The Ground Truth about Legal Hallucinations

June 10, 2024

Extended Abstract

Despite their uncanny ability to generate plausible and fluent text, LLMs remain intrinsically incapable of understanding such text or of evaluating the veracity or correctness of their outputs. Their propensity to "hallucinate," to generate text or responses that seem syntactically sound, fluent, and plausible but are factually incorrect or nonsensical is often regarded as a side-effect their language generation objective. It can also be regarded as one of the main roadblocks to their integration into legal workstreams.

The general assumption underlying discussions of hallucinations, at least in technical literature, is that the generated output can be evaluated with reference to some ground truth, a verifiable set of facts or generally accepted knowledge. In such instance, hallucinations are generally synonymous with incorrect or false statements. When deploying LLMs for tasks involving the application of substantive legal knowledge, however, it is often difficult to compare the output to a ground truth and thus confidently declare that it constitutes a hallucination. In the case of many legal tasks, such as legal QA or contract drafting, there may be no single, accepted ground truth. It is often difficult to unequivocally state what the law is or what the answer should be, especially in complex domains governed by a multitude of legal sources.

"Legal hallucinations" are extremely difficult to detect, especially given that the generated output often appears helpful, relevant, and informative. This promotes unnecessary risk-taking in an area where even the smallest mistake can snowball into lawsuits or result financial losses. The difficulty of detecting and hence measuring "legal hallucinations" also hinders reliable evaluations of a model's suitability to assist lawyers.

The final paper will briefly address of the sources of hallucinations, with particular emphasis on the low-quality of the legal knowledge stored in the parameters of LLMS. It will also confront the inconsistent definitions and taxonomies of hallucinations. The traditional meaning of the term connotes a detachment from reality and thus implies easy detection. The opposite is true of hallucinations generated by LLMs, where even nonsensical output may have the semblance of veracity and authority.

Despite their ubiquity, the categories of open-domain and closed-domain hallucinations remain ill-defined and are applied inconsistently, introducing additional confusion into an already complex field. In the legal context, it is less important to determine whether the output is consistent with the training corpus or with the prompt. Training corpora abound in incorrect information, prompts often contain false premises. What matters is whether the output is consistent with legal doctrine and presents a feasible solution to the legal problem at hand. Arguably then, it may often be more advantageous that the LLMs disregard their parametric knowledge and/or the prompt. Taxonomies aside, the question remains: what should the generated output be compared against in the absence of a clear ground truth? Who determines whether the output constitutes a hallucination? If subjective opinions come into play, is it even possible to speak of hallucinations? In the legal context, unless the generated answer is nonsensical ("contracts are stochastic parrots!") or contradicts legal doctrine ("breach of contract is a punishable offence!"), determining whether a particular output constitutes a hallucination is challenging if not impossible. The problem is unrelated to the technical attributes of LLMs but derives from the nature of legal knowledge and the difficulty of applying the law.

The main focus of the paper will be to demonstrate the need for a domain-specific approach to "legal hallucinations." Existing literature generally addresses the problem in scenarios where the generated output can be evaluated with reference to a ground truth or where a deviation from such ground truth is tolerable or even desirable. In the context of high-risk domains, as exemplified by law, traditional technical approaches are difficult to apply and may lead to an unintended obfuscation of the risks of using LLMs. Apart from criticizing existing attempts to categorize hallucinations, the paper will demonstrate the practical impossibility of developing methodologies to reliably detect and measure the existence of hallucinations in those instances, where the term implies a deviation from a ground truth. The correctness of the generated output cannot be open to interpretation. In the legal area, however, it frequently is. References to "correctness" are best replaced with "legal feasibility." An answer is legally feasible when it remains

within the constraints of the law, or the range of possible legal approaches to a problem or question and when it represents the output of logical reasoning. While the introduction of a new term does not solve the problem of hallucinations, at least it distances the discussion away from the misleading concept of "correctness" and emphasizes the subjective nature of hallucinations.

The challenges of detecting legal hallucinations are illustrated with two situations involving legal question answering. Situation 1 concerns questions about objective legal facts, such as “does case X exist” or “what does case X say?” Situation 2 concerns questions about a legal problem, such as “what happens if contractual performance involves the commission of illegal acts” or “are indemnities primary or secondary obligations under the Laws of England and Wales”? Situation 1 can be broadly associated with legal research. Situation 2 resembles the provision of legal advice, which requires not just knowledge of the law but also the ability to reason about the law.

In Situation 1, the generated answer can be evaluated as either correct or not. There is an undisputable ground truth. When inquiring about the existence of a case, its main ruling or the judge who decided it, there is only one possible correct answer. The existence and contents of legal sources, such as cases or statutes, are objective facts and are easy to verify. They also constitute a ground truth, against which the output can be evaluated. If the generated output misrepresents or contradicts the contents (or very existence) of a legal source, the LLM can be said to hallucinate. To clarify, Situation 1 does not require an interpretation or a reconciliation of legal sources. It focuses on their existence and contents. A study examined whether LLMs could generate accurate information concerning legal sources equated legal hallucinations with output inconsistent with legal facts, such as statutes and cases. While commendable for its disciplined approach, the study presents a very narrow view of legal hallucinations as it only measures the correctness of responses to verifiable questions about federal court cases and does not require any interpretation of the law, reconciliation of legal sources or legal reasoning.

Situation 2 involves questions that can have more than one correct answer or answers that cannot be evaluated with reference to a single legal ground truth. Situation 2 requires knowledge of the law and legal reasoning, including the interpretation and reconciliation of legal rules. It introduces subjective undertones as well as references to competence and domain-expertise. Many legal questions can have multiple answers that reflect different legal approaches to the application of legal principles or different lines of reasoning leading to different outcomes. The interpretation of primary legal sources, such as cases and statutes, is often subject to vigorous disagreements culminating in court appeals, scholarly articles and academic conferences. The text of a statute or a case is a question of fact. A legal source “says what it says.” The meaning and practical implications of a legal source can, however, be amendable to multiple interpretations. Even codified legal rules are intrinsically open-textured, dynamic, uncertain and incomplete. Consequently, the existence of hallucinations may lie in the eye of the beholder and depend on his or her knowledge, understanding and interpretation of the law.

To clarify: the fact that many answers to legal questions cannot be evaluated with reference to a ground truth does not mean that it is impossible to evaluate such answers in general. It means, however, that such evaluation is extremely challenging. Absent a legal ground truth, one can only speak of evaluating – and ranking - the generated answers on the basis of individual preferences based on legal expertise. Such ranking would, however, be highly subjective and extremely resource intensive as it would require domain experts agreeing which answers are, for example, closest to established legal doctrine and thus most likely to succeed in court. This casts doubts on the practicality of deploying LLMs in legal tasks that involve questions of substantive law, which lack a ground truth. It may be less resource intensive (speak: costly) to employ an experienced lawyer to handle a specific legal query than to verify the generated answer, not to mention - to employ a group of experts to evaluate and rank its quality.

Given the absence of a “legal ground truth,” determining whether the generated answer constitutes a hallucination or an unusual but feasible approach to a legal problem requires an understanding how it was arrived at. Explainability is, however, unrelated to the ability to determine which words activate specific connections in the neural network or to evaluate the low-level calculations involved in token prediction. It concerns the reasoning underlying the output as well as the sources relied on. Unfortunately, LLMs are inherently incapable of explaining the provenance of their outputs. They can be prompted to provide an explanation, but such “explanation” will be the result of a language generation task, not a retrospective exploration of the model’s reasoning.

Any attempts to mitigate hallucinations and to improve the quality of the outputs generated by LLMs, must be preceded by a clear definition of the problem and an acknowledgment of certain domain-specific challenges.

References

1. Adam Bouyamourn, Why LLMs Hallucinate, And How To Get (Evidential) Closure: Perceptual, Intensional and Extensional Learning for Faithful Natural Language Generation (2023) ACL

2. Matthew Dahl et al., Large Legal Fictions: Profiling Legal Hallucinations in Large Language Models, ARXIV (Jan. 4, 2024) https://arxiv.org/pdf/2401.01301.pdf

3. Ronald Dworkin, No Right Answer? 53 N.Y.U. L. REV. 1 (1978)

4. H.L.A. HART, THE CONCEPT OF LAW 127–28 (1961)

5. Ari Holtzman et al., The Curious Case of Neural Text Degeneration, ARXIV (Oct. 2, 2023), https://arxiv.org/pdf/2310.01693.pdf.

6. Lei Huang et al., A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. 2023. arXiv: 2311.05232

7. Ziwei Ji et al.,. "Survey of Hallucination in Natural Language Generation". In: ACM Computing Surveys 55.12 (Mar. 2023).

8. Samia Kabir, et al., Who Answers It Better? An In-Depth Analysis of ChatGPT and Stack Overflow Answers to Software Engineering Questions, ARXIV (Aug. 10, 2023) https://arxiv.org/pdf/2308.02312.pdf.

9. Sayash Kapoor, Peter Henderson, Arvind Narayanan, Promises and pitfalls of artificial intelligence for legal applications, ARXIV (Jan. 10, 2024) https://arxiv.org/pdf/2402.01656.pdf.

10. Noam Kolt, Predicting Consumer Contracts, 37 BERKELEY TECH. L. J. 71 (2021)

11. Stephen Lin et al., TruthfulQA: Measuring How Models Mimic Human Falsehoods, PROC. 60th ANN. MEET. ASSOC'N. COMPUTATIONAL LINGUISTICS 3214 (2021)

12. Joshua Maynez et al., On Faithfulness and Factuality in Abstractive Summarization, PROC. 58TH ANN MEETING ASSOC'N COMPUTATIONAL LINGUISTICS 1906 (2020).

13. Michael O'Neill Mark Connor, Amplifying Limitations, Harms and Risks of Large Language Models, ARXIV (Jul. 6, 2023) https://arxiv.org/pdf/2307.04821.pdf.

14. Lichao Sun et al., TrustLLM: Trustworthiness in Large Language Models, ARXIV (Jan. 25, 2024), https://arxiv.org/pdf/2401.05561.pdf

15. Alex Tamkin et al., Understanding the Capabilities, Limitations, and Societal Impact of Large Language Models, ARXIV (Feb. 4, 2021), https://arxiv.org /abs /2102.02503.

16. Adam Tauman, Kalai Santosh S. Vempala, Calibrated Language Models Must Hallucinate (Dec. 5, 2023)

17. S. M Towhidul Islam Tonmoy, et al. A Comprehensive Survey of Hallucination Mitigation Techniques in Large Language Models, ARXIV (Jan. 8, 2024) https://arxiv.org/pdf/2401.01313.pdf

18. Varshney, Neeraj et al. A Stitch in Time Saves Nine: Detecting and Mitigating Hallucinations of LLMs by Validating Low-Confidence Generation, ARXIV (Aug. 12, 2023) https://arxiv.org/pdf/2307.03987.pdf

19. Pranav Narayanan Venkit et al., "Confidently Nonsensical?": A Critical Survey on the Perspectives and Challenges of 'Hallucinations' in NLP (11 April, 2024)

20. Boxin Wang et al, DECODING TRUST: A Comprehensive Assessment of Trustworthiness in GPT Models, ARXIV (Jan. 5, 2024), https://arxiv.org/pdf/2306.11698.pdf.

21. Ziwei Xu et al, Hallucination is Inevitable: An Innate Limitation of Large Language Models (22 Jan 2024) arXiv:2401.11817v1 [cs.CL] 22 Jan 2024

22. Jiaxin Zhang et al., 2023a. SAC: Reliable hallucination detection in black-box language models via semantic-aware cross-check consistency. In Findings of the Association for Computational Linguistics: EMNLP 2023, pages 15445–15458, Singapore

23. Muru Zhang et al., How Language Model Hallucinations Can Snowball, ARXIV (May. 22, 2023) https://arxiv.org/abs/2305.13534

24. Yue Zhang et al., Siren's Song in the AI Ocean: A Survey on Hallucination in Large Language Models. 2023. arXiv: 2309.01219.