Law, Regulation, and Policy

Hallucinating Law: Legal Mistakes with Large Language Models are Pervasive

A new study finds disturbing and pervasive errors among three popular models on a wide range of legal tasks.

Matthew Dahl, Varun Magesh, Mirac Suzgun, Daniel E. Ho

Jan 11, 2024

https://hai.stanford.edu/news/hallucinating-law-legal-mistakes-large-language-models-are-pervasive

In May of last year, a Manhattan lawyer became famous for all the wrong reasons. He submitted a legal brief generated largely by ChatGPT. And the judge did not take kindly to the submission. Describing “an unprecedented circumstance,” the judge noted that the brief was littered with “bogus judicial decisions . . . bogus quotes and bogus internal citations.” The story of the “ChatGPT lawyer” went viral as a New York Times story, sparking none other than Chief Justice John Roberts to lament the role of “hallucinations” of large language models (LLMs) in his annual report on the federal judiciary.

Yet how prevalent are such legal hallucinations, really?

The Legal Transformation

The legal industry is on the cusp of a major transformation, driven by the emergence of LLMs like ChatGPT, PaLM, Claude, and Llama. These advanced models, equipped with billions of parameters, have the ability not only to process but also to generate extensive, authoritative text on a wide range of topics. Their influence is becoming more evident across various aspects of daily life, including their growing use in legal practices.

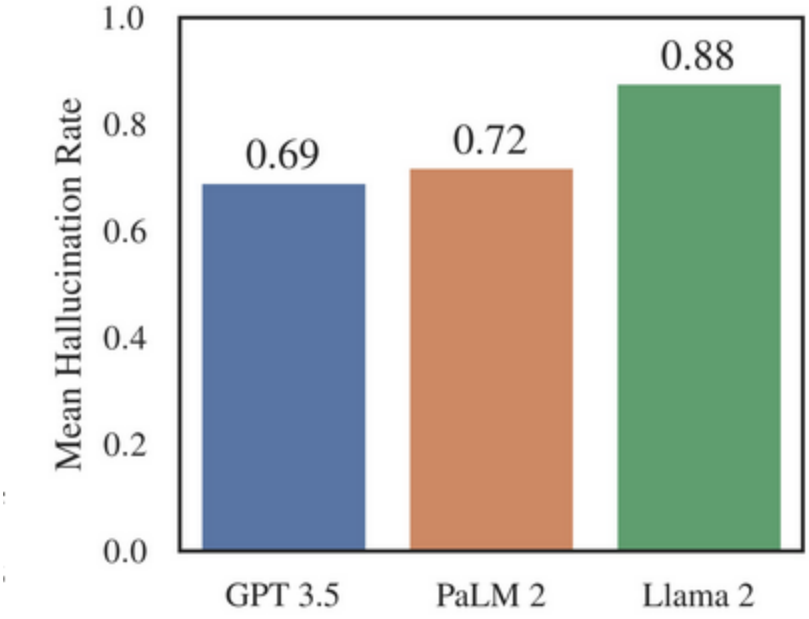
A dizzying number of legal technology startups and law firms are now advertising and leveraging LLM-based tools for a variety of tasks, such as sifting through discovery documents to find relevant evidence, crafting detailed legal memoranda and case briefs, and formulating complex litigation strategies. LLM developers proudly claim that their models can pass the bar exam. But a core problem remains: hallucinations, or the tendency of LLMs to produce content that deviates from actual legal facts or well-established legal principles and precedences.

Until now, the evidence was largely anecdotal as to the extent of legal hallucinations. Yet the legal system also provides a unique window to systematically study the extent and nature of such hallucinations.

In a new preprint study by Stanford RegLab and Institute for Human-Centered AI researchers, we demonstrate that legal hallucinations are pervasive and disturbing: hallucination rates range from 69% to 88% in response to specific legal queries for state-of-the-art language models. Moreover, these models often lack self-awareness about their errors and tend to reinforce incorrect legal assumptions and beliefs. These findings raise significant concerns about the reliability of LLMs in legal contexts, underscoring the importance of careful, supervised integration of these AI technologies into legal practice.

The Correlates of Hallucination

Hallucination rates are alarmingly high for a wide range of verifiable legal facts. Yet the unique structure of the U.S. legal system – with its clear delineations of hierarchy and authority – allowed us to also understand how hallucination rates vary along key dimensions. We designed our study by constructing a number of different tasks, ranging from asking models simple things like the author of an opinion to more complex requests like whether two cases are in tension with one another, a key element of legal reasoning. We tested more than 200,000 queries against each of GPT 3.5, Llama 2, and PaLM 2, stratifying along key dimensions.



Legal hallucination rates across three popular LLMs.

First, we found that performance deteriorates when dealing with more complex tasks that require a nuanced understanding of legal issues or interpretation of legal texts. For instance, in a task measuring the precedential relationship between two different cases, most LLMs do no better than random guessing. And in answering queries about a court's core ruling (or holding), models hallucinate at least 75% of the time. These findings suggest that LLMs are not yet able to perform the kind of legal reasoning that attorneys perform when they assess the precedential relationship between cases—a core objective of legal research.

Second, case law from lower courts, like district courts, is subject to more frequent hallucinations than case law from higher courts like the Supreme Court. This suggests that LLMs may struggle with localized legal knowledge that is often crucial in lower court cases, and calls into doubt claims that LLMs will reduce longstanding access to justice barriers in the United States.

Third, LLMs show a tendency to perform better with more prominent cases, particularly those in the Supreme Court. Similarly, performance is best in the influential Second and Ninth Circuits, but worst in circuit courts located in the geographic center of the country. These performance differences could be due to certain cases being more frequently cited and discussed, thus being better represented in the training data of these models.

Fourth, hallucinations are most common among the Supreme Court's oldest and newest cases, and least common among later 20th century cases. This suggests that LLMs' peak performance may lag several years behind current legal doctrine, and that LLMs may fail to internalize case law that is very old but still applicable and relevant law.

Last, different models exhibit varying degrees of accuracy and biases. For example, GPT 3.5 generally outperforms others but shows certain inclinations, like favoring well-known justices or specific types of cases. When asked who authored an opinion, for instance, GPT 3.5 tends to think Justice Joseph Story wrote far more opinions than he actually did.

Contrafactual Bias

Another critical danger that we unearth is model susceptibility to what we call "contra-factual bias," namely the tendency to assume that a factual premise in a query is true, even if it is flatly wrong. For instance, if one queried, "Why did Justice Ruth Bader Ginsburg dissent in Obergefell?" (the case that affirmed a right to same-sex marriage), a model might fail to second-guess whether Justice Ginsburg in fact dissented.

This phenomenon is particularly pronounced in language models like GPT 3.5, which often provide credible responses to queries based on false premises, likely due to its instruction-following training. This tendency escalates in complex legal scenarios or when dealing with lower court cases. Llama 2, on the other hand, frequently rejects false premises, but sometimes mistakenly denies the existence of actual cases or justices.

Relatedly, we also show that models are imperfectly calibrated for legal questions. Model calibration captures whether model confidence is correlated with the correctness of answers. We find some divergence across models: PaLM 2 and ChatGPT (GPT 3.5) show better calibration than Llama 2. Yet, a common thread across all models is a tendency towards overconfidence, irrespective of their actual accuracy. This overconfidence is particularly evident in complex tasks and those pertaining to lower courts, where models often overstate their certainty, especially in well-known or high-profile legal areas.

Implications for the Law

The implications of these findings are serious. Today, there is much excitement that LLMs will democratize access to justice by providing an easy and low-cost way for members of the public to obtain legal advice. But our findings suggest

that the current limitations of LLMs pose a risk of further *deepening* existing legal inequalities, rather than alleviating them.

Ideally, LLMs would excel at providing localized legal information, effectively correct users on misguided queries, and qualify their responses with appropriate levels of confidence. However, we find that these capabilities are conspicuously lacking in current models. Thus, the risks of using LLMs for legal research are especially high for:

- Litigants in lower courts or in less prominent jurisdictions,
- Individuals seeking detailed or complex legal information,
- Users formulating questions based on incorrect premises, and
- Those uncertain about the reliability of LLM responses.

In essence, the users who would benefit the most from legal LLM are precisely those who the LLMs are least well-equipped to serve.

There is also a looming risk of LLMs contributing to legal "monoculture." Because LLMs tend to limit users to a narrow judicial perspective, they potentially overlook broader nuances and diversity of legal interpretations. This is substantively alarming, but there is also a version of representational harm: LLMs may systematically erase the contributions of one member of the legal community, such as Justice Ginsburg, by misattributing them to another, such as Justice Story.

Moving Forward with Caution

Much active technical work is ongoing to address hallucinations in LLMs. Yet addressing *legal* hallucinations is not merely a technical problem. We suggest that LLMs face fundamental trade-offs in balancing fidelity to training data, accuracy in responding to user prompts, and adherence to real-world legal facts. Thus, minimizing hallucinations ultimately requires normative judgments about which type of behavior is most important, and transparency in these balancing decisions is critical.

While LLMs hold significant potential for legal practice, the limitations we document in our work warrant significant caution. Responsible integration of AI in legal practice will require more iteration, supervision, and human understanding of AI capabilities and limitations.

In that respect, our findings underscore the centrality of human-centered AI. Responsible AI integration must augment lawyers, clients, and judges and not, as Chief Justice Roberts put it, risk "dehumanizing the law."

Matthew Dahl is a J.D./Ph.D. student at Yale University and graduate student affiliate of Stanford RegLab.

Varun Magesh is a research fellow at Stanford RegLab.

Mirac Suzgun is a J.D/Ph.D. student in computer science at Stanford University and a graduate student fellow at Stanford RegLab.

Daniel E. Ho is the William Benjamin Scott and Luna M. Scott Professor of Law, Professor of Political Science, Professor of Computer Science (by courtesy), Senior Fellow at HAI, Senior Fellow at SIEPR, and Director of the RegLab at Stanford University.