International Journal of Legal Information 2025, pp. 1–12.
doi:10.1017/jli.2025.10052

Evaluating AI in Legal Operations: A Comparative Analysis of Accuracy, Completeness, and Hallucinations in ChatGPT-4, Copilot, DeepSeek, Lexis+ AI, and Llama 3

Bakht Munir*, Muhammad Zubair Abbasi†, W. Blake Wilson‡, and Allen Colombo Jr.#

Abstract

The proliferation of Artificial Intelligence (AI) is significantly transforming conventional legal practice. The integration of AI into legal services is still in its infancy and faces challenges such as privacy concerns, bias, and the risk of fabricated responses. This research evaluates the performance of the following AI tools: (1) ChatGPT-4, (2) Copilot, (3) DeepSeek, (4) Lexis+ AI, and (5) Llama 3. Based on their comparison, the research demonstrates that Lexis+ AI outperforms the other AI solutions. All these tools still encounter hallucinations, despite claims that utilizing the Retrieval-Augmented Generation (RAG) model has resolved this issue. The RAG system is not the driving force behind the results; it is one component of the AI architecture that influences but does not solely account for the problems associated with the AI tools. This research explores RAG architecture and its inherent complexities, offering viable solutions for improving the performance of AI-powered solutions.

Keywords: Hallucinations, LLMs, RAG, ChatGPT, Lexis +AI, DeepSeek, Copilot

Introduction

With the widespread adoption of AI in the legal profession, certain critical challenges, one of which is hallucinations—a phenomenon where AI models perpetuate plausible but inaccurate responses—came to the fore, necessitating verification of their generated content.[1] Advanced AI solutions have revolutionized the legal domain, including legal education, research, and practice, and have performed exceptionally on law school and bar exams as well as with legal analysis.[2] AI enables machines to simulate human intelligence, inspiring them to learn patterns

* The University of Kansas School of Law. Email: bakht.munir@ku.edu.
† Department of Law and Criminology, School of Law and Social Sciences, Royal Holloway, University of London. Email: Zubair.abbasi@rhul.ac.uk.
‡ The University of Kansas School of Law. Email: wmblakewilson@gmail.com.
# The University of Kansas School of Law. Email: allen.colombo@KU.edu.

We are deeply grateful to Prof. Richard E. Levy, Prof. Stephen J. Ware, Prof. Kyle Velte, Prof. Alexander I. Platt, and Prof. Laura Clark Fey of the KU Law School, along with Prof. Paul D. Callister of the UMKC School of Law, for their invaluable insights and thorough review of this manuscript. Their constructive feedback and suggestions have greatly enhanced the quality of our work. We sincerely appreciate their time, effort, and expertise, which have significantly contributed to the development of this research. Additionally, we are profoundly thankful to the KU School of Law for its unwavering support in helping us accomplish this work.

[1] M. Dahl, M.V. Magesh, M. Suzgun, and D.E. Ho. 2024. "Large Legal Fictions: Profiling Legal Hallucinations in Large Language Models." preprint, arXiv, arXiv:2401.01301.

[2] Choi, J.H., K.E. Hickman, A.B. Monahan, and D. Schwarcz. 2022. "ChatGPT Goes to Law School." *Journal of Legal Education* 71 (3): 387; Choi, J.H., and D. Schwarcz. 2023. "AI Assistance in Legal Analysis: An Empirical Study," https://

from the data they are trained on, solve problems, and make decisions. Lawyers are employing AI to augment legal services, and many of the world's leading firms have started using AI in rendering legal services.[3]

Nevertheless, these tools are not completely without risk and embrace ethical challenges, such as bias, data privacy, and fictitious outcomes, necessitating counter-verification of their generated content in the forms of citations, case law, statutes, quotations from superior court opinions, and legal arguments.[4] In response to users' prompts, AI generates new original content based on the machine-learning model that mirrors human intelligence and decision-making competence. With the development of Large Language Models (LLMs), these generative AI tools have gained more significance, creating human-like text based on the vast data on which these models are being trained. LLMs generate content by forecasting the next element in the sequence without any certainty of its accuracy, which could lead to generating misinformation.[5] Based on this probabilistic nature, in response to users' input, AI models may generate erroneous outcomes, which often sound plausible, and users assume they are accurate.[6]

ChatGPT, for instance, is an LLM developed by OpenAI, which secured itself as the fastest-growing consumer application after its release in November 2022.[7] Following the trend, other tech companies have launched AI tools, such as Microsoft Copilot, Lexis+ AI, and Llama. However, these tools are in their infancy and are susceptible to producing inaccurate or made-up responses, necessitating their assessment and reliability.

Though AI tools augment legal services, they pose risks like malpractice and legal liability and cause a miscarriage of justice if presented or relied upon without verification of their precision. AI solutions tend to produce plausible but inaccurate responses, which are referred to as hallucinations. Legal hallucinations refer to the phenomenon where AI models perpetuate fictitious precedents, misinterpret legal doctrines and statutes, provide inaccurate legal advice, or generate fabricated legal content. In the given scenario, the end user is liable to validate the outcomes of AI tools before making them part of their court filings or tendering AI-based legal advice.

There are various contributing factors leading to hallucinations: (1) limitations in training data, where AI models are trained on deficient or prejudiced datasets, and their generated content can be erroneous; (2) overconfidence in prediction, where AI models may produce results or generate content, often fabricating details instead of accepting uncertainty; and (3) model complexity, where due to the AI models' complexity and extensive data processing requirements, problems like overfitting, training data gaps, over-optimization, and a lack of grounding in real-world knowledge, can make LLMs prone to hallucinations.[8]

LLMs by design cannot differentiate between the truthfulness and falsehood of the generated content because these models mirror patterns from the data they are trained on without recognizing their precision. Hence,

---

papers.ssrn.com/sol3/papers.cfm?abstract_id=4539836; Livermore, M.A., F. Herron, and D.N. Rockmore. "Language Model Interpretability and Empirical Legal Studies" (Virginia Public Law and Legal Theory Research Paper No. 2023-69); Lo, C.K. 2023. "What is the impact of ChatGPT on education? A rapid review of the literature." *Education Sciences* 13 (4): 410; Nay, J.J., D. Karamardian, S.B. Lawsky, W. Tao, M. Bhat, R. Jain, A.T. Lee, J.H. Choi, and J. Kasai. 2023. "Large Language Models as Tax Attorneys: A Case Study in Legal Capabilities Emergence." preprint, arXiv, arXiv:2306.07075; Rodgers, I., J. Armour, and M. Sako. 2023. "How Technology Is (or Is Not) Transforming Law Firms." *Annual Review of Law and Social Science* 19: 299.

[3] Henry, J. 2024. "We Asked Every Am Law 100 Law Firm How They're Using Gen AI. Here's What We Learned." *ALM/Law.com*. https://www.law.com/americanlawyer/2024/01/29/we-asked-every-am-law-100-firm-how-theyre-using-gen-ai-heres-what-we-learned/.

[4] Avery, J.J., P.S. Abril, and A. del Riego. 2023. "ChatGPT, Esq.: Recasting Unauthorized Practice of Law in the Era of Generative AI." *Yale Journal of Law & Technology* 26: 64; Cyphert, A.B. 2021. "A Human Being Wrote This Law Review Article: GPT-3 and the Practice of Law." *UC Davis Law Review* 55: 401; Walters, E. 2018. "The Model Rules of Autonomous Conduct: Ethical Responsibilities of Lawyers and Artificial Intelligence." *Georgia State University Law Review* 35: 1073; Yamane, N. 2020. "Artificial intelligence in the legal field and the indispensable human element legal ethics demands." *Georgetown Journal of Legal Ethics* 33: 877.

[5] Stryker, Cole, and Mark Scapicchio. *What is Generative AI?*. IBM, accessed Apr. 25, 2025, https://www.ibm.com/think/topics/generative-ai.

[6] Feuerriegel, S., J. Hartmann, C. Janiesch, and P. Zschech. 2024. "Generative AI." *Business & Information Systems Engineering* 66: 111–26.

[7] Hu, K. "ChatGPT Sets Record for Fastest-Growing User Base – Analyst Note." Reuters, Feb. 2, 2023, https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/.

[8] Khan, A. "Top 5 Examples of AI Hallucinations and Why They Matter." *Lettria*, Oct. 1, 2024, https://www.lettria.com/blogpost/top-5-examples-ai-hallucinations.

https://doi.org/10.1017/jli.2025.10052 Published online by Cambridge University Press

accuracy in these models' responses is often inadvertent.[9] The models are not infallible and can perpetuate fabricated responses, creating real-world challenges where accuracy is paramount. Considering their fallibility, it is critical to understand the nature of their responses and the end user's corresponding duty to double-check these models' outcomes before using or relying on them.

RELEVANT LITERATURE

By applying Retrieval-Augmented Generation (RAG), a technique that combines the capabilities of a pre-trained LLM with an external data source, legal enterprises like LexisNexis and Thomson Reuters have claimed, though not supported by any empirical evidence, to have significantly mitigated the risk of hallucinations.[10]

A recent study proved that commercially available RAG-based, AI-driven legal tools still hallucinate. The study was conducted on Lexis+ AI, offered by LexisNexis; Westlaw AI-Assisted Research, produced by Thomson Reuters; Ask Practical Law AI, another Thomson Reuters product; and ChatGPT, offered by OpenAI. The research exhibited that 16.67 percent (1 out of 6) of Lexis+ AI and Ask Practical Law AI queries produced misleading or false information, whereas the rate for Westlaw was 33.33 percent (1 out of 3), although it was less prone to hallucinate in comparison with GPT-4. The frequency of accuracy reported in these tools was 65 percent for Lexis+ AI, 41 percent for Westlaw, and 19 percent for Practical Law AI.[11]

LLMs can transform legal operations, but this potential is hindered by their tendency to hallucinate, which is a critical challenge to the widespread adoption of LLMs. Another significant study examined hallucinations in open-domain settings, where LLMs were required to handle a wide range of legal prompts to provide exact responses to open-ended questions. The study revealed legal hallucinations of 69 percent with ChatGPT 3.5 and 88 percent with Llama 2. The results indicated that LLMs often provide superficially legitimate but incorrect replies to counterfactual legal queries. The research proved that LLMs such as ChatGPT are susceptible to generating responses incompatible with prevalent doctrines and case law. AI models aspire to provide legal services and information, but their inherent shortcomings in providing reliable and accurate information greatly obstruct this intent.[12]

Hallucinations in legal operations are a real-world challenge and should not be considered merely a theoretical concern. Plausible but fictional outcomes could necessitate novel methods for employing LLMs in situations where accuracy is paramount and inaccuracy is unacceptable, such as in legal, financial, and medical scenarios. According to a study undertaken on GPT 3.5, Llama 2, and PaLM 2, LLMs hallucinate from 69 percent to 88 percent of the time when responding to specific legal queries. These models are oblivious to their errors and bolster incorrect legal presumptions, fostering concerns about their authenticity and reliability and underscoring the significance of vigilant adoption of AI-powered solutions into the legal province.[13]

For example, the following instance demonstrates the breadth of hallucinations in legal operations: A summarization tool was employed to condense a lengthy document. AI fabricated certain legal terms and omitted critical features, resulting in a summary draft that misconstrued the original draft. Hallucinations of this kind could have severe repercussions where precision is inevitable.[14]

Fabricating citations is another critical instance of hallucinations. AI models can often hallucinate by creating fake references, apparently real, with all the information about the title, author, journal, periodical, or, in the instance of court precedent, the names of parties with a complete citation, etc. However, upon meticulous

[9] O'Brien, M. "Chatbots Sometimes Make Things Up. Is AI's Hallucination Problem Fixable?." AP News, Aug. 1, 2023.

[10] "Introducing AI-Assisted Research: Legal Research Meets Generative AI." *Thomson Reuters*, Nov. 15, 2023, https://legal.thomsonreuters.com/blog/legal-research-meets-generative-ai.

[11] Magesh, V., F. Surani, M. Dahl, M. Suzgun, C.D. Manning, and D.E. Ho. 2024. *Hallucination-Free? Assessing the Reliability of Leading AI Legal Research Tools*. preprint, arXiv, arXiv:2405.20362; Surani, F., and D.E. Ho. "AI on Trial: Legal Models Hallucinate in 1 out of 6 (or More) Benchmarking Queries." Stanford HAI (May 23, 2024), https://hai.stanford.edu/news/ai-trial-legal-models-hallucinate-1-out-6-or-more-benchmarking-queries.

[12] Dahl, Magesh, Suzgun, and Ho. "Large Legal Fictions" (n 1).

[13] Dahl, M., V. Magesh, M. Suzgun, and D.E. Ho. "Hallucinating Law: Legal Mistakes with Large Language Models Are Pervasive." Stanford Law School (Jan. 11, 2024), https://law.stanford.edu/2024/01/11/hallucinating-law-legal-mistakes-with-large-language-models-are-pervasive.

[14] Ibid.

https://doi.org/10.1017/jli.2025.10052 Published online by Cambridge University Press

examination, the source does not exist but is hypothetically generated. AI models would likely struggle with more complex or nuanced *Bluebook* citations.

In addition, while utilizing AI solutions, legal professionals may face challenges in comparing cases where the court reached opposite conclusions due to the nuanced nature of legal reasoning and interpretation. AI systems might encounter difficulties in understanding context, as cases depend on specific facts and jurisdictional nuances. Likewise, the legal language in opinions can be complex, often filled with jargon, especially when cases arrive at contrasting conclusions owing to jurisdictional differences, statutory interpretations, or factual circumstances.

For example, in a motion to dismiss filed in the US District Court for the District of Connecticut, a lawyer might like to compare two court cases reaching opposite conclusions on the same issue. After researching cases, the lawyer might cite *Castagno v. Wholean*[15] (reversing the dismissal of a grandparent's visitation petition for lack of jurisdiction in the absence of death, divorce, or a child custody proceeding) and *Frame v. Nehls*[16] (holding that the state's grandparent visitation statute did not apply in a paternity action because it was not an action for child custody). Though both cases involve grandparents' visitation rights, they differ in their legal interpretations, underscoring the importance of jurisdictional and statutory context in legal analysis.

Some researchers argue that AI-powered solutions might take over most of the tasks typically performed by lawyers, potentially multiplying access to justice at the cost of legal jobs.[17] However, these studies mainly focus on AI alone and miss the most significant part: the human interaction with these AI systems. Due to their inherent limitations, it is more likely that AI will assist humans rather than replace them.[18] Current research on AI often overlooks the interaction between humans and machines, missing the most probable application of AI in the legal field.

METHODS

Quantitative and comparative methods have been employed to conduct the proposed research, which was conducted between December 2024 and April 2025.[19] To check the accuracy, completeness, and hallucinations produced by ChatGPT-4, Copilot, DeepSeek, Lexis+ AI, and Llama 3, we prepared fifty legal questions divided into the following categories: constitutional doctrines; general law questions regarding new legal developments; questions requiring description; close-ended questions; and false propositions. We prompted each tool with the same questions and compared its responses with the correct answers in a spreadsheet. We calculated the performance of each tool separately in percentages for accuracy, incomplete response, and fabricated outcome. At the end of the analysis, we also demonstrated these tools with the help of charts. To quantify the performance of these AI tools, we employed three variables to evaluate the performance measure and accuracy-to-hallucination ratio. Through pseudocode, we represented the responses of these tools.

[15] *Castagno v. Wholean*, 684 A.2d 1181 (Conn. 1996).

[16] *Frame v. Nehls*, 550 N.W.2d 739 (Mich. 1996).

[17] Beioley, K., and C. Criddle. "Allen & Overy introduces AI chatbot to lawyers in search of efficiencies." *Financial Times*, Feb. 14, 2023; De Cremer, D., N.M. Bianzino, and B. Falk. 2023. "How Generative AI Could Disrupt Creative Work." *Harvard Business Review* 13: 8; Hatzius, J. "The Potentially Large Effects of Artificial Intelligence on Economic Growth (Briggs/Kodnani)." Goldman Sachs (Mar. 26, 2023); Schwarcz, D., and J.H. Choi. 2023. "AI Tools for Lawyers: A Practical Guide." *Minnesota Law Review Headnotes* 108: 1.

[18] Kolata, G. "When Doctors Use a Chatbot to Improve Their Bedside Manner." *International New York Times NA-NA*, June 12, 2023; Crootof, R., M.E. Kaminski, W. Price, and I. Nicholson. 2023. "Humans in the Loop." *Vanderbilt Law Review* 76: 429; Scholtes, J.J., and G. Vance. "AI+ Human: A Bright Future For Legal Co-Pilots." *Legal Tech Bridge*, Sept. 18, 2023, https://www.legaltechbridge.com/en/ai-human-a-bright-future-for-legal-co-pilots; Wendel, W.B. 2019. "The Promise and Limitations of Artificial Intelligence in the Practice of Law." *Oklahoma Law Review* 72: 21; Yamane, "Artificial Intelligence in the Legal Field" (n 4).

[19] Research on Copilot began on December 24, 2024, ChatGPT on December 26, 2024, Lexis+ AI on December 26, 2024, DeepSeek on January 29, 2025, and Llama 3 on January 30, 2025.

https://doi.org/10.1017/jli.2025.10052 Published online by Cambridge University Press

AI TOOLS: COMPARATIVE PERFORMANCE ANALYSIS

Given their probabilistic nature, AI solutions are prone to producing fabricated outcomes despite advancements in their performance. The following segment quantifies responses of ChatGPT-4, Copilot, DeepSeek, Lexis+ AI, and Llama 3 in terms of their accuracy, incomplete responses, and the hallucinations they create. Through pseudocode and mathematical equations, we drew a comparison of these tools by calculating the accuracy-to-hallucination ratio, providing insight into the decision-making tasks of these tools.

Mathematical Investigation

To mathematically represent the performance of AI-powered legal tools in terms of accuracy, incomplete responses, and hallucinations, we defined the following variables:

A_i: Accuracy of the tool i, expressed as a percentage.
I_i: Percentage of incomplete responses generated by the tool i.
H_i: Percentage of hallucinations produced by the tool i.

Where $i \in \{1,2,3,4,5\}$ corresponds to the AI tools:

ChatGPT-4
Copilot
DeepSeek
Lexis+ AI
Llama 3

A performance vector can capture each AI tool's performance (A_i, I_i, H_i), where:

ChatGPT-4: $(A_1, I_1, H_1) = (30\ percent, 34\ percent, 36\ percent)$.
Copilot: $(A_2, I_2, H_2) = (52\ percent, 22\ percent, 26\ percent)$.
DeepSeek: $(A_3, I_3, H_3) = (50\ percent, 20\ percent, 30\ percent)$.
Lexis+ AI: $(A_4, I_4, H_4) = (58\ percent, 22\ percent, 20\ percent)$.
Llama 3: $(A_5, I_5, H_5) = (36\ percent, 28\ percent, 36\ percent)$.

Total Performance Measure

To evaluate the overall performance of each AI tool in terms of accuracy, incomplete responses, and hallucinations, we defined a total performance measure P_i for each tool:

$$P_i = A_i + I_i + H_i$$

This equation summarizes the three key performance metrics, allowing for a holistic comparison of each tool's behavior.

Accuracy-to-Hallucination Ratio

A more insightful evaluation of the balance between accuracy and hallucination rates can be made by calculating the accuracy-to-hallucination ratio R_i:

$$R_i = \frac{A_i}{H_i}$$

This ratio helps to quantify how well each tool performs in terms of accuracy relative to the hallucinations it generates. A higher value of R_i indicates that the tool maintains better accuracy with fewer hallucinations, while a lower value suggests a higher propensity for hallucinations compared to its accuracy.

https://doi.org/10.1017/jli.2025.10052 Published online by Cambridge University Press

Comparative Analysis

The above metrics and ratios can be used to perform comparative analyses between the AI tools, offering insights into their strengths and weaknesses in legal decision-making tasks. Graphical representations, such as bar charts or radar plots, can further aid in visualizing the performance differences across accuracy, incomplete responses, and hallucinations for each tool.

Pseudocode Representation

```
Define performance for each tool as (accuracy, incompleteness, hallucinations)

ChatGPT = (30, 34, 36)
Copilot = (52, 22, 26)
DeepSeek = (50, 20, 30)
Lexis+AI = (58, 22, 20)
Llama 3 = (36, 28, 36)
List of tools
Tools = [ChatGPT, Copilot, DeepSeek, Lexis+ AI, Llama 3]
Function to calculate the total performance
Function TotalPerformance(tool):
   Return tool[0] + tool[1] + tool[2]
Function to calculate accuracy-to-hallucination ratio
Function AccuracyToHallucination(tool):
   If tool[2] > 0:
         Return tool[0] / tool[2]
   Else:
         Return 0
Loop through tools and print results
For each tool in Tools:
     Print "Total Performance:", TotalPerformance(tool)
     Print "Accuracy-to-Hallucination:", AccuracyToHallucination(tool)
End
```

Graphical Representation

The following charts demonstrate the frequency of accuracy and incomplete responses, as well as the tendency to generate hallucinations in these AI-powered solutions:

Figure 1 represents accuracy in AI-generated legal content through tools employed in the survey: the survey shows that Lexis+ AI secured the highest accuracy at 58 percent, followed by Copilot at 52 percent, and DeepSeek at 50 percent. Llama 3 provided 36 percent accuracy, whereas ChatGPT-4 secured the lowest accuracy at 30 percent.

Figure 2 exhibits incomplete responses: DeepSeek outperformed all the comparative tools by producing 20 percent incomplete responses. Copilot and Lexis+ AI yielded 22 percent incomplete responses. Llama 3 and ChatGPT-4 perpetuated 28 percent and 34 percent incomplete responses, respectively. However, the risk associated with the use of DeepSeek is critical. A recent study reported huge security risks concerning the use of DeepSeek, where a significant data breach exposed the sensitive information of over one million users. ClickHouse, DeepSeek's database, remained unprotected, permitting full control over database operations. This exposed a breach of sensitive information, such as chat histories and other critical data, which raised concerns regarding data management, practices, and compliance with privacy laws.[20] Consequently, several countries and organizations have banned

[20] Phillips, G. "Massive DeepSeek data leak exposes sensitive info for over 1 million users—what you need to know." Tom's Guide, Feb. 4, 2025, https://www.tomsguide.com/computing/online-security/one-million-sensitive-records-exposed-in-mass-deepseek-data-leak.

https://doi.org/10.1017/jli.2025.10052 Published online by Cambridge University Press

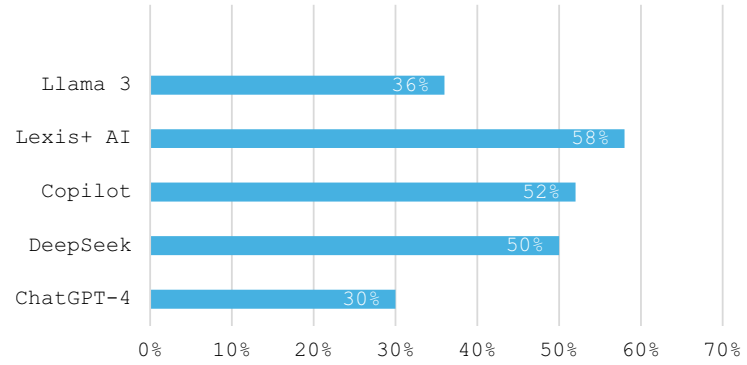

Figure 1. Percentage of Accuracy.

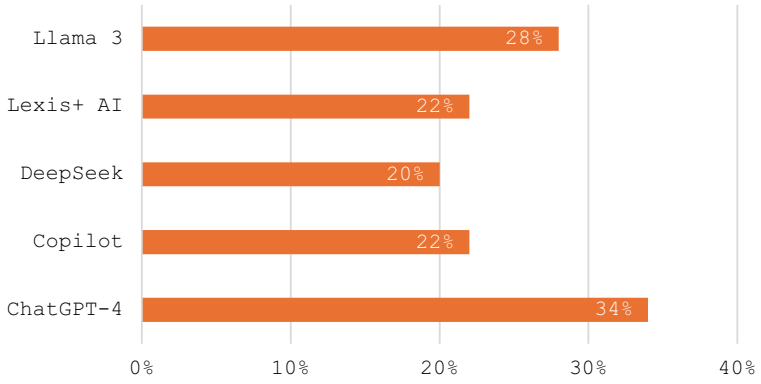

Figure 2. Percentage of Incomplete Responses.

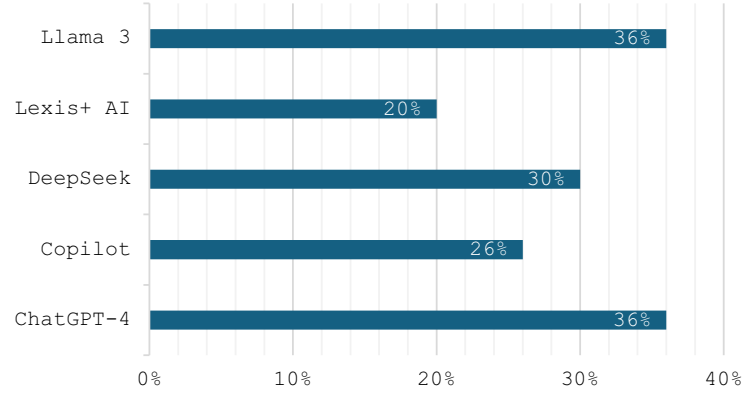

Figure 3. Percentage of Legal Hallucinations.

DeepSeek. Experts have criticized its security protocols, highlighting the potential for data exploitation by criminals, necessitating robust security protocols in AI-driven platforms to protect users and maintain trust.[21]

Figure 3 indicates how often these tools generate hallucinations through conceivable but fictional responses. The susceptibility to hallucinations is increased when an inaccurate fictitious prompt is given: ChatGPT-4 and Llama 3 produced 36 percent fictitious responses. DeepSeek produced 30 percent fabricated responses. Copilot had

[21] Ibid.

https://doi.org/10.1017/jli.2025.10052 Published online by Cambridge University Press

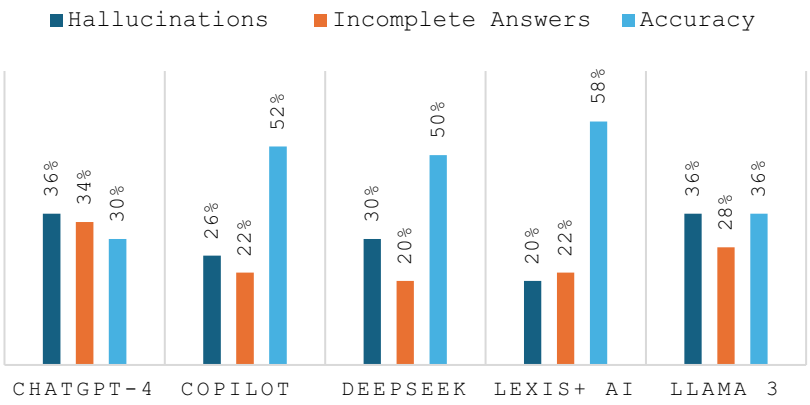

Figure 4. Proportion of Responses.

26 percent made-up responses, while Lexis+ AI outperformed all the comparative tools by reporting the lowest number of fabricated responses at 20 percent.

Figure 4 shows the proportions of responses observed in these tools, characterized as hallucinations, incomplete responses, and accuracy: ChatGPT-4 reported 30 percent accuracy, 34 percent incomplete responses, and 36 percent fabricated responses. Copilot produced 52 percent correct, 22 percent incomplete, and 26 percent fabricated responses. DeepSeek had 50 percent precision, 20 percent incomplete, and 30 percent fabricated responses. Lexis+ AI demonstrated 58 percent precision, 22 percent incomplete, and 20 percent fabricated responses, while Llama 3 achieved 36 percent correct, 28 percent incomplete, and 36 percent fabricated legal content. It was also observed that case law from lower courts was subject to more hallucinations than the superior courts. Interestingly, the frequency of hallucinations multiplied with a wrong or fictitious prompt. The tools advance and construct fabricated content in support of made-up citations instead of correcting or negating the existence of the proposed authority.

The Application and Implications of the Rag Model

Commercial services use Retrieval-Augmented Generation (RAG) to enhance their capabilities. All the AI tools deployed in this research leverage the RAG model, tailored to their use and strengths. ChatGPT-4, for instance, utilizes it to retrieve relevant information from external sources to improve conversational accuracy. Copilot uses it to ground its responses collected from credible, authoritative sources. DeepSeek models support RAG capabilities and are designed to combine retrieval-based and generative approaches to produce more accurate and relevant results.

Llama 3 uses RAG for handling large datasets for precise responses. Lexis+ AI uses GraphRAG—an advanced method to combine graph databases and RAG—to deliver high-quality insights with coherent and contextually relevant outcomes. It helps reduce AI hallucinations by leveraging structured data represented as a graph.[22] Though AI tools adapt the RAG model to their exclusive requirements, the core objective of combining retrieval with generation remains constant across all these tools.

In this technique, a prompt proposed to an LLM is first transferred to a retrieval system to search for a text-based source database, which creates a list of relevant documents. The retrieved data with the prompt is referred to the LLM to base its response exclusively on those documents, which we may refer to as an external database. By basing an LLM response exclusively on documents found, the chances of hallucinations could be significantly reduced, and it may also address any knowledge gap in an LLM by producing a comprehensive, accurate, and relevant response. The RAG opens new avenues to a more transparent and conceivable interaction with these models by allowing the end users to consider the same source that the LLM used to craft a response. RAG anchors LLMs' responses to

[22] "What is GraphRAG? Is it Better Than RAG?." *CapeStart Blog*, accessed Apr. 24, 2025, https://www.capestart.com/resources/blog/what-is-graphrag-is-it-better-than-rag/.

https://doi.org/10.1017/jli.2025.10052 Published online by Cambridge University Press

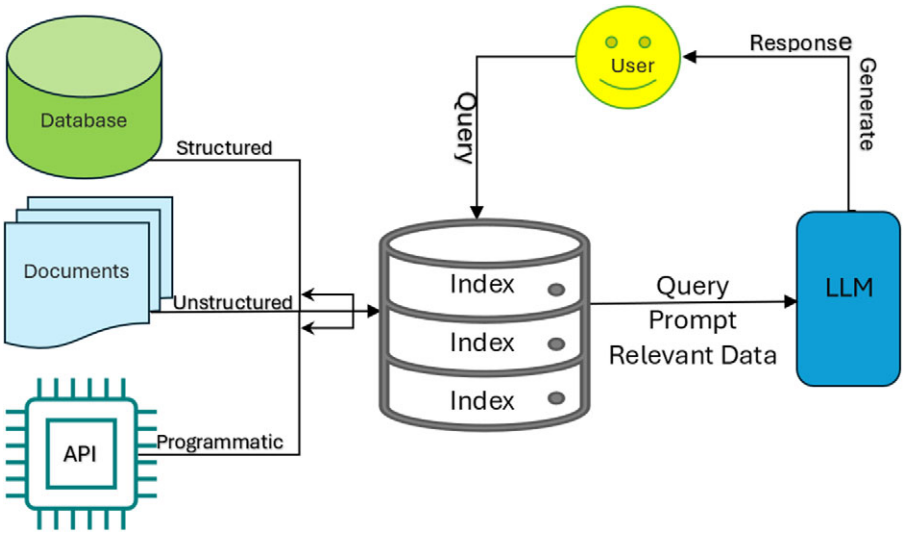

Figure 5. RAG Model.

information retrieved from an external database, thus ensuring the byproduct is factual and verifiable. Though RAG furthers understanding and trust in the LLMs, it limits the creativity of the generated response by design.[23] The above diagram conceptualizes RAG architecture:

Searching for a particular principle of US law, for instance, may involve the following steps: First, the process begins with a user inputting a prompt into the system, as demonstrated in Figure 5. In the second step, the prompt is forwarded to a retrieval system assigned to search a large database to identify relevant keywords, principles, or documents commensurate with the query. In the third step, the retrieved data and the prompt are transferred to an LLM, which uses the referred data to produce a well-informed, coherent, and contextually relevant response in the given case-relevant principle. In the fourth step, the LLM's generated response is delivered to the user within the same chatbot interface where the user initially entered their prompt.

Though the application of RAG models has contributed to the accuracy of responses, these models have confronted certain inherent challenges: complexity and the extraction of data from various sources—specifically, unstructured data, which could be complex.[24] RAG models manage both retrieval and generation, making them more complicated than the standard generative models.[25] Accuracy and scalability, which ensure the relevancy and accuracy of the retrieved data, are critical since poor retrieval can lead to the incorporation of inaccurate and irrelevant information in the model's response.[26]

Similarly, scaling the RAG model to handle voluminous data efficiently is challenging while concurrently dealing with the storage and retrieval of vector embeddings.[27] Some of these challenges are the following: 1) latency —the RAG model can slow down the performance of the AI tool, as the retrieval process can cause latency, making the RAG model slower;[28] 2) creativity—based on their reliance on the retrieved data, the models' creativity may be constrained by the scope and quality of the data being accessed;[29] 3) ambiguities—RAG may struggle with complex

[23] "AI Hallucinations (Why Would I Lie?)." *BitLaw*, Forsgren Fisher McCalmont DeMarea Tysver LLP, accessed Apr. 24, 2025, https://www.bitlaw.com/ai/hallucinations-and-AI.html.

[24] Fatima, F. 2024. "12 RAG Framework Challenges to Build Production-Ready LLM Applications." *Data Science Dojo Blog*, Mar. 28, 2024, https://datasciencedojo.com/blog/rag-framework-challenges-in-llm/.

[25] Solawetz, J. "The Pros and Cons of RAG Systems and Fine-Tuning in Natural Language Processing." *Arcee Blog*, Aug. 8, 2023, https://www.arcee.ai/blog/the-pros-and-cons-of-rag-systems-and-fine-tuning-in-natural-language-processing.

[26] Narendran, N. "How RAG Architecture Overcomes LLM Limitations." *New Stack Blog*, May 3, 2024, https://thenewstack.io/how-rag-architecture-overcomes-llm-limitations/.

[27] Fatima, F. "12 RAG Framework Challenges" (n 24).

[28] Emanuilov, S. "RAG Limitations." *Unfold AI Blog*, June 1, 2024, https://unfoldai.com/rag-limitations/#Latencysensitivity.

[29] Solawetz, J. "The Pros and Cons of RAG Systems" (n 25).

https://doi.org/10.1017/jli.2025.10052 Published online by Cambridge University Press

words having multiple meanings and contexts, leading to the retrieval of irrelevant or incorrect information placed into the model's response;[30] and 4) dependency—RAG models largely depend on the data's quality and comprehensiveness, which could lead to poor model performance where the dataset is biased, inaccurate, or incomplete.[31] A data validation process with vigilant and persistent improvements can diminish these challenges and improve AI tools' self-learning.

Domain-specific and geographical scenarios are the least addressed real-world challenges to RAG models' performance of the high-quality structured data in diverse domains like law, finance, and healthcare, which might be intricate, necessitating specific terminologies that might affect retrieval and generation quality.[32] The efficacy of RAG models is subject to their understanding of domain-specific knowledge, which may require fine-tuning and intelligent training at the production level. Geographical differences in language and dialects are another impediment to the performance of the RAG models. A system trained on specific data from one region may struggle with queries from another region, owing to linguistic or dialectal variations. Cultural context is essential for generating relevant and precise responses. So, the RAG models should be sensitive to cultural nuances that change across various geographical regions. Localized data sources significantly contribute to the RAG models' performance.[33] For instance, a model trained and designed for use in Pakistan would perform better if it had access to local information and databases. Likewise, a system designed for the US, trained on local datasets and information sources, might perform adversely in other geographical regions or when generating responses about other jurisdictions.

The following techniques can help optimize the domain- and region-specific complexities and challenges of the RAG models' performance: (1) Fine-tuning models on region- and domain-specific datasets, including curated datasets reflecting specific needs of the focused domain or region, can considerably improve their performance; (2) A hybrid approach to combining rule-based methods for structured data with machine-learning technologies for unstructured data can help largely resolve domain-specific issues;[34] and (3) Employing RAG systems on optimized infrastructure, such as using cloud-based platforms like AWS, Google Cloud, or Microsoft Azure, along with hardware optimization, can significantly enhance query processing and the overall performance of the models across different domains and geographical scenarios.

RESULTS AND DISCUSSION

Considering the responses to the fifty legal prompts, Lexis+ AI outperformed other AI comparative tools, but all these tools still hallucinate due to their probabilistic nature. Unlike ChatGPT-4, Copilot, DeepSeek, and Llama 3, which are general-purpose AI models, Lexis+ AI specializes in legal operations. Lexis+ AI is trained in legal-specific data, which enhances its accuracy and reliability for legal research and analysis. It integrates products like case law searching, legal drafting, and analytics within a single platform, making it more productive for legal professionals. With the additional features of security and data privacy, Lexis+ AI offers a unique user interface tailored to legal professionals, making it easier to navigate and use for specific legal tasks.[35]

Despite all these features, our research demonstrates that Lexis+ AI still perpetuates hallucinations and incomplete responses. Hence, its generated content cannot be relied upon or presented without counter-verification through conventional validation methods.

Employment of the RAG system overcomes the chances of hallucinations; however, all these tools are still prone to fabricated responses at different rates. Like other modern language models, DeepSeek is built on transformer architecture, which enables the model to produce proficient results by fusing advanced neural networks.

---

[30] Emanuilov, S. "RAG Limitations," (n 28).

[31] Ibid.

[32] Emanuilov, S. "Enhancing Domain-Specific RAG Systems." *Unfold AI Blog*, Aug. 18, 2024, https://unfoldai.com/rag-systems-evaluations/.

[33] Sunita Nadampalli, A.G., and Hamid Shojanazeri. "Improve RAG Performance with Torch compile on AWS Graviton Processors." *PyTorch Blog*, Dec. 20, 2024, https://pytorch.org/blog/improve-rag-performance/.

[34] Emanuilov, S. "Enhancing Domain-Specific RAG Systems" (n 32).

[35] "Legal AI vs. ChatGPT: What Makes Them Different?." *LexisNexis Blog*, Oct. 17, 2023, https://www.lexisnexis.com/community/insights/legal/b/thought-leadership/posts/legal-ai-vs-chatgpt-what-makes-them-different.

https://doi.org/10.1017/jli.2025.10052 Published online by Cambridge University Press

These models are mainly trained with a combination of supervision and learning from human feedback, tailored to human preferences and values.[36]

Moreover, the responses generated by these comparative tools substantiated variations with domain and region-specific queries, necessitating fine-tuning and infrastructural optimization of RAG models, including hardware optimization through graphics processing units (GPUs) or AWS Graviton processors, cloud-based solutions, efficient data management, and intelligent data training at the production level. The following segment concludes the research by further contributing to mitigating factors of the prospective hallucinations.

RECOMMENDATIONS AND CONCLUSION

The following are some recommendations for developers of AI models and legal professionals who are dealing with these models as end users:

(1) The AI tools' responses are significantly grounded in the quality of training data. Designers should ensure high-quality training data through training samples and subsequent testing through supervised and self-supervised learning. Blind datasets should be augmented based on their performance at the production level. Self-supervised learning is a process whereby a model generates labeled data from the trained data during the learning process without relying on externally provided labels, leveraging intrinsic structures within the input data to create deep-training signals. The complexities of the data queried at the production level, which models could not resolve, should be introduced and incorporated into the training datasets of supervised networks. Additionally, the complexities of self-labeling, a key aspect of self-supervised learning, should be improved at the production level.

(2) Overfitting is another frequent problem in machine learning. A model may perform exceptionally with training data but poorly with testing or validation data because of overfitting.[37] The latter happens when a model memorizes the training data, has too many parameters, is trained for too long, or is attuned to specifics and noise to such an extent that it fails to perform well on new unseen data. This makes the model overly complex and captures the random fluctuations in the training data rather than the desired output. The data for learning needs to present all levels of challenges and complexities as presented in production. Real-world search inquiries must be frequently added to the training of LLMs so that the networks learn the intrinsic patterns. Limited data availability can also lead to overfitting, so the dataset needs to be augmented intelligently to enhance the network's performance.

(3) Imposing limitations on AI outcomes can contribute to LLMs' performance. The application of sophisticated artificial neural network models on specific legal data and their combinations can improve the performance of the AI models.

(4) Developers should continuously improve and update the AI models by subsequent testing and getting feedback from end users through a process we may refer to as human-in-the-loop (HITL).

(5) Human oversight is recommended to validate the responses of AI tools by involving an expert in local laws. For instance, Leap AI has already initiated an optional verification process in which the AI-generated content is sent to an expert attorney to validate the accuracy of the AI-generated content.

(6) Developers may also consider fine-tuning models for specific legal tasks by adjusting the models' parameters to learn new tasks.

(7) By designing AI tools to justify and give reasons for their decision-making, developers can significantly enhance the reliability of the generated content. The models' self-explanatory attributes should be part of the RAG architecture. Based on its probabilistic nature and black box system that operates without providing information regarding its internal operation for decision-making, reasoned decision-making can significantly contribute to the overall efficacy and performance of AI tools.

(8) Designers should arrange online training sessions and guidelines on using AI tools for legal services without compromising accuracy, which is paramount in law.

[36] Team, G.E. "DeepSeek: A Comprehensive Guide." GetGuru, May 7, 2025, https://www.getguru.com/reference/deepseek.

[37] "What is Overfitting?." *IBM Blog*, Oct. 15, 2021, https://www.ibm.com/think/topics/overfitting.

https://doi.org/10.1017/jli.2025.10052 Published online by Cambridge University Press

(9) AI should assist but not replace humans, and legal professionals must be cautious about using AI tools for legal assistance. AI solutions should be employed to augment legal services, but not at the cost of subjective judgments.
(10) AI-generated content must be counter-verified through cross-referencing with traditional databases, peer reviews, consultation with senior colleagues, and other rounds of swift checking before presenting in the courts.
(11) Legal professionals should master precise prompt-crafting skills that help AI yield accurate responses.
(12) Legal professionals must be abreast of AI innovations and best practices with continuous audits of AI tools, ensuring their compatibility with legal standards, relevant laws, and policies.
(13) To avoid the apprehension of malpractice and civil liability, legal professionals must disclose to their clients the use of AI and keep the court updated on the extent of their reliance on AI, which is part of their ethical and professional responsibilities.
(14) Legal professionals should coordinate with developers to update the AI tools and report when these tools produce fabricated or inaccurate responses.

To conclude, AI is in its transitional phase of revolutionizing legal services. It exhibits ethical challenges like data privacy, bias, and fabricated responses. Existing research has demonstrated that Lexis+ AI is the most reliable and authentic tool for legal assistance, but it is still subject to hallucinations. The likelihood of fictitious responses amplifies where the dataset these AI tools are trained on is inaccurate, biased, or otherwise flawed, making an inverse relationship between hallucinations and accuracy. With the recommendations given, the overall performance of AI tools can be significantly improved to overcome the existing issue of hallucinations.

https://doi.org/10.1017/jli.2025.10052 Published online by Cambridge University Press