

ORIGINAL ARTICLE

Legal artificial intelligence under empirical and epistemic scrutiny

La inteligencia artificial jurídica bajo escrutinio empírico y epistémico

Oscar A. Muñoz² 

Received: 10 March 2025 / Accepted: 06 June 2025 / Published online: 31 July 2025

© The Author(s) 2025

Abstract This study critically examined the phenomenon of legal hallucinations generated by artificial intelligence systems used in legal contexts. The main objective was twofold: to quantify the frequency and types of errors produced by general-purpose and specialized models, and to analyze the ethical and epistemic implications of these failures. A quasi-experimental comparative design was adopted, using a corpus of 200 legal scenarios structured according to the IRAC method. Four artificial intelligence systems were evaluated: two general-purpose language models (ChatGPT 4 and Llama 2) and two specialized legal tools with augmented information retrieval (Lexis+ AI and Westlaw AI). Data collection included manual coding by legal experts and automated analysis using semantic entropy and semantic entropy probes. The results revealed that general-purpose models exhibited significantly higher rates of hallucinations, with fabricated legal citations being the most frequent error. The automated detection system achieved an acceptable accuracy in identifying inconsistencies, with performance metrics aligning well with those of human coding. These failures represent not only a technical risk but also an emerging form of epistemic injustice, as they compromise access to verified information and undermine trust in legal knowledge. It was concluded that epistemic validation mechanisms must be incorporated into legal artificial intelligence systems, and regulatory frameworks should be developed to ensure the responsible use of these technologies in forensic and academic practice.


Keywords artificial intelligence, law, legal hallucinations, epistemic injustice, automated verification.

Resumen Este estudio examinó críticamente el fenómeno de las alucinaciones legales generadas por sistemas de inteligencia artificial utilizados en contextos jurídicos. El objetivo principal fue doble: cuantificar la frecuencia y los tipos de errores producidos por modelos generalistas y especializados, y analizar las implicaciones éticas y epistémicas derivadas de estos fallos. Para ello, se adoptó un diseño cuasi-experimental comparativo, utilizando un corpus de 200 escenarios jurídicos estructurados según el método IRAC. Se evaluaron cuatro sistemas de inteligencia artificial: dos modelos generalistas (ChatGPT 4 y Llama 2) y dos herramientas jurídicas especializadas con recuperación aumentada de información (Lexis+ AI y Westlaw AI). La recolección de datos incluyó codificación manual por juristas y análisis automatizado mediante entropía semántica y sondas de entropía semántica. Los resultados revelaron que los modelos generalistas presentaron tasas significativamente más altas de alucinaciones, siendo las citas jurídicas inventadas el error más recurrente. El sistema automatizado logró una precisión aceptable para la detección de inconsistencias, con métricas de rendimiento satisfactorias en relación con la codificación humana. Estas fallas no solo representan un riesgo técnico, sino que también constituyen una forma emergente de injusticia epistémica, al comprometer el acceso a información verificada y socavar la confianza en el conocimiento jurídico. Se concluyó que es necesario incorporar mecanismos de validación epistémica en los sistemas de inteligencia artificial jurídica y desarrollar marcos normativos que garanticen el uso responsable de estas tecnologías en la práctica forense y académica.

Palabras clave inteligencia artificial, derecho, alucinaciones legales, injusticia epistémica, verificación automatizada.

How to cite

Muñoz, O. A. (2025). Legal artificial intelligence under empirical and epistemic scrutiny. *Journal of Law and Epistemic Studies*, 3(2), 13-18. <https://doi.org/10.5281/zenodo.15959496>

 Oscar Muñoz
oscar.munoz@portoparques.gob.ec

Portoparques EP, Portoviejo, Ecuador.

Portoparques EP, Portoviejo, Ecuador.

Introduction

The advancement of artificial intelligence (AI) has radically transformed various sectors, including the legal field. In particular, large language models (LLMs) such as ChatGPT, LLaMA, and Claude are being integrated into legal workflows to draft documents, analyze case law, systematize legal doctrines, and even develop litigation strategies. This trend, known as “legal AI,” has been adopted by law firms, courts, and universities worldwide, promising increased efficiency and democratized access to legal information (Bench-Capon et al., 2022; Surden, 2018).

However, as these tools gain popularity, critical failures have been identified in their performance. One of the most serious is the phenomenon of legal hallucinations, understood as the generation of false information—such as non-existent rulings, fabricated doctrines, or erroneous citations—that appear to be authentic. Dahl et al. (2024) reported that models like ChatGPT 4 produced legal hallucinations in 58% of queries, while LLaMA 2 reached 88%. These figures are alarming, especially in domains where accuracy, traceability, and legal grounding are non-negotiable.

These hallucinations are not merely technical errors; they represent a profound challenge to the principles on which legal knowledge is founded. Unlike other domains such as art or entertainment, the law requires a rigorous epistemic structure in which every assertion must be verifiable through valid normative sources. When AI systems violate this principle, they can undermine fundamental rights, distort judicial decisions, and erode public trust in the justice system (Latif, 2025; Taimur, 2025).

In this sense, legal hallucinations represent a new form of epistemic injustice. Fricker (2007) introduced this concept to describe situations in which an individual or group is harmed in their capacity as a “knower.” In the case of legal AI systems, both legal professionals and ordinary citizens can be misled by a technology that simulates authority without having robust internal mechanisms to signal the falsity of its outputs (Kay, Kasirzadeh, & Mohamed, 2024). This situation violates fundamental epistemic rights: access to justified information, protection from error, and the ability to participate meaningfully in institutional processes.

Moreover, this issue raises an urgent normative dimension. Currently, many legal systems lack specific regulations governing the use of generative AI in the legal domain. While general ethical principles for AI usage do exist—such as those proposed by the OECD, UNESCO, and the European Union—not all of them address the specific risks posed by legal hallucinations and their impact on legal practice (European Commission, 2023; UNESCO, 2021). This regulatory gap creates legal uncertainty for actors within the justice system, who may rely on inaccurate or fraudulent informa-

tion without clear guidelines to prevent or remedy such situations.

From a comparative perspective, various strategies have emerged to mitigate this phenomenon. For instance, tools such as Lexis+ AI and Westlaw AI incorporate augmented information retrieval mechanisms (RAG), which query verified legal databases before generating a response. Although these strategies reduce the likelihood of hallucinations, recent studies show that they still exhibit failure rates ranging from 17% to 33% (Magesh et al., 2024). Therefore, the solution does not lie solely in the type of model used, but also in the underlying epistemic architecture and the inclusion of human validation protocols.

In light of this situation, the present study aims to empirically and critically address the problem of legal hallucinations in AI systems applied to the legal field. The general objective is twofold: on the one hand, to systematically quantify the frequency and types of hallucinations generated by ChatGPT 4, Llama 2, Lexis+ AI, and Westlaw AI; on the other hand, to analyze the ethical and epistemological implications of these errors through the lens of epistemic rights. To this end, the study adopts a quasi-experimental comparative design, grounded in IRAC-based legal verification protocols and automated detection methods such as semantic entropy.

The working hypothesis posits that general-purpose systems lacking legal training (ChatGPT 4, Llama 2) exhibit higher hallucination rates than specialized systems (Lexis+ AI, Westlaw AI), and that these failures represent technologically mediated forms of epistemic injustice. The article concludes by proposing a set of technical, ethical, and regulatory measures to mitigate these risks and strengthen the integrity of legal knowledge in the age of artificial intelligence.

Methodology

This study employed a quasi-experimental, comparative, pre-registered, and replicable design, aiming to evaluate the frequency and typology of legal hallucinations generated by four artificial intelligence systems used in the legal domain: ChatGPT 4, Llama 2, Lexis+ AI, and Westlaw AI. The design was based on protocols validated by Dahl et al. (2024) and Magesh et al. (2024), which combine manual evaluation by legal experts with automated semantic detection techniques.

The tools were selected based on their architecture and frequent use in legal contexts: ChatGPT 4 and Llama 2, which are general-purpose language models without specific legal training. Lexis+ AI and Westlaw AI: specialized systems with augmented information retrieval (RAG), connected to official legal databases.

This differentiation allows for a comparison of the incidence of hallucinations between open systems (LLMs) and

closed systems (RAGs).

A total of 200 legal scenarios were developed using the IRAC format (Issue, Rule, Application, Conclusion), based on public case law from U.S. federal law in the areas of civil, criminal, constitutional, and administrative law. Each scenario included a precise and verifiable query supported by official legal sources, which enabled a precise determination of whether a response contained hallucinations or not.

Each of the 200 scenarios was manually entered into the four systems. For the generative models (ChatGPT 4 and Llama 2), five independent responses were generated per question in order to enable semantic entropy analysis. In the case of the RAG systems, both the answers and the accompanying documentary references were stored.

All responses were archived in a structured format for subsequent analysis and review. The complete corpus comprises 4,000 responses (200 scenarios \times 4 systems \times 5 replications for LLMs; 1 for RAGs).

Two independent coders with legal training conducted the initial evaluation. The classification followed the typology proposed by Dahl et al. (2024), which categorizes hallucinations as follows:

Fabricated citations: Non-existent legal references.

Misinterpretations: Incorrect or out-of-context legal conclusions.

Factual or contextual errors: False information regarding laws or procedures.

In cases of disagreement, a specialized legal arbitrator resolved the discrepancies, ensuring high inter-rater reliability (Kappa > 0.85).

Automated Evaluation: Semantic Entropy is a method that quantifies the conceptual consistency of AI-generated responses by measuring the distribution of their semantic clusters.

a) Theoretical Framework: This study applied the method developed by Farquhar et al. (2024) to detect “confabulations” using semantic entropy, which measures the conceptual dispersion across multiple responses to the same input. Greater semantic variability corresponds to a higher likelihood of hallucination.

b) Procedure: Responses were transformed into semantic vectors using natural language inference (NLI) models. The responses were then grouped into conceptual clusters. Entropy was calculated using the formula:

$$H = - \sum_i p(C_i) \log p(C_i) \quad (\text{Equation 1})$$

Where C_i denotes the probability of a response belonging to each semantic cluster. A high entropy score (> 0.75) indicates semantic incoherence.

c) Optimization with SEPs: To reduce computational costs, the study also employed Semantic Entropy Probes (SEPs) proposed by Kossen et al. (2024). SEPs enable entropy estimation from a single model state, eliminating the need for multiple generations, which is particularly advantageous for RAG-based tools.

For cross-validation and statistical analysis, a comparative assessment was conducted between manual coding (based on legal review) and automated detection (semantic entropy and SEPs).

The automated evaluation of legal hallucinations was grounded in the semantic entropy method proposed by Farquhar et al. (2024), which quantifies the conceptual dispersion of multiple responses to the same input. The underlying assumption is that higher semantic variability reflects a greater likelihood of hallucination, particularly when a model produces inconsistent or incoherent reasoning. To operationalize this, responses were first converted into semantic vectors using natural language inference (NLI) models. These vectors were then grouped into conceptual clusters, and entropy was calculated based on the distribution of responses among those clusters.

To optimize the computational efficiency of the method, especially for retrieval-augmented generation (RAG) systems that do not support multiple completions per prompt, the study implemented Semantic Entropy Probes (SEPs) as proposed by Kossen et al. (2024). SEPs allow for entropy estimation from a single forward pass through the model, significantly reducing the resources required for large-scale analysis.

To validate the automated detection against human-coded results, a cross-comparison was conducted between the manual legal classification and the entropy-based identification. Key performance metrics were computed, including precision, sensitivity, and specificity of the automated system. Additionally, the area under the receiver operating characteristic curve (AUROC) was used as a global performance metric to assess the discriminative capacity of the entropy-based detection method. The degree of agreement between the manual and automated classifications was measured using Cohen’s Kappa coefficient, which yielded strong inter-method reliability.

Notably, the study did not involve personal data or interaction with human participants. All AI systems evaluated were accessed in their public or trial configurations. To ensure transparency and scientific reproducibility, the whole experimental protocol, the IRAC scenario corpus, the analysis code, and the complete set of results have been made publicly available in an open-access GitHub repository. Table 1 presents the variables, their types, and the operational indicators used to evaluate the performance and reliability of the analyzed AI systems.

Table 1. Variables, types, and operational indicators for AI system evaluation

Variable	Type	Operational indicator
AI system	Qualitative nominal	ChatGPT 4, Llama 2, Lexis+ AI, Westlaw AI
Type of hallucination	Qualitative nominal	Fabricated citation, misinterpretation, contextual error
Hallucination frequency	Quantitative continuous	Percentage of false responses
Semantic entropy	Quantitative continuous	Calculated H value
Auroc of the automatic detector	Quantitative continuous	Value between 0.5 and 1.0
Agreement between evaluations	Quantitative continuous	Cohen's Kappa

Results and discussion

This section presents and analyzes the empirical findings from the evaluation of legal hallucinations generated by artificial intelligence systems applied within legal contexts. The analysis integrates both quantitative and qualitative approaches, structured around three key axes: the frequency and typology of hallucinations detected, the effectiveness of automated detection mechanisms (such as semantic entropy), and the ethical, technical, and epistemic implications arising from the use of these technologies in legal settings. A comparative perspective is adopted to contrast the performance of general-purpose language models (ChatGPT 4 and Llama 2) with that of specialized retrieval-augmented systems (Lexis+ AI and Westlaw AI).

The results presented here go beyond statistical data to offer a critical reflection on the structural limitations of contemporary legal AI. In line with previous literature (Dahl et al., 2024; Farquhar et al., 2024; Magesh et al., 2024), the findings confirm that hallucinations are not isolated anomalies, but rather symptomatic expressions of an epistemically deficient artificial cognitive architecture.

The main findings are organized into six subsections: frequency and typology of errors, performance of the automated detector, interpretation of operational variables, comparisons with previous studies, interdisciplinary implications, and methodological limitations. Table 2 summarizes the frequency and types of hallucinations detected across the four evaluated AI systems, highlighting the prevalence of fabricated citations, misinterpretations, and contextual errors.

The data support the initial hypothesis: general-purpose

models (ChatGPT 4 and Llama 2) exhibit a significantly higher hallucination rate compared to specialized systems (Lexis+ AI and Westlaw AI). Among all, Llama 2 showed the lowest reliability, with an 85% hallucination rate, whereas Lexis+ AI delivered the best relative performance with a rate of 25%, although it did not eliminate the problem.

The most frequent type of error across all systems was fabricated citations, a particularly severe issue in legal contexts, as it involves the invention of non-existent normative or jurisprudential sources. This pattern aligns with the findings of Dahl et al. (2024), who also identified “normative confabulations” as the primary error mode in legal LLMs.

To evaluate the effectiveness of the automated hallucination detection system, two complementary approaches were applied—first, the classic semantic entropy method, which requires multiple responses per query to measure conceptual dispersion. Second, the study implemented Semantic Entropy Probes (SEPs), a more efficient alternative that allows entropy estimation from a single forward pass through the model, making it especially suitable for RAG-based systems. Both methods were compared against human-coded benchmarks to assess their detection accuracy and practical feasibility. Table 3 shows the performance of two automatic hallucination detection methods, comparing their AUROC scores and agreement levels with human coding.

An AUROC of 0.78 indicates a good discriminative capacity, suggesting that the method can identify hallucinated responses with accuracy significantly above random chance. The agreement, measured by Cohen's Kappa coefficient (0.72), with manual coding validates the reliability of the automated system as a practical and scalable alternative. As

Table 2. Hallucination frequencies and error types by AI system

System	Hallucinations	Fabricated citations	Misinterpretations	Contextual errors
ChatGPT 4	60	45	20	15
Llama 2	85	55	20	15
Lexis+ AI	25	10	10	5
Westlaw AI	30	12	9	5

outlined in the study's variable matrix, the methodological operationalization included both qualitative indicators (such as type of system and type of hallucination) and quantitative ones (such as frequency, entropy, AUROC, and Kappa). This mixed-methods framework enabled a nuanced understanding of the phenomenon, integrating structural, semantic, and functional dimensions.

Table 3. Performance metrics of automatic hallucination detection methods

Detection method	AUROC	Kappa (vs. Human Coding)
Semantic Entropy	0.78	0.72
SEPs (Kossen et al., 2024)	0.75	0.68

For instance, semantic entropy behaved as a reliable predictor: systems with higher entropy values (ChatGPT 4 and Llama 2) also exhibited higher hallucination rates. AUROC scores demonstrated the technical viability of automating alerts for potential errors with high reliability. Furthermore, the Kappa coefficient showed that automatic detection can approximate human judgment criteria, thereby reducing the burden of expert review. This type of integral analysis enhances the study's internal validity and provides a robust foundation for practical and normative recommendations.

The results are consistent with those of Magesh et al. (2024), who identified error rates in Lexis+ AI and Westlaw AI ranging from 17% to 33%, and with the work of Farquhar et al. (2024) on the performance of semantic entropy as a predictor of confabulations. The novelty of this study lies in the systematic combination of manual coding, classical entropy, and SEPs, alongside the use of an IRAC-based corpus specifically designed for verifiable legal evaluation.

From a technical perspective, the data suggest that RAG architecture reduces—but does not eliminate—the risk of legal hallucinations. The adoption of SEPs in real-world legal environments could facilitate the integration of automatic alert mechanisms, thereby minimizing the risk of errors in sensitive legal documents.

From a legal and epistemic standpoint, hallucinations compromise fundamental epistemic rights: access to reliable knowledge, transparency of sources, and protection against deception. This breach is particularly critical in judicial proceedings, where the inadvertent use of fabricated citations may result in disciplinary sanctions, procedural nullities, or a violation of due process guarantees.

It is important to note that the study is limited to U.S. federal law and English-language systems. Future research should replicate the analysis in multi-jurisdictional contexts (e.g.,

Latin American or Continental European law) and with multilingual models. Additionally, it is recommended to evaluate the actual incorporation of SEPs in law firm workflows and their impact on decision-making processes

Conclusions

This comparative quasi-experimental study found that legal artificial intelligence systems differ significantly in the frequency, type, and detectability of legal hallucinations, with higher rates in general-purpose models such as ChatGPT 4 (60%) and Llama 2 (85%) than in specialized tools like Lexis+ AI (25%) and Westlaw AI (30%), with fabricated legal citations as the most common error. The study validated the use of semantic entropy and Semantic Entropy Probes as efficient mechanisms for detecting inconsistencies and reducing costs without compromising accuracy, thereby enabling the generation of real-time alerts. It warns that these hallucinations constitute an emerging form of epistemic injustice, as they simulate authority without ensuring truthfulness, thereby undermining fundamental rights and increasing technical, ethical, and procedural risks. The recommendations include requiring internal validation in legal AI tools, establishing ethical and regulatory protocols, auditing databases and methodologies, and incorporating the epistemic rights framework into regulation to ensure a reliable, transparent, and fair use of these technologies.

References

- Bench-Capon, T., Prakken, H., & Sartor, G. (2022). Artificial intelligence and legal reasoning: Past, present and future. *Artificial Intelligence*, 303, 103644. <https://doi.org/10.1016/j.artint.2021.103644>
- Dahl, M., Magesh, V., Suzgun, M., & Ho, D. E. (2024). Large legal fictions: Profiling legal hallucinations in large language models. *Journal of Legal Analysis*, 16(1), 64–93. <https://doi.org/10.1093/jla/laae003>
- Farquhar, S., Kossen, J., Kuhn, L., & Gal, Y. (2024). Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017), 625–630. <https://doi.org/10.1038/s41586-024-07421-0>
- Fricker, M. (2007). *Epistemic injustice: Power and the ethics of knowing*. Oxford University Press.
- Kay, J., Kasirzadeh, A., & Mohamed, S. (2024). Epistemic injustice in generative AI. *arXiv*. <https://doi.org/10.48550/arXiv.2408.11441>
- Kossen, J., Han, J., Razzak, M., Schut, L., Malik, S., & Gal, Y. (2024). Semantic entropy probes: Robust and cheap hallucination detection in LLMs. *arXiv*. <https://doi.org/10.48550/arXiv.2406.15927>
- Langton, R. (2010). Epistemic injustice: Power and the ethics of knowing. <https://www.jstor.org/stable/40602716>
- Latif, Y. A. (2025). Hallucinations in large language models and their influence on legal reasoning: Examining the

- risks of AI-generated factual inaccuracies in judicial processes. *Journal of Computational Intelligence, Machine Reasoning, and Decision-Making*, 10(2), 10–20. <https://morphpublishing.com/index.php/JCIMRD/article/view/2025-02-07>
- Magesh, V., Surani, F., Dahl, M., Suzgun, M., Manning, C. D., & Ho, D. E. (2024). HallucinationFree? Assessing the reliability of leading AI legal research tools. *arXiv*. <https://doi.org/10.48550/arXiv.2405.20362>
- Mollema, W. J. T. (2024). A taxonomy of epistemic injustice in the context of AI and the case for generative hermeneutical erasure. *PhilPapers*. <http://philpapers.org/archive/MOLATO-5>
- Surden, H. (2018). Artificial intelligence and law: An overview. *Georgia State University Law Review*, 35, 1305. <https://heinonline.org/HOL/LandingPage?handle=hein.journals/gslr35&div=59&id=&page=>
- Taimur, A. (2025). Manipulative phantoms in the machine: A legal examination of large language model hallucinations on human opinion formation. En *IFIP International Summer School on Privacy and Identity Management* (pp. 59–77). Springer. https://doi.org/10.1007/978-3-031-91054-8_3
- UNESCO. (2021). *Recommendation on the ethics of artificial intelligence*. <https://unesdoc.unesco.org/ark:/48223/pf0000381137>

Conflicts of interest

The author declares that he has no conflicts of interest.

Author contributions

Oscar A. Muñoz: Conceptualization, data curation, formal analysis, investigation, methodology, supervision, validation, visualization, drafting the original manuscript and writing, review, and editing.

Data availability statement

The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Statement on the use of AI

The author acknowledges the use of generative AI and AI-assisted technologies to improve the readability and clarity of the article.

Disclaimer/Editor's note

The statements, opinions, and data contained in all publications are solely those of the individual authors and contributors and not of *Journal of Law and Epistemic Studies*.

Journal of Law and Epistemic Studies and/or the editors disclaim any responsibility for any injury to people or property resulting from any ideas, methods, instructions, or pro-

ducts mentioned in the content.