RETRIEVAL-AUGMENTED GENERATION: STRUCTURALLY ENABLING RELIABLE AI OUTPUT

BY TIAN BO ZHANG*

TABLE OF CONTENTS

* Staff Editor, Georgetown Law Technology Review; Business Law Scholar, Georgetown Law; J.D., Georgetown Law (2026); B.Com, University of Ottawa (2022).

INTRODUCTION

Imagine a coworker who is peerless in speed, responsiveness, and breadth of knowledge, yet possesses a catastrophic character flaw: they will occasionally, with absolute conviction, state a falsehood as an undeniable fact in a manner completely indistinguishable from their usual brilliant demeanor. In a vacuum, their productivity is a marvel - you might continue work with this individual on low-stakes administrative tasks where their speed is an asset, but would you ever allow this person on a "life-or-death" matter for the firm?

This is the central paradox facing white-collar workers in 2026. As investment in artificial intelligence has blossomed into a suite of tools now ubiquitous in professional workflows, practitioners feel a mounting pressure from both clients and firms to incorporate generative AI into their daily practice[1]. The allure is undeniable: Large Language Models (LLMs) offer a level of efficiency and versatility that feels like a cognitive superpower. Like our hypothetical coworker, LLM tools are peerless in speed, responsiveness, and breadth of knowledge – and they're always only one click away. We have a deep, practical desire to rely on these tools, because why not make our lives easier? Yet, we are frequently reminded that these models are prone to the same catastrophic flaw: they lie to us, they "hallucinate."

An "AI hallucination" is the generation of text that is contextually plausible but factually unfaithful.[2] For the legal industry, especially for unassuming lawyers who might choose to rely on the output of these LLMs, the consequences of such hallucinations are dire. The most notorious instance occurred in the district court case, *Mata v. Avianca*, where attorneys submitted a federal brief containing citations to six non-existent judicial decisions. When the court questioned the validity of these cases, the

---

[1] Roy Strom, *Clients Push Big Law Firms to Use Generative AI for Cost Savings*, BLOOMBERG L., Sept. 11, 2025, https://news.bloomberglaw.com/business-and-practice/clients-push-big-law-firms-to-use-generative-ai-for-cost-savings [https://perma.cc/6LDG-YE5M] (noting law firms are pressured by clients to disclose AI strategy as prerequisite to winning business, certain clients requiring AI for specific tasks like e-discovery).

[2] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Delong Chen, Weliang Dai, Ho Shu Chan & Andrew Madotto, *Survey of Hallucination in Natural Language Generation,* ACM COMPUT. SURV., Feb, 2022, at 4, https://doi.org/10.48550/arXiv.2202.03629 (defining a hallucinated text as "text that gives the impression of being fluent and natural despite being unnatural and unfaithful.").

lawyers, demonstrating a misplaced trust in the technology, insisted they had performed "reasonable diligence" by asking ChatGPT to verify the citations.[3] But in reality, the AI had doubled down on the hallucination, falsely confirming to the lawyers that the cases could be found on traditional legal databases like Westlaw and LexisNexis.[4]

The resulting fallout was a watershed moment for legal technology. The presiding judge ruled that the attorneys had acted in bad faith, violating Federal Rule of Civil Procedure Rule 11. The incident also reached the highest level of the American legal system when Chief Justice John Roberts addressed it in his 2023 Year-End Report on the Federal Judiciary. The Chief Justice warned that the habit of submitting AI-generated briefs citing non-existent cases is "always a bad idea," noting that while AI has great potential, it cannot replace the human "duty of candor" to the court.[5]

Thus, for the legal industry, reliability is the key bottleneck preventing AI from transitioning from a novelty into a professional-grade tool. If the legal profession is to move forward with AI, we require an architecture that prioritizes and ensures veracity and truth, thus moving past the fears of professional censure over convenience. Today, the most robust structural solution to this crisis of trust is Retrieval-Augmented Generation (RAG). To understand why RAG is the contemporary antidote to AI's "lying problem," one must first examine the structural reasons why a standard, static AI model is essentially designed to hallucinate.

I. LIMITATIONS OF CONVENTIONAL LLM ARCHITECTURE

To understand why a Large Language Model (LLM) hallucinates, one must first understand how it "knows" anything at all. In a conventional, non-RAG architecture, an AI's knowledge is stored entirely within its parametric memory.[6] As the name suggests, this knowledge is embedded within the model's billions of output parameters, which are mathematical weights that represent patterns, facts, and relationships distilled from a massive corpus of training data.[7] This training process essentially converts

---

[3] *Mata v. Avianca, Inc*., 678 F. Supp. 3d 443, 458 (S.D.N.Y. 2023).

[4] *Id.*

[5] John G. Roberts, Jr., Chief Justice of the U.S., *2023 Year-End Report on the Federal Judiciary*, SUPREME COURT, Dec. 31, 2023, at 6, https://www.supremecourt.gov/publicinfo/year-end/2023year-endreport.pdf [https://perma.cc/4F7C-KRYD].

[6] *See* Amber L. Solberg, *Understanding LLMs*, 9 GEO. L. TECH. REV. 257, 262 (2025) (discussing the role of parameters in LLMs).

[7] *Id.*

vast quantities of textual knowledge into mathematical statistics. When prompted for a response by the user, LLMs repeatedly use these extracted weights to predict next token in a sequence – think "next word in a sentence. Thus, applying the law of large numbers, an LLM's output represents a probabilistic average of the knowledge and linguistic features present in its training data[8] and is the reason why LLMs are remarkably powerful when answering general inquiries with broad and surface-level summaries. But the very compression that makes an LLM a powerful synthesizer also makes it a poor lawyer: inherent in the process of AI knowledge synthesis is a dissociation between a source and its conceptual relation.[9]

This dissociation between an LLM and its source material is best understood through the lens of digital compression, like for instance, uploading a high definition 4K resolution video to a streaming platform like YouTube. To save space and bandwidth, the platform "compresses" the file into a lower resolution, such as 720p.[10] While the general shapes, colors, and dialogue remain

---

[8] AJ Alvero, Quentin Sedlacek, Maricela León & Courtney Peña, *Digital Accents, Homogeneity-by-Design, and the Evolving Social Science of Written Language*, 45 ANN. REV. APPLIED LINGUISTICS 50, 62 (2025), https://doi.org/10.1017/S0267190525000042 (discussing how ChatGPT's "Dominican accent" is a homogenized "probabilistic average" of the linguistic features in its training data).

[9] *See* Peter Johnston, *Retrieval-Augmented Generation (RAG): Towards a Promising LLM Architecture for Legal Work?*, HARV. J.L. & TECH. DIGEST (Apr. 2, 2025), https://jolt.law.harvard.edu/digest/retrieval-augmented-generation-rag-towards-a-promising-llm-architecture-for-legal-work, [https://perma.cc/HN4G-5LCX] (explaining that the LLM training process "embeds knowledge . . . into the model's output parameters" and that "parametric memory does not allow the LLM to reliably cite specific sources").

[10] *See* timeturner88, REDDIT (r/youtube), *I uploaded a very high quality video but it looks low quality on YouTube. What can I do?* (2021), https://www.reddit.com/r/youtube/comments/oorrsy/i_uploaded_a_very_high_quality_video_but_it_looks/; *see also* Grid4WillScot, OBS PROJECT FORUM (Windows Support), *Why does the quality of my videos being reduced after uploading them to YouTube?* (Jan. 30, 2019), https://obsproject.com/forum/threads/why-does-the-quality-of-my-videos-being-reduced-after-uploading-them-to-youtube.99787/ [https://perma.cc/HX76-NN9X]; biomagical, TOM'S GUIDE FORUM (Apps General Discussion), *Quality lose after uploading to youtube, info for my issue in post*, (Sept. 1, 2015), https://forums.tomsguide.com/threads/quality-lose-after-uploading-to-youtube-info-for-my-issue-in-post.86802/ [https://perma.cc/TE8X-SPZM].

recognizable, the fine details are permanently discarded.[11] After the compression, you can likely discern that the object in the far distance looks like a car, but you cannot "zoom in" on the resulting 720p video to read its 4K resolution license plate. Parametric memory functions as a form of lossy compression for human language, much like YouTube's compression of uploaded videos. During training, the LLM does not store the original document's details, but extracts the statistical pattern of the text and discards the rest to make the model small enough to run on a server.[12] The result is that the AI remembers the concept but loses the receipt.[13]

For legal practitioners, this architectural quirk is unacceptable. For example, an LLM trained on civil procedure may be able to flawlessly articulate the "minimum contacts" test for personal jurisdiction.[14] It will seem to understand the relationship between an out-of-state defendant and a forum state because it has seen those concepts grouped together millions of times. However, because the model never processed the specific metadata of its training set, it often cannot reliably tie that rule back to the common law reasoning articulated in *International Shoe*. When a user demands a citation, the model is forced to engage in a "statistical guess." It knows what a citation should look like, and it knows the "neighborhood" of the law it is discussing, but it lacks the deterministic link to the source.[15] Thus, despite employing sophisticated teams of "AI Engineers" to create proprietary algorithms in an attempt to circumvent the hallucination issue, law firms like Morgan & Morgan have discovered that no amount of

---

[11] *See* timeturner88, REDDIT (r/youtube), *I uploaded a very high quality video but it looks low quality on YouTube. What can I do?* (2021), https://www.reddit.com/r/youtube/comments/oorrsy/i_uploaded_a_very_high_quality_video_but_it_looks/; *see also* Grid4WillScot, OBS PROJECT FORUM (Windows Support), *Why does the quality of my videos being reduced after uploading them to YouTube?* (Jan. 30, 2019), https://obsproject.com/forum/threads/why-does-the-quality-of-my-videos-being-reduced-after-uploading-them-to-youtube.99787/ [https://perma.cc/HX76-NN9X]; biomagical, TOM'S GUIDE FORUM (Apps General Discussion), *Quality lose after uploading to youtube, info for my issue in post*, (Sept. 1, 2015), https://forums.tomsguide.com/threads/quality-lose-after-uploading-to-youtube-info-for-my-issue-in-post.86802/ [https://perma.cc/TE8X-SPZM].

[12] *See* Johnston, *supra* note 9.

[13] *Id.*

[14] *Int'l Shoe Co. v. Washington*, 326 U.S. 310, 316 (1945) (holding that due process requires "certain minimum contacts" with the forum state such that the maintenance of the suit does not offend "traditional notions of fair play and substantial justice").

[15] *See* Johnston, *supra* note 9.

"prompting" and sophisticated user instruction can force a conventional LLM model to retrieve information it never actually stored. The AI will ultimately still be making predictive linguistic analysis to give you your citation.[16]

In short, conventional LLM architecture is systemically incompatible with a profession that relies on pinpoint citation to precedent. Thus, as proposed in the seminal 2020 paper by Patrick Lewis and his team at Meta AI, the solution is not to train bigger models, but to move toward a more suitable architecture.[17]

II. MECHANICS OF RETRIEVAL-AUGMENTED GENERATION

In a RAG-enabled system, the architecture inherently facilitates accurate citations by decoupling the LLMs knowledge function from its reasoning function.[18] Unlike conventional LLMs, which, as previously discussed, converts contextual knowledge into weights and compresses the results into its fixed parametric memory, a RAG-enabled LLM uses an additional layer of non-parametric memory to store and retrieve external data.[19] This transformation is akin to a cross-disciplinary litigator moving from a solo-practice to a premier large law firm. At their new firm, the litigator is handed the exact, highly relevant case files by their new assistant and a team of dedicated, expert librarians. Unburdened by the need to do their own legal research, the litigator is free to excel at what they do

---

[16] *See* Debra Cassens Weiss, *No. 42 Law Firm by Head Count Sanctioned Over Fake Case Citations Generated by AI*, ABA J. (Feb. 10, 2025), https://www.abajournal.com/news/article/no-42-law-firm-by-headcount-could-face-sanctions-over-fake-case-citations-generated-by-chatgpt [https://perma.cc/6EU5-5CM7] (reporting on the $5,000 in sanctions imposed on attorneys from Morgan & Morgan).

[17] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel & Douwe Kiela, *Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks*, 33 ADVANCES NEURAL INFO. PROCESSING SYS. (2020), https://doi.org/10.48550/arXiv.2005.11401 (proposing hybrid models combining parametric and non-parametric memory as a solution to solve the issues of pre-trained LLMs).

[18] *See* Qian Song, Qingyao Ai, Hongning Wang, Yiding Liu, Haitao Li, Su Weihang, Yiqun Liu, Tat-Seng Chua & Shaoping Ma, *Decoupling Knowledge and Context: An Efficient and Effective Retrieval Augmented Generation Framework via Cross Attention*, in *Proceedings of the 2025 Int'l World Wide Web Cong.* 1422, 1423 (2025), https://doi.org/10.1145/3696410.3714608 (explaining that RAG decouples an LLM's internal knowledge from the context data that is fed to it).

[19] *Id*.

best: reading, summarizing, and synthesizing the provided evidence.

The RAG architecture surrounding the LLM mirrors these advantages by attempting to supply the reasoning engine with curated, highly relevant sources. The structural isolation unburdens the LLM, allowing it to focus on reasoning while more suitable methods for citation and data retrieval support its efforts. This process generally occurs in three stages:

A. Retrieval of "Relevant" Information

To identify the most relevant contextual data, the architecture must first emulate the concept of "relevance." Many leading researchers and commercial RAG products offer different solutions for exactly how relevance is measured, but almost all of them implement a variation of "vector space scoring,"[20] which posits that groups of words and their relationships can be measured as coordinates in a multi-dimensional grid. Modern RAG systems use "vector embedding" to implement this thesis: every paragraph, or "chunk," of every document fed to the final model is first passed through an embedding technique that converts aspects of the text into a high-dimensional vector. This creates a multi-dimensional "space" on which every chunk of data can exist relative to one another and where each dimension represents an aspect of semantic similarity.[21] The system can then algorithmically compare the similarity between words and sentences: the closer their dimensions are to one another in this multi-dimensional semantic space, the more related the texts are to one another.[22]

---

[20] *See* Christopher D. Manning, Prabhakar Raghavan & Hinrich Schütze, *An Introduction to Information Retrieval* 109 (2008) (defining the "vector space model" as a framework where documents and queries are represented as vectors in a multi-dimensional space to determine relevance through mathematical distance); *see also* Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang & Haofen Wang, *Retrieval-Augmented Generation for Large Language Models: A Survey,* arXiv, Dec. 2023, at 3, https://doi.org/10.48550/arXiv.2312.10997 (discussing that retrieval encodes queries into vector representation to compute similarity scores).

[21] See Solberg, *supra note* 6, at 262 (explaining the "embedding layer" as the stage where tokens are converted into vectors and describing vectors as lists of numbers that capture the meaning and relationship of text in high-dimensional space).

[22] *See* Manning et al., *supra* note 20.

However, the number of similarity dimensions in commercial RAG products can total well over one thousand.[23] At this scale, straight-line distances become easily distorted by differences in text length. To solve this issue, systems calculate the "angle" separating the vectors rather than their exact dimensional distance in a formula known as "Vector Similarity."[24] By focusing purely on the angle, the system evaluates the directional alignment of the core concepts, ensuring that a short user query can perfectly match a lengthy, highly relevant document as long as their underlying meanings point in the same direction.[25] The formula is as follows:

$$Similarity\ Score = \cos(\theta) = \frac{\text{Vector (A)} * \text{Vector (B))}}{(\| \text{Vector(A)} \| * \| \text{ Vector (B)} \|)}$$

The numerator (Vector (A) * Vector(B)) is a dot product[26] of the vectors of the documents being compared and measures how much the document vectors "overlap."[27] If the vectors point in the same direction across many dimensions (e.g., "fiduciary duty" and "Delaware" will have a high dot product), the dot product will be high and will represent a high similarity score.

The denominator (‖Vector (A)‖*‖ Vector(B)‖) is the product of the Euclidian lengths of the document vectors and included to "normalize" the dot product, ensuring that the formula accounts for magnitude rather than length. This allows the system to recognize that a 10-word summary of a rule and a 500-word explanation of that same rule are semantically identical, even if their "size" is different.[28]

---

[23] *See* Solberg, *supra* note 6, at 262–63 (mentioning how LLMs are defined by the sheer volume of parameters and vectors).

[24] *See* Manning et al., *supra* note 20, at 121 (discussing how cosine similarity became the standard method of quantifying similarity between two documents due to document length).

[25] *Id.*

[26] In linear algebra, a dot product multiplies two vectors to produce a single value that represents the geometric angle between the two vectors if when plotted on a multi-dimensional plane. This resulting value generally represents how closely the two vectors line up. For more information and a helpful introduction to these topics, please visit *Dot Products*, KHAN ACADEMY, https://www.khanacademy.org/math/multivariable-calculus/thinking-about-multivariable-function/x786f2022:vectors-and-matrices/a/dot-products-mvc [https://perma.cc/93KQ-77AD].

[27] *See* Manning et al., *supra* note 20, at 121.

[28] *Id.* (discussing how two documents with similar contents might have a significance vector difference due to their difference in length and the effect of the denominator is to normalize the vectors).

The result of this formula is a similarity score with the cosine range between -1 and 1. A similarity score of 1 means the vectors are identical while a score of 0 means they are completely unrelated (orthogonal).[29] A system can then rank documents in its library based on their similarity scores to the task at hand to filter in and filter out documents appropriately. This is a critically important feature for LLMs in the legal industry because it allows end users to set a similarity threshold for acceptance (e.g. 0.8 similarity score). If no documents or sources can be retrieved that fit the threshold, the LLM will not accept those documents as sources and thus, will hallucinate answers based on deficient sources.[30]

B. LLM AUGMENTATION

Once the system identifies the most relevant chunks of text, it enters the Augmentation phase. This is where the AI's parametric memory (reasoning function) is augmented by the newly identified non-parametric resources.[31]

The system places these retrieved snippets into the LLM's "context window." The context window functions as the AI's active, short-term working memory.[32] The system then "stuffs" the LLM with an internal prompt that contains instructions acting as a behavioral leash. A basic RAG system prompt might look like this:

> You are an expert legal researcher. Below are several excerpts that are relevant to your task. Using only these excerpts, answer the user's question. If the excerpts do not contain the answer, you must state that you cannot find the information. Do not use outside knowledge and only write grounded answers.[33]

---

[29] *Id.* at 123 (discussing what a 0.999 similarity score means vs a 0.888 through a hypothetical with three vectorized novels).

[30] *See* Gao et al., *supra* note 20, at 10 (noting that the "ChatLaw" application filters out documents with poor relevance through similarity critique).

[31] *Id.* at 3.

[32] *What Is a Context Window?*, IBM, https://www.ibm.com/think/topics/context-window [https://perma.cc/8XJR-KLGT] (defining the context window as the amount of text the model can consider or "remember" at any one time).

[33] *Build a Retrieval-Augmented Generation (RAG) Agent with NVIDIA Nemotron*, NVIDIA DEVELOPER, https://developer.nvidia.com/blog/build-a-rag-agent-with-nvidia-nemotron [https://perma.cc/XW74-5EML] (providing a template system prompt that instructs the AI: "Based on the provided context, answer the following question . . . If the answer is not in the context, say 'I don't know'").

C. GROUNDED SYNTHESIS & TEXT GENERATION

The final stage is Generation. Here, the LLM applies its linguistic logic to the retrieved texts while adhering to the system prompt previously given. [34] Because the "knowledge" is sitting right in front of the model, the LLM no longer needs to "guess" the next word based on probability. Instead, it engages in "grounding," where the LLM will treat the specified text as the ground truth.

The grounding command in the system prompt forces every sentence produced by the AI to be tethered to a specific piece of evidence in the context window or to be an "I don't know."[35] This creates a functional audit trail that is the core benefit of using a RAG system. Through the grounding command, reliability is baked into the architecture, ensuring that professionals can reasonably trust that output of the AI is based on fact, not hallucination, if they have supplied the correct inputs.[36]

To reemphasize the importance of this feature, consider the following hypothetical: You ask an AI program to summarize and list all the district court interpretations of a new FTC ruling. In a non-RAG LLM, you should know to never ask the AI, "Where did you learn this specific fact?" because the answer is buried in billions of invisible weights. Any attempt to dig further will result in predictive analysis of what a district court citation should look like. In an RAG-enabled LLM, the system either knows exactly which vector coordinates (court dockets) are relevant or stops you before you have had a chance to inquire about its sources by answering that there is not enough data to generate an answer. Thus, an RAG-enabled system, regardless of whether it produced a conclusive response, is already more useful because its answer is inherently more reliable.

III. MODERN DAY IMPLEMENTATIONS

This architecture of accountability leads directly to the current landscape of legal technology, where RAG-enabled AI has been utilized by leading firms to power the industry's most sophisticated tools. The most successful implementations of RAG have focused on integration and ease of use: the two industry leaders, Thompson Reuters & Harvey, have found massive success, winning over

---

[34] *See* Gao et al., *supra* note 20, at 3.

[35] *See* NVIDIA DEVELOPER, *supra* note 33.

[36] *See* Lewis et al., *supra* note 17, at 9–10 (discussing the societal impact of an AI system grounded in fact).

lawyers and established law firms by integrating themselves into a lawyer's workflow seamlessly.[37]

For example, Thomson Reuters' CoCounsel represents an initial success in semantic legal research. By grounding the LLM in a closed universe of verified case law, it avoids the possibility of hallucinations while allowing for natural language queries regarding complex legal standards and the creation of synthesized memos that are grounded in accurate, pinpoint citations.[38] Similarly, in the realm of corporate law, Harvey utilizes RAG-enabled pipelines to interrogate massive data rooms resulting in contracts with top law firms like Latham & Watkins LLP and Paul, Weiss LLP.[39] By decoupling the AI's logic from its memory, these systems allow users to ask specific questions across its customers' private, proprietary data, retrieving relevant clauses and summarizing obligations with direct links to the source documents.

IV. WEAKNESSES & DIRECTION FOR GROWTH

However, even the most robust RAG architecture is still nascent technology—its current implementation reveals significant bottlenecks. The first and most significant is the "Garbage In, Garbage Out" problem: a RAG system is only as effective as its retriever and library of materials.[40] If the vector math fails to pull

---

[37] *One Million Professionals Turn to CoCounsel as Thomson Reuters Scales AI for Regulated Industries*, THOMSON REUTERS, https://www.thomsonreuters.com/en/press-releases/2026/february/one-million-professionals-turn-to-cocounsel-as-thomson-reuters-scales-ai-for-regulated-industries [https://perma.cc/ZN42-MQA4] (credits its own success to operating inside research, drafting and compliance platforms and other established professional systems); *Harvey Accelerates Enterprise AI With Agent-Powered Platform and Microsoft 365 Copilot*, HARVEY, https://www.harvey.ai/blog/agent-platform-and-microsoft-365-copilot [https://perma.cc/2N9K-YV8Y] (crediting its own success to Harvey's direct plug into Microsoft word, enabling smooth workflows).

[38] *Id.*

[39] *Latham Announces Firmwide Deployment of Harvey*, LATHAM & WATKINS, https://www.lw.com/en/news/2025/08/latham-announces-firmwide-deployment-of-harvey [https://perma.cc/8WW6-Y4PH] *Paul, Weiss Partners With Harvey AI on New AI Workflows Innovation*, PAUL, WEISS, https://www.paulweiss.com/insights/firm-news/paul-weiss-partners-with-harvey-ai-on-new-ai-workflows-innovation [https://perma.cc/4X9N-2LQT]; *How Legal Teams are Working Better With 25,000+ Workflows*, HARVEY, https://www.harvey.ai/blog/25000-custom-workflows [https://perma.cc/GWS7-EDDV] (Harvey attributing its success to custom workflows to structure massive datasets).

[40] Shabna Thattantavida & Gourab Sarkar, *Enhance RAG Performance Through Intelligent Chunking Strategies*, IBM DEVELOPER

the correct document due to a vague query or poor optical character recognition (OCR), the LLM will be "grounded" in irrelevant data. This results in a "Grounded Hallucination," where the AI provides a perfectly accurate summary of the wrong document.

Secondly, RAG LLMs do nothing to address the "Lost in the Middle" phenomenon.[41] As AI tools become more reliable, lawyers might be inclined to rely on them for larger and more complex textual tasks, only to experience that their LLMs' reasoning performance dips significantly when they need it most. This performance degradation occurs when an LLM becomes highly adept at using information at the very beginning or end of a long prompt but frequently ignores the evidence in the middle.[42] Since longform content is the primary work product of the legal profession, it is anticipated that legal professionals will require further attention from the AI industry to address this issue.

Finally, recent studies have found that RAG introduces a new problem known as the "Chunking Dilemma," or contextual fragmentation.[43] Because documents are often broken into smaller segments to fit into vector databases, the AI can lose the "connective tissue" of a complex legal argument; if a crucial, analogous line of reasoning appears on page five but the "relevant text" occurs on page fifty, the retriever may only pull the latter. This leaves the LLM unable to understand complex relationships.[44] These weaknesses highlight that while RAG bridges the gap

---

(Feb. 2, 2025), https://developer.ibm.com/articles/awb-enhancing-rag-performance-chunking-strategies/ [https://perma.cc/5SC3-E4JN] ("the principle of garbage in, garbage out applies—poorly segmented data leads to suboptimal results.").

[41] *See* Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni & Percy Liang, *Lost in the Middle: How Language Models Use Long Contexts*, 12 TRANS. ASSOC. COMPUTATIONAL LINGUISTICS, July 2023, at 2, https://doi.org/10.48550/arXiv.2307.03172 (finding that LLM performance is highest when relevant information occurs at the very beginning or end of the input context, but significantly degrades when models must access information in the middle of long contexts).

[42] *Id.*

[43] *See generally* Amine Kobeissi & Philippe Langlais, *Decomposing Retrieval Failures in RAG for Long-Document Financial Question Answering*, ARXIV, Feb. 23, 2026, https://doi.org/10.48550/arXiv.2602.17981.

[44] Alice Gomstyn & Amanda Downie, *RAG Problems Persist – Here Are Five Ways to Fix Them*, IBM DEVELOPER, https://www.ibm.com/think/insights/rag-problems-five-ways-to-fix [https://perma.cc/5WBF-5VUJ] ("The RAG challenges routinely confronting users include . . . inability to understand complex relationships").

between toy and tool, it does not remove the essential human-in-the-loop requirement for professional verification.

CONCLUSION

In the end, the transition from parametric memory to Retrieval-Augmented Generation (RAG) is about fixing a fundamental character flaw. For years, we have been dazzled by AI's speed but paralyzed by its habit of lying with a straight face. Even though the technology is not perfect, this move from statistical guessing to the evidence-based reasoning of RAG is a step in the right direction and is a milestone in the coming AI revolution.