The Ground Truth About Legal Hallucinations
(a work in-progress)

Eliza MIK

1. Introduction

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 natural but are factually incorrect or nonsensical[1] is often regarded as a side-effect their language generation objective.[2] It can also be regarded as one of the main roadblocks to their integration into legal workstreams.[3] The general assumption underlying discussions of hallucinations, at least in technical literature, is that the generated output can be evaluated with reference to a ground truth, a verifiable set of facts or generally accepted knowledge.[4] In such instance, hallucinations are generally synonymous with incorrect or false statements.[5] 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, especially in complex domains governed by a multitude of legal sources.

In the legal context, the problem is not that LLMs hallucinate but that it is extremely challenging to detect and to measure "legal hallucinations," especially in light of the fact that the generated output often appears helpful, relevant, and informative.[6] 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 measuring "legal hallucinations" also precludes reliable evaluations of a model's suitability to assist lawyers.

The final 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 their subjective, expertise-dependent nature.

---

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

[2] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. *The Curious Case of Neural Text Degeneration,* ARXIV (Oct. 2, 2023), https://arxiv.org/pdf/2310.01693.pdf.

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

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

[5] Ziwei Xu et al, Hallucination is Inevitable: An Innate Limitation of Large Language Models (22 Jan 2024); for a study of the definitions and common understanding of hallucinations in technical literature, see: Pranav Narayanan Venkit et al., "Confidently Nonsensical?": A Critical Survey on the Perspectives and Challenges of 'Hallucinations' in NLP (11 April, 2024)

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

2. Sources of Hallucinations

When extolling the future potential of LLMs it is often forgotten what those systems are designed and trained to do. An LLM "is simply a probability distribution D over sequences of tokens, i.e., words or other character sequences. Any [LLM] which predicts every string with positive probability will necessarily hallucinate with positive probability."[7] LLMs do not make decisions or judgements. They cannot reason and have no common sense – they make statistically-informed guess about the next word.[8] They are insensitive to the truth of the word sequences they predicts.[9] For an LLM, there simply is no right or wrong answer.[10] Apart from the simple fact that LLMs are trained and designed to predict the next word based on the probability distributions learned during training, technical scholarship has singled out multiple sources of hallucinations.[11] Such sources range from unreliable and outdated training data and/or problems with the input text in the prompt, [12] as well as "collisions" between the parametric knowledge that the LLM has acquired during pre-training and the new knowledge that it is fine-tuned with[13] or provided by means of retrieval augmented generation, or "RAG," during inference.[14] Hallucinations can thus be seen as a natural consequence of the LLM's language generation objective. It must be remembered that given their statistical nature, LLMs do not operate at the level of semantic information.[15] They model word distributions in a language, but they cannot understand language.[16] Unsurprisingly, it has been established that LLMs are sensitive to output probability and "perform better when the correct answer is a high-probability string than when it is a low-probability string, even in deterministic situations where the answer could be determined without considering probability."[17] Consequently, LLMs seem to "prefer" common outputs to correct outputs: the number of times a particular word string expressing a given fact or factoid was mentioned in the training corpus is more important than its correctness. [18]

The point here is not, however, to delve into the sources of hallucinations in LLMs. This seems to be a purely technical problem, a problem that must commence with the simple acknowledgement that even under ideal conditions, all autoregressive transformers will hallucinate. Arguably, *all* that LLMs do when generating text is hallucinate! The point here is that apart from the technical characteristics of LLMs, additional challenges stem from the characteristics of certain specialised domains. One such domain is law, and the the problem with detecting or even defining hallucinations in this domain lies in the very nature of legal knowledge.

3. Definitions

The traditional definition of *hallucinate* is "to seem to see, hear, feel, or smell something that does not exist, usually because of a health condition or because you have taken a drug."[19] Traditional hallucianations are associated with distortions in perceptions that, in principle, lead to patently fantastical or unrealistic visions. They denote something that does not exist. The new, additonal definition recently introduced by the Cambridge Dictionary associates hallucination with artificial

[7] For a detailed technical explanation why hallucinations are technically inevitable given that the task of language generation is based on word prediction, see: Adam Tauman, Kalai Santosh S. Vempala, Calibrated Language Models Must Hallucinate (Dec. 5, 2023)

[8] Dagmara Panas et al., Can Large Language Models put 2 and 2 together? Probing for Entailed Arithmetical Relationships (??)

[9] Murray Shanahan, 'Talking about large language models' (2020) arXiv preprint arXiv:2212.03551

[10] 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. at 3, 7.

[11] Nouha Dziri et al., 2022. On the Origin of Hallucinations in Conversational Models: Is it the Datasets or the Models? arXiv. https: //doi.org/10.48550/ARXIV.2204.07931 Version Number: 1; Lei Huang et al., 'A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions' (2023) arXiv: 2311.05232

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

[13] Zorik Gekhman et al., 'Does Fine-Tuning LLMs on New Knowledge Encourage Hallucinations?' (13 May 2024).

[14] Alex Mallen et al., *When Not to Trust Language Models: Investigating Effectiveness of Parametric and Non-Parametric Memories,* PROC. 61st ANNUAL MEETING ASSOC. FOR COMPUTATIONAL LINGUISTICS 9802–9822 (2023)

[15] Adam Bouyamourn, Why LLMs Hallucinate, And How To Get (Evidential) Closure: Perceptual, Intensional and Extensional Learning for Faithful Natural Language Generation (2023) ACL 3183.

[16] caveat

[17] R. Thomas McCoy et al., *Embers of Autoregression: Understanding Large Language Models Through the Problem They are Trained to Solve*, ARXIV (Sep. 24, 2023), https://arxiv.org/pdf/2309.13638.pdf. 20

[18] R. Thomas McCoy et al., 23, 24

[19] Cambridge Dictionary

intelligence that **"**produces false information."[20] Similarly, the Merriam-Webster dictionary defines hallucinations as "a plausible but false or misleading response generated by an artificial intelligence algorithm."[21] At present, despite the sensory connotations of the term, "hallucinations" are associated with false, incorrect, or nonsensical statements generated by LLMs.[22] Other formulations refer to "false information not supported by the input,"[23] "exceedingly confident, yet erroneous, assertions,"[24] "plausible-sounding but unfaithful or nonsensical information."[25] It has also been suggested to replace the term hallucination with "soft bullshit" given that LLMs generate their outputs "without concern for its truth" and "without any intent to mislead the audience about the utterer's attitude towards truth."[26] Leaving aside the aforementioned suggestion that may be unbecoming for corporate clients and media campaigns, it is necessary to acknowledge – and maybe to question - the big conceptual jump from "something that does not exist" to false or nonsensical information, from something that originates in the sensory input of humans to something that denotes the erroneous or misleading output of a machine learning model. One might ask: why is this a problem? On a conceptual level, the introduction of this term seems to be a cheap marketing ploy aimed to mask the shortcomings of LLMs while implying that they possess human-like capabilities. To state that a system hallucinates sounds better than acknowledging that the system produces incorrect output[27] and plays well with the popular narrative concerning the imminent arrival of AGI. On a practical level, it masks the dangers of integrating LLMs into workflows, particularly in high-risk areas like law, finance or medicine. In principle, hallucinations in the traditional sense of the term, are easy to detect. It is usually evident when someone hallucinates or that a particular statement (including the lyrics of certain songs!) might be the result of an altered cognitive state. In the context of LLMs, however, the "hallucinated text" may still seem plausible, relevant, and informative.[28] In the context of LLMs, it may not be immediately apparent that the model is producing output that is detached from reality. As described below, what constitutes a nonsensical answer to a legal question may be far more difficult to determine than commonly assumed.

In sum, the continued use of the term "hallucination" is undesirable not only because of its anthropomorphic undertones but mainly because it misrepresents the difficulty of their detection. Average users who do not know the technical underpinnings and hence the risks of using LLMs, may assume that in the context of LLMs hallucinations are immediately apparent and, just like their "traditional equivalents," easy to detect.

4. Classifications

Despite the lack of a precise definition, there are continued efforts to introduce basic classifications of hallucinations. These efforts are particularly relevant in the context of legal hallucinations as they illustrate the fundamental challenges of establishing a clear point of reference, "something" that the generated output can be compared with or evaluated against. At present, it is common to distinguish between closed-domain and open-domain hallucinations. Closed-domain hallucinations are usually associated with output that deviates from or conflicts with the text contained in the prompt. Classic examples are summarization, translation, or

[20] Cambridge Dictionary has announced hallucinate as the Word of the Year for 2023, see: https://www.cambridge.org/news-and-insights/hallucinate-is-cambridge-word-of-the-year-2023

[21] Dictionary.com also picked "hallucinate" as its word of the year in 2023. The definition, when it comes to AI, means: "to produce false information contrary to the intent of the user and present it as if true and factual." Aliza Chasan, *Why dictionary.com's word of the year is "hallucinate,"* CBS NEWS (December 12, 2023), https://www.cbsnews.com/news/dictionary-com-word-of-the-year-hallucinate-ai/

[22] For a study of the definitions and common understanding of hallucinations in technical literature, see: Pranav Narayanan Venkit et al., "Confidently Nonsensical?": A Critical Survey on the Perspectives and Challenges of 'Hallucinations' in NLP (11 April, 2024)

[23] Yijun Xiao and William Yang Wang. 2021a. On hallucination and predictive uncertainty in conditional language generation. *arXiv preprint arXiv:2103.15025*.

[24] Jiaxin Zhang et al., '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. Association for Computational Linguistics.

[25] Ziwei Ji et al., 'Towards mitigating llm hallucination via self reflection,' in: *The 2023 Conference on Empirical Methods in Natural Language Processing.*

[26] M. T. Hicks et al., 'ChatGPT is Bullshit' (2024) 26 Ethics and Information Technology 38 .

[27] JOY BUOLAMWINI, UNMASKING AI (2023) 56

[28] Ji Ziwei et al., *Survey of Hallucination in Natural Language Generation*, 55 ACM COMPUTING SURVEYS 284, 1-38, 4, 5 (2023).

transcription tasks, where the generated output is inconsistent with the text provided by users.[29] It follows that LLMs can be said to hallucinate when they generate an incorrect or incomplete summarization or translation of a legal documents provided in the prompt.[30] Such discrepancies are particularly problematic in tasks like summarizing judicial opinions or extracting key points from a brief.[31] At first glance, one might assume that closed-domain hallucinations are relatively easy to detect because the generated output can be evaluated with reference to the text in the prompt – and the contents of the prompt are *by definition* provided by and hence known to the persons using the LLM. Arguably then, establishing the existence of a closed-domain hallucination requires a comparison between the text in the prompt and the generated text. While it cannot be denied that such comparison is likely to reveal hallucinations, it must be appreciated that comparing user input with model output can be extremely resource intensive as it would require a thorough verification whether the generated output constitutes, for example, a *correct* summarisation or translation of the input text. The fact that the text of the prompt is provided by the user and hence known does not imply that the LLM will efficiently process all such text and will do so correctly! After all, LLMs are known to ignore parts of longer prompts[32] and, even when provided with clear task instructions in the prompt, to remain exceedingly affected by the contents of their training corpora.[33] In other words, despite a clear point of reference, establishing the existence of closed-domain hallucinations may require a careful, resource-intensive comparison that may exceed the competence of the average user. The second category, open-domain hallucinations, involves outputs that that contradict or do not derive from the LLM's training corpus. [34] Some confusion seems to have been inadvertently introduced by Agarwal et al, who approached correctness as a comparison with ground-truth answers and distinguished it from groundedness, which required a comparison with the training data.[35] Regrettably, Agarwal et al defined open-domain hallucinations as fabricated text with little or no grounding in the training corpora. This reasoning is unconvincing if only due to the fact that given the lack of access to such corpora, it becomes impossible to confidently establish whether a certain output constitutes an open-domain hallucination. Along the same lines, the very concept of groundedness seems redundant. What is the point of having a point of reference – in this case: the training corpus – if its *contents* are difficult if not impossible to establish?
Other categorizations refer to output that conflicts with previously generated information, contradicts established facts[36] or is logically inconsistent.[37] Hallucination have also been classified either phenomenally or mechanically. The former approach classifies hallucination based on outcomes, while the latter focuses on training and deployment methodologies.[38]

The majority of the above classifications focuses on the consistency of the generated output with either the prompt or the training corpus. They are, however, neutral as to the domain in which the LLM is applied. For present purposes, these classifications are less helpful because, apart from lack of access to the training corpora and the LLM's tendency to ignore parts of the prompt, the main problem in the legal domain is that may be difficult to determine *what* the generated output should be compared against. Moreover, it seems more important to establish whether the generated output is incorrect or inconsistent with established legal doctrine than to determine whether it is inconsistent with the training corpus or the prompt. In this sense, the division into open- and closed-domain hallucinations may be of little practical relevance in the legal area – at

[29] Sebastien Bubeck et al., *Sparks of Artificial General Intelligence: Early experiments with GPT-4,* ARXIV (Apr. 13, 2023), https://arxiv.org/pdf/2303.12712.pdf. 82; Lei Huang et al., *A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions* (2023).

[30] Aniket Deroy, Kripabandhu Ghosh & Saptarshi Ghosh, *How Ready are Pre-trained Abstractive Models and LLMs for Legal Case Judgement Summarization?* LEGALAIIA (2023); Jaromir Savelka et al., *Explaining Legal Concepts with Augmented Large Language Models (GPT-4)*, ARXIV (Jun. 22, 2023) https://arxiv.org/pdf/2306.09525.pdf.

[32] Nelson F. Liu et al., *Lost in the Middle: How Language Models Use Long Contexts* (July 2023) arxiv.

[33] Zhaofeng Wu, *Reasoning or Reciting? Exploring the Capabilities and Limitations of Language Models Through Counterfactual Tasks* (28 March, 2024) https://arxiv.org/abs/2307.02477

[34] Ayush Agrawal et al., *Do Language Models Know When They're Hallucinating References?*, ARXIV (May 29, 2023), https://arxiv.org/abs/2305.18248; Adam Tauman, Kalai Santosh S. Vempala, *Calibrated Language Models Must Hallucinate* (Dec. 5, 2023).

[35] Ayush Agrawal et al., *Do Language Models Know When They're Hallucinating References?*, ARXIV (May 29, 2023), https://arxiv.org/abs/2305.18248

[36] 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.

[37] Lei Huang et al., *A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions* (2023) arXiv: 2311.05232.

[38] for an alternative, particularly nuanced classification see: Vipula Rawte et al., *The Troubling Emergence of Hallucination in Large Language Models – An Extensive Definition, Quantification, and Prescriptive Remediations* (2023) arXiv: 2310.04988.

least when compared to the problems ahead! The limited usefullness of this division is particularly visible in one of the first studies of hallucinations in the legal domain, which focused on *open-domain* hallucinations. The latter were defined as output that contradicts or does not directly derive from the LLM's training corpus.[39] To complicate matters, Dahl et al. also stated that such output "should be logically derivable from the content of its training corpus, regardless of whether the content of the corpus is factually or objectively true."[40] This statement acknowledges that training corpora often contain low-quality and incorrect information and, inadvertently, that consistency with such corpora may be a source of hallucinations in itself! At the same time, however, Dahl et al. also associated legal hallucinations with "factual infidelity between an LLM's response and the controlling legal landscape."[41] This approach is unsatisfactory for two reasons.

First, it cannot be assumed that the training corpus contains all the relevant legal information that could form the "legal landscape" in a given jurisdiction and, more importantly, that even if such information was contained therein it may not be represented in a manner that is statistically relevant. In other words, the mere presence of all the relevant legal information in the training corpus (assuming for the sake of argument that all such information could in fact be contained therein) does not guarantee that such information will form part of the LLMs parametric knowledge and affect the generated output. To be stored in the model's parameters, a word string representing specific information must not just be present but also frequent in the training corpus. It is widely known that, at inference time, LLMs are less likely to utilize low-probability word strings that they encountered during training.[42] It is also known that specialized, long-tail information is usually underrepresented in an LLM's parametric knowledge.[43] In sum, it is statistically unlikely that training corpora contain a sufficient amount of legal information that *could* constitute the legal landscape of a given jurisdiction.

Second, as training corpora are known to abound in incorrect or outdated information, an inconsistency with or deviation from such corpora need not result in a "factual infidelity" with the legal landscape. Legal hallucinations cannot be associated with infidelity to the training corpus *and* to the "legal landscape." There needs to be clarity as to *what* the output must conflict with to be considered a legal hallucination: the training corpus, which may contain information that is not objectively true and may not adequately represent the laws of a particular jurisdiction or with information that exists outside of such corpus but represents objective, undisputed facts. As Dahl et al. emphasized the importance of "factual and textual accuracy,"[44] their definition and undisciplined use of the term open-domain hallucination is unfortunate. This, however, is only the beginning of the difficulties.

Before proceeding, it is important to briefly revisit the concept of groundedness, which is often introduced as a variable in establishing the existence of hallucinations. A recent study by Magesh et al., who assessed the reliability of LLMs augmented with retrieval from auxiliary databases containing curated legal information, defined hallucinations as "outputs that are demonstrably false."[45] At the same time, the study enriched the concept by continuing the theme of groundedness. In their study, however, groundedness was evaluated not in relation to the training corpora but in relation to legal documents retreived from the aforementioned databases. According to Magesh et al, "a response is correct if it is both factually correct and relevant to the query. A response is incorrect if it contains any factually inaccurate information."[46] Correct repsonses were additionally evaluated with regard to their groundedness: "A response is grounded if the key factual propositions in its response make valid references to relevant legal documents. A response is ungrounded if key factual propositions are not cited. A response is

[39] Matthew Dahl et al.
[40] Matthew Dahl et al.
[41] Matthew Dahl et al.
[42] R. Thomas McCoy et al.,
[43] Nikhil Kandpal et al., '*Large language models struggle to learn long-tail knowledge'* in: Andreas Krause et al., Proceedings of the 40th International Conference on Machine Learning, volume 202 of Proceedings of Machine Learning Research, pages 15696–15707. PMLR, 23–29 Jul 2023.
[44] Matthew Dahl et al.
[45] Varun Magesh et al., *Hallucination-Free? Assessing the Reliability of Leading AI Legal Research Tools* (May 30, 2024) https://arxiv.org/abs/2405.20362.
[46] Magesh et al. above.

misgrounded if they are cited but misinterpret the source or reference an inapplicable source."[47] This approach evaluated the quality of the retrieval system and the ability of a given language model to efficiently utilise the additional information. Acknowledging the relevance and correctness of the retrieved legal documents, hallucinations were defined as follows: "a response is considered hallucinated if it is either incorrect or misgrounded. If a model makes a false statement or falsely asserts that a source supports a statement, that constitutes a hallucination."[48] Although its general usability within the confines of the study cannot be discarded, the more elaborate definition of halucinations proposed by Magesh et al must be approached with some caution if the aim is to delineate the concept of hallucinations for the needs of the legal profession, as well as for the purposes of this paper. In principle, the concept of groundedness is useful only to the extent that the source texts to which the generated output is compared to – in which it is *grounded* - are not only known and available but also correct. If those source text are not available, the situation is comparable to open domain hallucinations which required a comparison of the generated output with the training corpora. If the source text is not curated and cannot guaranteed to be correct, then grounding the generated output in such text may result in hallucinations. Most importantly, whether the generated output is grounded or not may not be as important if a particular task requires complex reasoning with regards to the source text.

In sum, the proposed division into open- and closed-domain hallucinations as well as the concept of groundedness are useful in general but, unless used in carefully delineated scenarios measuring a particular skill, likely to introduce additional confusion. In high-risk areas such as law, it is less important whether the generated output is grounded in the prompt, in the training data or in an auxiliary database, then whether it is objectively correct or, as described below, legally feasible. Fidelity to the law seems more important than fidelity to the training corpus or to the prompt. Training corpora abound in dubious sources and incorrect information, prompts often contain incorrect instructions or 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, in the legal context it may often be more advantageous that the LLM disregarded its parametric knowledge and/or the prompt.

5. Absence of Ground Truth

Leaving aside classifications of hallucinations based on a reference text, the more pertinent question in the legal domain is: 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. To recall, the main focus of this paper is to demonstrate the need for a domain-specific approach to "legal hallucinations." Existing technical literature addresses the problem of hallucinations in scenarios where the generated statements 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 and legal services, traditional technical approaches are difficult to apply and may lead to an unintended obfuscation of the risks of using LLMs.

An aside: it is difficult to confidently state whether we should be talking about the challenges of *detecting* hallucinations or of the challenges of defining the concept in the first place. After all, before "something" can be detected, there must be clarity as to what this "something" is. These challenges 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

[47] Magesh et al. above.
[48] Magesh et al. above.

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.

Situation 1

In the first situation, it is possible to provide a single answer and to evaluate such answer as either correct or not. The question has only one correct answer. There is an undisputable ground truth. For example, when inquiring about the existence of a case, its main ruling or the name of the judge who delivered the opinion of the court, there is only one possible correct answer. These questions concern objective facts and are easy to verify. The existence and contents of legal sources, such as cases, statutes and academic treatises are facts. They also constitute a ground truth, against which the output can be evaluated as either correct or not.[49] If the generated output misrepresents or contradicts the contents (or very existence) of a legal source, the LLM hallucinates. To clarify, the present situation does not require an interpretation or a reconciliation of legal sources. It focuses on the existence and contents of a legal source. It must be acknowledged that knowledge of the law – or legal expertise - is not necessarily synonymous with knowledge concerning the existence and contents of legal sources. Being able to reliably answer questions about legal sources is not synonymous with being able to solve legal problems. There is a difference between stating the law and applying the law.

The aforementioned study of legal hallucinations, examined whether LLMs could generate accurate information concerning legal sources.[50] Legal hallucinations were equated with output inconsistent with legal facts, such as statutes and cases.[51] Dahl et al. emphasized the need to ensure "adherence to the source text" as "unfaithful or imprecise interpretations of law can lead to nonsensical—or worse, harmful, and inaccurate—legal advice or decisions.[52] The study itself focused on outputs that constituted answers to open-ended legal queries. These "open-ended queries" did not, however, require any interpretation of the law, reconciliation of legal sources or legal reasoning but involved verifiable questions about the existence and contents of federal court cases. The generated outputs were evaluated against an undisputable, objective legal ground truth. As US cases follow a standard structure, Dahl et al. relied on "tabular metadata that is recorded in legal databases on these dimensions to create knowledge queries that simulate basic legal research tasks for each case."[53] The study involved fourteen tasks resembling basic legal research of various complexity. All tasks involved a ground truth and were objectively measurable. In all tasks, the LLMs had to rely exclusively on their parametric knowledge. Low complexity tasks did not require higher-order legal reasoning but, for example, the determination whether a case existed or the provision of the name of the court that ruled on it after being provided with case citation. Moderate complexity tasks required knowledge of a case's substantive content, which had to be obtained from specific portions of its text. For example, given a case citation, the model had to supply a case that had been cited in the opinion or indicate the year it was overruled. High complexity tasks presupposed rudimentary legal reasoning skills *and* information that was not readily available in existing legal databases like LexisNexis. These tasks required an LLM to synthesize legal information "out of unstructured legal prose."[54] For example, provided with a case citation, the LLM had to state its subsequent procedural history or its central holding.[55] It was established that hallucinations increased with the complexity of the task and that LLMs were, in principle, unable to assess relationships between cases, a skill indispensable in basic legal research. The number of hallucinations also varied by court as well as by jurisdiction. LLMs were more familiar prominent precedents than with cases from smaller courts. Prominent precedents at the federal level are, by definition, mentioned more frequently than cases from smaller courts. The study was confined to "simple" hallucinations that concerned objectively verifiable facts. It

[49] Zhang et al., *supra* note **Error! Bookmark not defined.**, at 3.
[50] Dahl et al., *supra* note 37, at 1.
[51] Dahl et al., *supra* note 37, at 1.
[52] Dahl et al., *supra* note 37, at
[53] Dahl et al., *supra* note 37, at 2
[54] Dahl et al., *supra* note **Error! Bookmark not defined.**, at
[55] Dahl et al., *supra* note **Error! Bookmark not defined.**, at 4,5,6.

did not engage with more difficult legal questions but, indirectly, confirmed that LLMs are not capable of basic legal reasoning that would be expected from a junior lawyer or even a paralegal engaged to locate relevant legal sources.

Situation 2

This situation involves questions that can have more than one correct answer or answers that cannot be evaluated with reference to a single 'legal ground truth.'[56] Situation 2 requires knowledge of the law and legal reasoning, including the interpretation and reconciliation of legal rules.[57] 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 text of a statute or a case is a question of fact. A legal source "says what it says." It has a certain date of publication and entry into force; it has a certain content, and it is "produced" by an identifiable person or body. These facts are ground truths. In contrast, the meaning and practical implications of a legal source can, however, be amendable to multiple interpretations. Even codified legal rules, such as those found in statutory instruments and regulations, are intrinsically open-textured, dynamic, uncertain and incomplete.[58] 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. Many legal principles are intrinsically open-textured, dynamic, and capable of multiple interpretations.[59] They can be uncertain, incomplete, and, to complicate matters, capable of different formulations. It follows that the principles relevant to a particular legal question may be scattered across multiple legal sources that are often inconsistent and contradictory. As each law student will confirm, even finding and reconciling the relevant legal sources is a skill acquired over multiple years. Once the foregoing observations about the nature of legal knowledge are combined with the fact that experienced lawyers often adopt different approaches to a particular legal problem and may differ in their interpretation of an identical set of facts, it quickly becomes apparent that any reference to a ground truth is an exercise in futility. There simply may not be a single point of reference, a perfect answer, when evaluating an output that was generated by an LLM in response to a legal question involving the application of the law. Legal problems are not algorithmic and capable of a straightforward answer. According to Dahan et al., this may be the reason why a common answer provided by lawyers is 'it depends.'[60]

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 subjective and resource intensive as it would require a panel of domain experts agreeing which answers are, for example, closest to established legal doctrine and thus most likely to succeed in court. The very need for such resource intensive and subjective evaluations 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

[56] Ronald Dworkin, *No Right Answer?*, 53 N.Y.U. L. REV. 1 (1978), republished in RONALD DWORKIN, A MATTER OF PRINCIPLE 119–45 (1985) (observing that most legal cases have no single correct answer); D Litowitz, *Dworkin and Critical Legal Studies on Right Answers and Conceptual Holism* (1994) 18(2) LEGAL STUDIES FORUM 135; KARL LLEWELLYN, JURISPRUDENCE: REALISM IN THEORY AND PRACTICE (2011).

[57] I deliberately use the broad term 'legal rule' to encompass legal principles, concepts, and policies, irrespective of their source, *see*: R DWORKIN, TAKING RIGHTS SERIOUSLY 22 (1977).

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

[59] W TWINNING & DAVID MIERS, HOW TO DO THINGS WITH RULES 122, 137 (2010); Brian Bix, *H.L.A. Hart and the "Open Texture" of Language*, 10 LAW & PHIL. 51, 52–55 (1991), republished in BRIAN BIX, LAW, LANGUAGE, AND LEGAL DETERMINACY ch. 1 (1995).

[60] Samuel Dahan et al., *Lawyers Should Not Trust AI: A call for an Open-source Legal Language Model* (August 28, 2023). Queen's University Legal Research Paper, Available at SSRN: https://ssrn.com/abstract=4587092 or http://dx.doi.org/10.2139/ssrn.4587092

and rank its quality. At the same time, as "tasks that are harder to evaluate also tend to be those that would lead to the most significant changes in the legal profession."[61]

The output may seem correct but, upon closer examination, may turn out be legally indefensible. The output may also seem incorrect but turn out to be a viable albeit unusual legal approach to the legal task at hand. In both instances, the output requires extensive review and expert verification. Even then, however, legal opinions may differ as to its quality. Absent a single "legal ground truth," which would provide a clear point of reference against which one could compare to generated output, 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. Regardless of whether an answer to a legal question was provided by a human or generated by an LLM, its legal feasibility depends on the opinion *and expertise* of those who evaluate such answer. Just like different lawyers may provide different answers to the same legal question, different lawyers may differ in their evaluation of the answer – irrespective of its provenance.

6. Eye of the Beholder

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. Experienced lawyers, academics and adjudicators will detect more hallucinations – or *potential* hallucinations - than law students. The existence *(or detection?)* of hallucinations may be a function of the user's competence. According to Heersmink et al, "generating a false answer or response is known as hallucinating. Again, recognising this depends on the knowledge, skills, and attitudes of the user. What's obviously false for one person, may not be so for another. A major problem is that, if one isn't knowledgeable on the topic in question, it is impossible to detect when it is hallucinating."[62] It follows that experienced lawyers, academics and adjudicators will detect more hallucinations – or *potential* hallucinations - than law students or non-lawyers. The latter are less likely to appreciate or even notice inconsistencies or "deviations" from what could be considered a legally feasible answer. In other words, non-lawyers and inexperienced lawyers are less likely to detect outputs that warrant further scrutiny and verification.[63] The problem is not necessarily that a legal question can have multiple correct answers but in that the person who posed the question may not appreciate its complexity and the resulting need to verify the generated answer.

Maybe the overwhelming enthusiasm surrounding LLMs derives from the fact that those who use LLMs are often incapable of evaluating the quality of their output? Maybe they too easily conflate linguistic competence with actual expertise in a given domain? The fact that the generated text "looks good" does not imply that it is "good." Regrettably, it has been established that humans often prefer generated responses even though such responses are, in a majority of cases, incorrect.[64] Furthermore, text generated by LLMs is often perceived as "clearer and more engaging" without "identifying any differences with regards to message's competence and trustworthiness."[65]

7. Explainability

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

[61] Sayash Kapoor, Peter Henderson, Arvind Narayanan, *Promises and pitfalls of artificial intelligence for legal application*s, ARXIV (Jan. 10, 2024) https://arxiv.org/pdf/2402.01656.pdf.

[62] Richard Heersmink et al., *A phenomenology and epistemology of large language models: transparency, trust, and trustworthiness* (2024) 26 Ethics and Information Technology 41; see also: Noam Kolt, *Predicting Consumer Contracts*, 37 BERKELEY TECH. L. J. 71 (2021) at 95.

[63] Boxi Cao et al., *Knowledgeable or educated guess? revisiting language models as knowledge bases*, PROC. 59TH ANN. MEETING ASS'N COMPUTATIONAL LINGUISTICS 1860 (2021); Yanai Elazar et al., *Measuring and improving consistency in pretrained language models*, 9 TRANS. ASSOC'N. COMPUTATIONAL LINGUISTICS 1012 (2021).

[64] 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.

[65] For a study of credibility perceptions of human-generated and computer-generated content in different user interface (UI) settings, see: Martin Huschens et al., *Do You Trust ChatGPT? – Perceived Credibility of Human and AI-Generated Content* (5 Sep 2023) arXiv:2309.02524v1 [cs.HC] 5 Sep 2023

understanding *how* it was arrived at.[66] After all, "depending on one's view of law, there may not be a single right answer. In which case, what matters is not getting to the right decision in any way, but getting to any decision in the right way."[67] The *potential* legal viability of an output requires explainability. To clarify: explainability is 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.[68] When establishing whether the generated output constitutes a hallucination, explainability concerns the *reasoning* underlying the output as well as the sources relied on. It does not address the question *how* the LLM works in technical terms but aims to establish *why* it generated a particular answer to a legal question. Prompted with the question "why did you generate the sentence, 'the judge will rule for the appelant?'" an LLM might respond: "after processing the convolutional layers, the activation for the *appelant* in the softmax layer was higher then for the *respondent*." Such explanation is less useful than one stating "given existing case law, especially cases [a, b, c, d] and a review of evolving judicial attitudes in similar cases, it is more likely that the court would rule for the appellant than for the respondent.

Unfortunately, LLMs are inherently incapable of explaining the reasoning underlying their outputs. They can be prompted to provide an explanation, but such "explanation" will also be the result of a language generation task – not a retrospective examination/exploration of the model's reasoning. Consequently, the explanation will have to be evaluated as to its substantive consistency with the output in the sense of providing "a plausible causal account of how y was derived from x and c."[69] It has been demonstrated that LLMs can generate plausible and superficially convincing explanations of their output - even where the output is incorrect or nonsensical.[70] After all, LLMs aim to maintain consistency with previously generated output, including earlier hallucinations, rather than correcting previously generated output.[71] Even in chain-of-thought prompting, a technique instructing models to provide reasoning steps underlying their output, LLMs can generate "fake" reasoning steps misrepresenting the "true reason" for their response.[72]

8. Interim Conclusions

Intuitively, one might suggest that the only way of counterbalancing the risks of hallucinations is to rigorously review the generated output. Given the length of such output as well as the inherent difficulty of establishing whether a particular legal proposition or solution are feasible or whether they constitute a hallucination, the idea of reviewing the output – however thoroughly - does not seem to address the fundamental problem: that of reliably detecting the existence of hallucinations. Moreover, a rigorous review would by definition be extremely resource intensive and defeat the very purpose of deploying LLMs - that of increasing efficiency in the performance of legal tasks. 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.

[66] Luciano Floridi, Massimo Chiriatti, *GPT-3: Its Nature, Scope, Limits, and Consequences*, 30 MINDS & MACH. 681, 687 (2020).
[67] Reuben Binns, *Analogies and Disanalogies between Machine-Driven and Human-Driven Legal Judgement*, 1 J. CROSS-DISCIPLINARY RES. COMPUTATIONAL L. 1, (2021).
[68] 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.
[69] Bubeck et al., note X, at
[70] Muru Zhang et al., *How Language Model Hallucinations Can Snowball,* ARXIV (May. 22, 2023) https://arxiv.org/abs/2305.13534
[71] *Id.*
[72] Miles Turpin, Julian Michael, Ethan Perez, Samuel R. Bowman, *Language Models Don't Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting,* ARXIV (Dec. 9, 2023) https://arxiv.org/pdf/2305.04388.pdf.