs [Guha, Nyarko, et al. 2023]. To construct valid evaluations, it is important to understand how language models are used. Unfortunately, we are entirely in the dark about how these models are being used in the real world. Since model developers do not share information about model use, we currently have few ways to study many important questions about language models.

Prompt sensitivity leads to several challenges in AI evaluation. First, in programmatic use, where developers are writing prompts for legal applications (instead of legal professionals or end users directly using a language model), performance could improve with better prompting, so measured results provide a lower bound of how well the tools work. In some cases, performance could also degrade as the model's behaviour changes over time [Chen et al. 2023; Narayanan and Kapoor 2023b]. Second, in use by legal professionals, prompt sensitivity means that results are conditional on users being trained on proper prompting techniques. Recent large-scale evaluations of language model performance start to expand the scope of evaluations on a wider range of legal tasks [Guha, Nyarko, et al. 2023], but even in these cases, benchmark creators pick a fixed set of prompts that are used across evaluations. It is possible that a user, particularly those not knowledgeable enough about the legal domain or the limitations of language models, could see drastically different performance on the same tasks if they do not craft their prompt in the same way as the evaluation benchmark. Even ordering few-shot examples in a prompt differently can affect performance by double-digit percentage points [Lu et al. 2021]. Third, in use by non-professionals, the state of evaluation is even worse. The lack of naturalistic datasets means that we cannot evaluate how often chatbots respond to legal questions with useful answers as opposed to irrelevant or inaccurate ones since we do not know how everyday users interact with these models in the real world.

Recommendations for developers of legal AI

Improve construct validity by involving legal experts in evaluation

Many current evaluations of large language models (LLMs) are general purpose: they measure the efficacy of language models on general tasks such as summarization, retrieval, or factuality. However, these evaluations do not tell us much about how LLMs can aid legal professionals in their day-to-day tasks. The involvement of legal experts in designing and conducting evaluations is necessary to improve the status quo [Hagan 2023]. Without their involvement, benchmarks for testing language models on legal tasks will likely suffer from construct validity issues.

Such evaluations can be both quantitative and qualitative. An interdisciplinary group of lawyers and AI experts created the LegalBench benchmark for evaluating language models on various legal reasoning tasks [Guha, Nyarko, et al. 2023]. This is an example of a quantitative evaluation created by professionals to measure the usefulness of generative AI in their profession. But there are reasons to think that qualitative studies of professionals and how they could use AI are likely to be even more useful, since these tools are so new that we still need consensus on what the right questions to ask are. To our knowledge, such qualitative studies have not yet been conducted for legal professionals. However, in other professions, notably medicine, several such studies have been conducted, which can inform such evaluations in the law [Abouammoh et al. 2023; Nayak et al. 2023; Noy and Zhang 2023].

Develop naturalistic evaluation methods

As outlined in our discussion of prompt sensitivity, a major limiting factor in current evaluations of language models is the lack of transparency around how users actually use these models on a day-to-day basis [Bommasani, Klyman, et al. 2023; Narayanan and Kapoor 2023a]. Without knowing how users interact with LLMs, it is hard to understand what limitations must be addressed and how evaluations can best be constructed to represent typical use cases. To improve the construct validity of current evaluations and prevent evaluations from falling prey to prompt sensitivity, researchers can conduct naturalistic evaluations of people using LLMs that closely model their use in the real

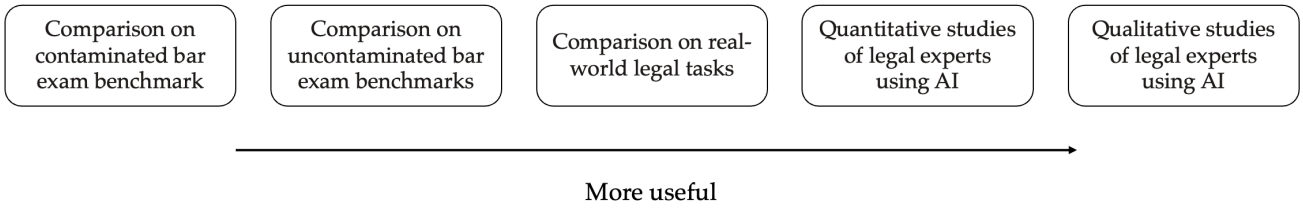

Figure 1: ***Types of evaluations of generative AI****. Current evaluations of AI are often based on exam benchmarks meant for humans, such as the bar exam, and suffer from contamination: overlaps between the training and evaluation datasets. Comparing the performance of these models on real-world tasks, especially those curated by legal experts, is more likely to be useful. Since the use of generative AI is nascent, qualitative studies that observe how legal experts use these tools for day-to-day tasks are likely to be a more useful, if expensive, way of evaluating these tools.*

world. For example, Zheng et al. [2023] released a dataset of user conversations with 25 different LLMs over three months. Similar datasets collecting real-world interactions with users asking legal questions would improve our understanding of how users use LLMs for legal tasks and, in turn, improve evaluations.

Communicate the limitations of current LLMs

Recent cases of lawyers misusing language models have made the headlines [Wagner 2023; Weiser 2023]. Language models can fabricate information even while presenting it authoritatively [Zhao et al. 2020]. When users are unaware of these limitations, it could result in severe professional damage. Several lawyers have been sanctioned for fabricating information in legal filings. Even when a language model is trained on accurate text, such as a filtered dataset of past legal documents, it is not guaranteed to produce accurate outputs [Dorf 2023]. These cases highlight the need for better communication of these limitations for end users by companies providing these services [Vincent 2023]. Developers have added some disclaimers to language models to reduce such errors. For example, OpenAI says, 'ChatGPT can make mistakes. Consider checking important information' in a small font at the bottom of the ChatGPT chatbox. Anthropic goes one step further. Its disclaimer is more clear about the limitations ('Claude is in beta release and may display incorrect or harmful information,'). When the output contains URLs, there is also a disclaimer about links potentially being inaccurate. Some judges have also issued chambers' rules to clarify how lawyers should explicitly account for their use of AI [Donald et al. 2022].

Use AI in narrow settings with well-defined outcomes and high observability of evidence.

In a more constrained, highly issue-specific and low-stakes setting, it may be possible to construct a thorough evaluation. One type of application that meets these criteria is checking errors in various legal documents and filings [Bommasani, Hudson, et al. 2021]. The Social Security Administration already uses a simple model to spot issues with decisions that might lead to a remand of the judgment on appeal [Glaze et al. 2021]. One mistake flagged by the automated system is when the adjudicator's opinion does not address a medical claim made in a benefits claim in their denial of benefits decision. Such a mistake would almost certainly result in a remand of the decision on appeal. The system does not need any additional information beyond the benefits claim and the decision text to make such an assessment. That is, the system operates under full observability, allowing thorough evaluations to be conducted.

Similar technology could be developed and deployed in a wide range of settings where easy-to-spot errors in initial filings are prevalent. In particular, over 86% of patent applications received at least one non-final rejection [Carley et al. 2015], so semi-automated checks for common errors could reduce costs to both the filing party and the United States Patent and Trademark Office. Nonetheless, even in these cases, automated judgments should be implemented with extreme caution. Deployments should be structured to favour helpful, informative recommendations to parties in a dispute rather than being used as a binding mechanism. And a thorough appeals process should be available.

These tasks are distinct from the more general case of using AI to predict court case outcomes, a more problematic application that we discuss in the next section. First, they are constrained to a single or small handful of issues, which makes it possible to sample sufficient data to cover typical use cases. Second, the model has (or should ideally have) access to the same information as the adjudicator. This is typically not true of general-purpose judgment prediction tasks.

AI for making predictions about the future

In recent years, over a hundred research papers have claimed to predict court outcomes using AI based on text from court proceedings [Medvedeva and Mcbride 2023]. Such predictive abilities could be useful to lawyers in guiding legal strategy or businesses to assess potential litigation risks. AI has also been used to make consequential decisions about people, most notably pre-trial detention and parole in criminal justice. In this section, we identify shortcomings that plague these applications and question the use of predictions in legal settings.

Predicting the outcomes of court decisions

Medvedeva and Mcbride systematically review 171 papers claiming to predict court decisions [Medvedeva and Mcbride 2023]. They find severe shortcomings in the literature they review. Their main finding is that the vast majority of papers claiming to predict the outcomes of court judgments do not try to solve this problem at all. In many cases, the papers solve a related but ultimately less helpful problem: they use the judgment text containing the final judgment to 'predict' the verdict. Since the text of the final judgment includes the verdict, these studies do not provide real-world evidence of the usefulness of AI in judgment prediction. In sum, only 12 of 171 papers (7%) end up carrying out their claimed task of predicting court decisions.

This study follows a smaller-scale study to evaluate predictions of court decisions, where Medvedeva, Wieling, et al. [2023] point out that such errors could be caused by insufficient knowledge of the datasets being used in judgment classification and inadequate steps taken to filter out information about the verdict from the dataset. This highlights the need for both legal and AI expertise for useful applications of AI in legal settings. Moreover, for the small minority of papers that actually predict court outcomes, the accuracy of the resulting models is much lower.

The low accuracy demonstrates that automating judgments from the text of legal cases is hard. This is not surprising: legal outcomes depend on the context and specifics of cases, the available documents might not comprise the entirety of the context of the case being adjudicated, and the specific judgment might depend on a specific judge's (or set of judges') interpretation of the arguments. In addition, there is significant variability across different jurisdictions, meaning the amount of data that can be used to train AI to automate judgments in any specific jurisdiction is small. Finally, the judgments made over time evolve with changes to the specific judges, the set of past cases comprising precedent, legislation and many other factors.

Medvedeva and Mcbride's findings also point to the problem of contamination. Since the text of the judgment also contains the verdict, the model essentially has access to the answers while making predictions—like teaching to the test, this vastly inflates the accuracy of the resulting models, leading to exaggerated performance estimates. In other cases, even if the final judgment text is excluded, the input to the prediction model uses the statement of facts prepared *after* the verdict. Since the verdict informs this statement of facts, it constitutes leakage. This is a well-known issue in machine learning. In traditional machine learning research, it is called *data leakage* or simply *leakage* [Kaufman et al. 2012], and it affects hundreds of papers across scientific fields. While there are no perfect solutions for fixing leakage, there are several steps researchers can take to prevent leakage in their models [Kapoor and Narayanan 2023].

This does not even begin to address the potential for biases, sensitivity to inputs and other challenges for evaluating legal judgment prediction tasks. The challenges with evaluation should limit where and how judgment prediction tasks are used. A well-evaluated judgment prediction sys-

tem could be used to better understand what properties of briefs could lead to poor outcomes (e.g., finding common errors). This would serve as a suggestion to attorneys that might miss common errors but not result in any binding outcome and could be ignored by the attorney.

Predictive AI for making decisions

In addition to research claiming to predict court outcomes using AI, AI-based predictions are also used to make decisions about people. We call such applications predictive AI. Predictive AI suffers from pervasive shortcomings that may nullify the claimed benefits. A closer account of these shortcomings can help us understand why such systems fail.

Low accuracy of deployed applications. A common application of predictive tools in criminal justice is to predict recidivism. A 2016 ProPublica investigation found that COMPAS, a widely used algorithm to predict the risk of recidivism for defendants, had twice as many false positives for Black defendants as White ones [Angwin et al. 2016]. Perhaps more surprisingly, the investigation found that the overall accuracy of the algorithm was only around 65%. In a follow-up study, Dressel and Farid [2018] found that this accuracy was no more accurate than predictions made by people without any background in criminal justice. Notably, the majority of defendants predicted to be at high risk of committing violent crimes do not go on to recidivate. These simple models distil into these few features a model of a person's entire future life for the next few years. They have no access to private information, like a defendant's mental state or intentions, nor can they model defendants' attempts to seek help.

Another problem is *distribution shift*: when the data used to train an ML model differs from the population on which the model is eventually deployed, models are unable to adapt well. A machine learning tool called Public Safety Assessment (PSA) is used in U.S. courts in over half the states. Like COMPAS, if the tool predicts that a defendant has a high risk of re-offending, bail could be denied. PSA is trained on data from 1.5 million cases across the country. But crime patterns in specific regions differ from nationwide averages in important ways, which means that it fails catastrophically in some areas. Corey [2019] highlights that in Cook County, Illinois, the rate of violent recidivism is *ten times* lower than the nationwide data that was used for training PSA. Distribution shift is an open research problem in machine learning [Geirhos et al. 2020], and affects most predictive AI applications where the population of interest differs from training data [Wang et al. 2024].

Where dynamics are known and stable over time, and information is readily available, prediction is possible—as in physical sciences, where we can build reliable approximations of aspects of the world that we are modelling. Yet this is not true of predictive AI in law, where fundamentally, most predictions will be about people and societies.

Predictive AI has even more limitations in practice. Vendors sell predictive AI based on the promise of full automation and elimination of jobs, but when it performs poorly, they retreat to the fine print, which says that the tool should not be used on its own. The responsibility for ensuring that a predictive AI system works well is spread thinly across multiple stakeholders, often deliberately [Martineau 2022]. The individual decisions made by these systems also tend not to be contestable by decision subjects, as vendors claim that the logic of the tool is a trade secret [Jackson and Mendoza 2020; Moore 2017]. In most cases, decision makers (such as court systems) do not develop predictive tools in house—tools that might be tailored to their specific needs and those of the populations that they serve. Instead, they purchase or license one-size-fits-all products from AI vendors. This exacerbates issues with evaluations since the decision subjects or civil rights advocates cannot easily push back against vendors' claims.

These issues are not specific to the examples we list above. In an analysis of eight predictive AI applications across domains, Wang et al. [2024] found that these issues are widespread in domains such as finance, insurance, child welfare and medicine, in addition to criminal justice. Given the propensity of such applications to failure, predictive AI in the legal domain needs to be held to a much higher standard to ensure that it functions as its developers claim. This requires much stronger transparency by the developers, clear mechanisms to ensure contestability to decision subjects, and evaluations that go beyond just the technical specifications of these tools into the societal impact of these tools.

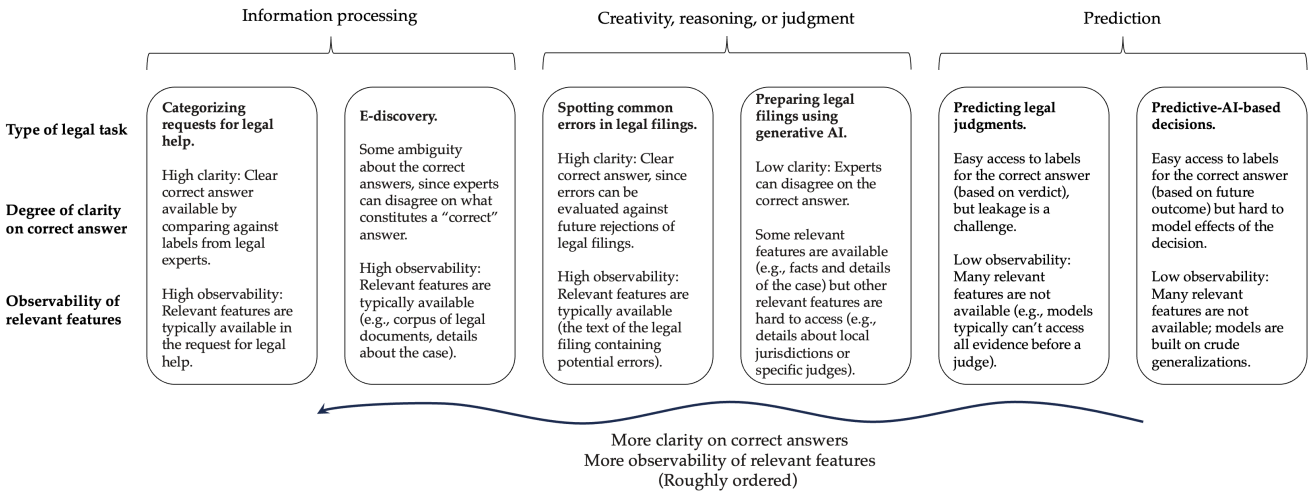

Figure 2: ***Variation in the difficulty of evaluating AI for legal tasks****. We categorise difficulty along two dimensions: clarity on correct labels and observability of relevant features. Some tasks, such as AI for categorizing requests for legal help by area of law, have clear correct answers, whereas for other tasks, such as preparing legal filings using AI, there is no clear right answer, which makes evaluation hard. Similarly, for some applications, all relevant features are available, such as for spotting common errors in legal filings. For others, relevant features are not (or cannot be) available, such as for predictive AI. As we proceed from right to left, the clarity of correct answers and observability of relevant features roughly increases.*

The use of AI for prediction, whether court decisions or recidivism, fundamentally differs from information processing tasks and tasks involving creativity, reasoning, or judgment. They attempt to predict the future without sufficient observability of relevant features and lack data to form a robust model of the world that would allow for accurate predictions. Instead, they rely on extremely rough generalizations and approximations using simple linear models (when the underlying dynamics are far from linear).

Conclusion

The effective deployment of AI in legal contexts requires shifting from technical evaluations to robust sociotechnical assessments carried out in the specific context in which an AI system would be deployed. While past machine learning applications did not consider such evaluations because they were cost prohibitive, this change is necessary due to the complex nature and societal impact of AI applications in the legal field. Figure 2 illustrates how the difficulty in evaluating legal applications of AI varies over the three types of tasks we discussed.

To answer the question 'What can I use an AI system for?', it is essential first to answer 'How was this AI system evaluated?'. Unfortunately, the current state of AI evaluations leaves much to be desired.

Acknowledgments

We thank Mireille Hildebrandt and the attendees of the Cross-disciplinary Research in Computational Law (CRCL 2023) conference for their feedback. We also thank attendees of the World Intellectual Property Organization Judges Forum and the Harvard Journal of Law and Technology speaker series for their feedback on talks based on this paper. We are grateful to Angelina Wang, Emily Cantrell, Matthew J. Salganik and Solon Barocas for the conversations and collaborations that informed this paper.

Parts of this paper are based on blog posts by two of the authors (`https://aisnakeoil.com`).

References

Noura Abouammoh et al.. 2023. 'Exploring Perceptions and Experiences of ChatGPT in Medical Education: A Qualitative Study Among Medical College Faculty and Students in Saudi Arabia'. Pages: 2023.07.13.23292624.

Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. *Machine Bias*. Retrieved Nov. 5, 2023 from `https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing`.

Anthropic. 2023. *Introducing Claude*. Retrieved Nov. 5, 2023 from `https://www.anthropic.com/index/introducing-claude`.

Rishi Bommasani, Drew A Hudson, et al.. 2021. 'On the opportunities and risks of foundation models'. arXiv:2108.07258 [cs].

Rishi Bommasani, Kevin Klyman, Shayne Longpre, Sayash Kapoor, Nestor Maslej, Betty Xiong, Daniel Zhang, and Percy Liang. 2023. 'The Foundation Model Transparency Index'. arXiv:2310.12941 [cs].

Joshua Browder. 2023a. *DoNotPay will pay any lawyer or person $1,000,000 with an upcoming case in front of the United States Supreme Court to wear AirPods and let our robot lawyer argue the case by repeating exactly what it says. (1/2)*. Retrieved Nov. 5, 2023 from `https://twitter.com/jbrowder1/status/1612312707398795264`.

Joshua Browder. 2023b. *Good morning! Bad news: after receiving threats from State Bar prosecutors, it seems likely they will put me in jail for 6 months if I follow through with bringing a robot lawyer into a physical courtroom. DoNotPay is postponing our court case and sticking to consumer rights:* Retrieved Nov. 5, 2023 from `https://twitter.com/jbrowder1/status/1618265395986857984`.

Tom Brown et al.. 2020. 'Language models are few-shot learners'. *Advances in neural information processing systems*, 33, 1877–1901.

Michael Carley, Deepak Hedge, and Alan Marco. 2015. 'What is the probability of receiving a US patent'. *Yale JL & Tech.*, 17, 203.

Ilias Chalkidis. 2023. 'ChatGPT may Pass the Bar Exam soon, but has a Long Way to Go for the LexGLUE benchmark'. arXiv:2304.12202 [cs].

Lingjiao Chen, Matei Zaharia, and James Zou. 2023. 'How is ChatGPT's behavior changing over time?' arXiv:2307.09009 [cs].

Audrey Ching and Donna Hershkowitz. 2023. *Report from the Alternative Pathway Working Group: Request to Circulate for Public Comment*. Los Angeles Office, California State Bar. Retrieved Nov. 8, 2023 from `https://www.courthousenews.com/wp-content/uploads/2023/09/california-bar-exam-alternative-proposal.pdf`.

Ethan Corey. 2019. *How a Tool to Help Judges May Be Leading Them Astray*. Retrieved Nov. 5, 2023 from `https://theappeal.org/how-a-tool-to-help-judges-may-be-leading-them-astray/`.

Andrew Deck. 2023. *AI translation is jeopardizing Afghan asylum claims*. Retrieved Jan. 8, 2024 from `https://restofworld.org/2023/ai-translation-errors-afghan-refugees-asylum/`.

Laurence Diver et al.. 2022. 'Typology of Legal Technologies: Cross-disciplinary Research in Computational Law (CRCL)'.

Hon. Bernice Bouie Donald, Hon. James C. Francis IV, Ronald J. Hedges, and Kenneth J. Withers. 2022. *Generative AI and Courts: How Are They Getting Along?* Retrieved Nov. 6, 2023 from `https://www.jamsadr.com/blog/2023/francis-james-pli-generative-ai-1023`.

DoNotPay. 2023a. *DoNotPay - The World's First Robot Lawyer*. Retrieved Nov. 5, 2023 from `https://web.archive.org/web/20230101170502/https://donotpay.com/`.

DoNotPay. 2023b. *DoNotPay - Your AI Consumer Champion*. Retrieved Nov. 5, 2023 from `https://web.archive.org/web/20230730013643/https://donotpay.com/`.

Michael C. Dorf. 2023. *Law-Specific Large Language Model Generative AI Interim Report: Lexis+AI Versus GPT-4*. Retrieved Nov. 6, 2023 from `https://www.dorfonlaw.org/2023/11/law-specific-large-language-model.html`.

Julia Dressel and Hany Farid. 2018. 'The accuracy, fairness, and limits of predicting recidivism'. *Science Advances*, 4, 1, eaao5580.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. 2020. 'Shortcut learning in deep neural networks'. *Nature Machine Intelligence*, 2, 11, 665–673.

Kurt Glaze, Daniel E Ho, Gerald K Ray, and Christine Tsang. 2021. 'Artificial Intelligence for Adjudication: The Social Security Administration and AI Governance'.

Martin Gramatikov, Rodrigo Núñez, Isabella Banks, Maurits Barendrecht, Jelmer Brouwer, Brittany Kauffman, and Logan Cornett. 2021. *Justice Needs and Satisfaction in the United States of America*. Retrieved Jan. 8, 2024 from `https://iaals.du.edu/publications/justice-needs-and-satisfaction-united-states-america`.

Neel Guha, Peter Henderson, and Diego Zambrano. 2022. 'Vulnerabilities in Discovery Tech'. *Harvard Journal of Law & Technology*, 35.

Neel Guha, Julian Nyarko, et al.. 2023. 'LegalBench: A Collaboratively Built Benchmark for Measuring Legal Reasoning in Large Language Models'. arXiv:2308.11462 [cs].

Margaret Hagan. 2023. 'Good AI Legal Help, Bad AI Legal Help: Establishing quality standards for responses to people's legal problem stories'. In: *JURIX*. Vol. 2023, 36th.

Horace He. 2023. *I suspect GPT-4's performance is influenced by data contamination, at least on Codeforces. Of the easiest problems on Codeforces, it solved 10/10 pre-2021 problems and 0/10 recent problems. This strongly points to contamination. 1/4 https://t.co/wm6yP6AmGx*. Retrieved Nov. 9, 2023 from `https://twitter.com/cHHillee/status/1635790330854526981`.

Eugenie Jackson and Christina Mendoza. 2020. 'Setting the Record Straight: What the COMPAS Core Risk and Need Assessment Is and Is Not'. *Harvard Data Science Review*, 2, 1.

Sayash Kapoor and Arvind Narayanan. 2023. 'Leakage and the reproducibility crisis in machine-learning-based science'. *Patterns*, 4, 9.

Shachar Kaufman, Saharon Rosset, Claudia Perlich, and Ori Stitelman. 2012. 'Leakage in data mining: Formulation, detection, and avoidance'. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 6, 4, 1–21.

Katherine Lee, Orhan Firat, Ashish Agarwal, Clara Fannjiang, and David Sussillo. 2019. *Hallucinations in Neural Machine Translation*. `https://openreview.net/forum?id=SkxJ-309FQ`.

Changmao Li and Jeffrey Flanigan. 2023. 'Task Contamination: Language Models May Not Be Few-Shot Anymore'. arXiv:2312.16337v1 [cs].

Daniel W. Linna Jr. 2021a. 'Evaluating Artificial Intelligence for Legal Services: Can "Soft Law" Lead to Enforceable Standards for Effectiveness?' *IEEE Technology and Society Magazine*, 40, 4, 37–51.

Daniel W. Linna Jr. 2021b. 'Evaluating legal services: The need for a quality movement and standard measures of quality and value'. In: *Research Handbook on Big Data Law*. Edward Elgar Publishing, 404–431.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2021. 'Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity'. arXiv:2104.08786.

Inbal Magar and Roy Schwartz. 2022. 'Data Contamination: From Memorization to Exploitation'. arXiv:2203.08242 [cs].

John Markoff. 2011. 'Armies of Expensive Lawyers, Replaced by Cheaper Software'. *The New York Times*.

Paris Martineau. 2022. *Toronto Tapped Artificial Intelligence to Warn Swimmers. The Experiment Failed*. Retrieved Nov. 5, 2023 from https://www.theinformation.com/articles/when-artificial-intelligence-isnt-smarter.

Eric Martínez. 2023. 'Re-Evaluating GPT-4's Bar Exam Performance'.

Masha Medvedeva and Pauline Mcbride. 2023. 'Legal Judgment Prediction: If You Are Going to Do It, Do It Right'. In: *Proceedings of the Natural Legal Language Processing Workshop 2023*. Association for Computational Linguistics, Singapore, 73–84.

Masha Medvedeva, Martijn Wieling, and Michel Vols. 2023. 'Rethinking the field of automatic prediction of court decisions'. *Artificial Intelligence and Law*, 31, 1, 195–212.

Meta. 2023. *Make-A-Video*. Retrieved Nov. 5, 2023 from https://makeavideo.studio/.

Taylor Moore. 2017. *Trade Secrets and Algorithms as Barriers to Social Justice*. Retrieved Jan. 8, 2024 from https://cdt.org/insights/trade-secrets-and-algorithms-as-barriers-to-social-justice/.

Arvind Narayanan and Sayash Kapoor. 2023a. *Generative AI companies must publish transparency reports*. Retrieved Nov. 5, 2023 from http://knightcolumbia.org/blog/generative-ai-companies-must-publish-transparency-reports.

Arvind Narayanan and Sayash Kapoor. 2023b. *Is GPT-4 getting worse over time?* Retrieved Jan. 10, 2024 from https://www.aisnakeoil.com/p/is-gpt-4-getting-worse-over-time.

Ashwin Nayak, Matthew S. Alkaitis, Kristen Nayak, Margaret Nikolov, Kevin P. Weinfurt, and Kevin Schulman. 2023. 'Comparison of History of Present Illness Summaries Generated by a Chatbot and Senior Internal Medicine Residents'. *JAMA Internal Medicine*, 183, 9, 1026–1027.

Shakked Noy and Whitney Zhang. 2023. 'Experimental evidence on the productivity effects of generative artificial intelligence'. *Science*, 381, 6654, 187–192.

OpenAI. 2023a. *DALL·E 3*. Retrieved Nov. 5, 2023 from https://openai.com/dall-e-3.

OpenAI. 2023b. 'GPT-4 Technical Report'. arXiv:2303.08774 [cs].

Stephanie Pacheco. 2023. *ANALYSIS: DoNotPay Lawsuits: A Setback for Justice Initiatives?* Retrieved Nov. 5, 2023 from https://news.bloomberglaw.com/bloomberg-law-analysis/analysis-donotpay-lawsuits-a-setback-for-justice-initiatives.

Deborah Raji, Emily Denton, Emily M. Bender, Alex Hanna, and Amandalynne Paullada. 2021. 'AI and the Everything in the Whole Wide World Benchmark'. *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, 1.

Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. 'Retrieval Augmentation Reduces Hallucination in Conversation'. In: *Findings of the Association for Computational Linguistics: EMNLP 2021*. Ed. by Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih. Association for Computational Linguistics, Punta Cana, Dominican Republic, 3784–3803.

Karen Sloan. July 19, 2023. *New bar exam gets lukewarm reception in previews*. (July 19, 2023). Retrieved Nov. 8, 2023 from https://www.reuters.com/legal/legalindustry/new-bar-exam-gets-lukewarm-reception-previews-2023-07-19/.

Stanford Legal Design Lab and Suffolk LIT Lab. 2018. *Learned Hands*. Retrieved Dec. 27, 2023 from https://learnedhands.law.stanford.edu.

Statista Research Department. 2023. *U.S.: number of lawyers 2007-2022*. Retrieved Nov. 5, 2023 from https://www.statista.com/statistics/740222/number-of-lawyers-us/.

Rachel L. Thomas and David Uminsky. 2022. 'Reliance on metrics is a fundamental challenge for AI'. *Patterns*, 3, 5, 100476.

James Vincent. 2023. *OpenAI isn't doing enough to make ChatGPT's limitations clear*. Retrieved Nov. 5, 2023 from https://www.theverge.com/2023/5/30/23741996/openai-chatgpt-false-information-misinformation-responsibility.

David Wagner. 2023. *This Prolific LA Eviction Law Firm Was Caught Faking Cases In Court. Did They Misuse AI?* Retrieved Nov. 5, 2023 from https://laist.com/news/housing-homelessness/dennis-block-chatgpt-artificial-intelligence-ai-eviction-court-los-angeles-lawyer-sanction-housing-tenant-landlord.

Angelina Wang, Sayash Kapoor, Solon Barocas, and Arvind Narayanan. 2024. 'Against predictive optimization: On the legitimacy of decision-making algorithms that optimize predictive accuracy'. *ACM Journal on Responsible Computing*, 1, 1, 1–45.

Benjamin Weiser. 2023. *Here's What Happens When Your Lawyer Uses ChatGPT*. Retrieved Nov. 5, 2023 from https://www.nytimes.com/2023/05/27/nyregion/avianca-airline-lawsuit-chatgpt.html.

Zheng Zhao, Shay B. Cohen, and Bonnie Webber. 2020. 'Reducing Quantity Hallucinations in Abstractive Summarization'. In: *Findings of the Association for Computational Linguistics: EMNLP 2020*. Ed. by Trevor Cohn, Yulan He, and Yang Liu. Association for Computational Linguistics, Online, 2237–2249.

Lianmin Zheng et al.. 2023. 'LMSYS-Chat-1M: A Large-Scale Real-World LLM Conversation Dataset'. arXiv:2309.11998 [cs].