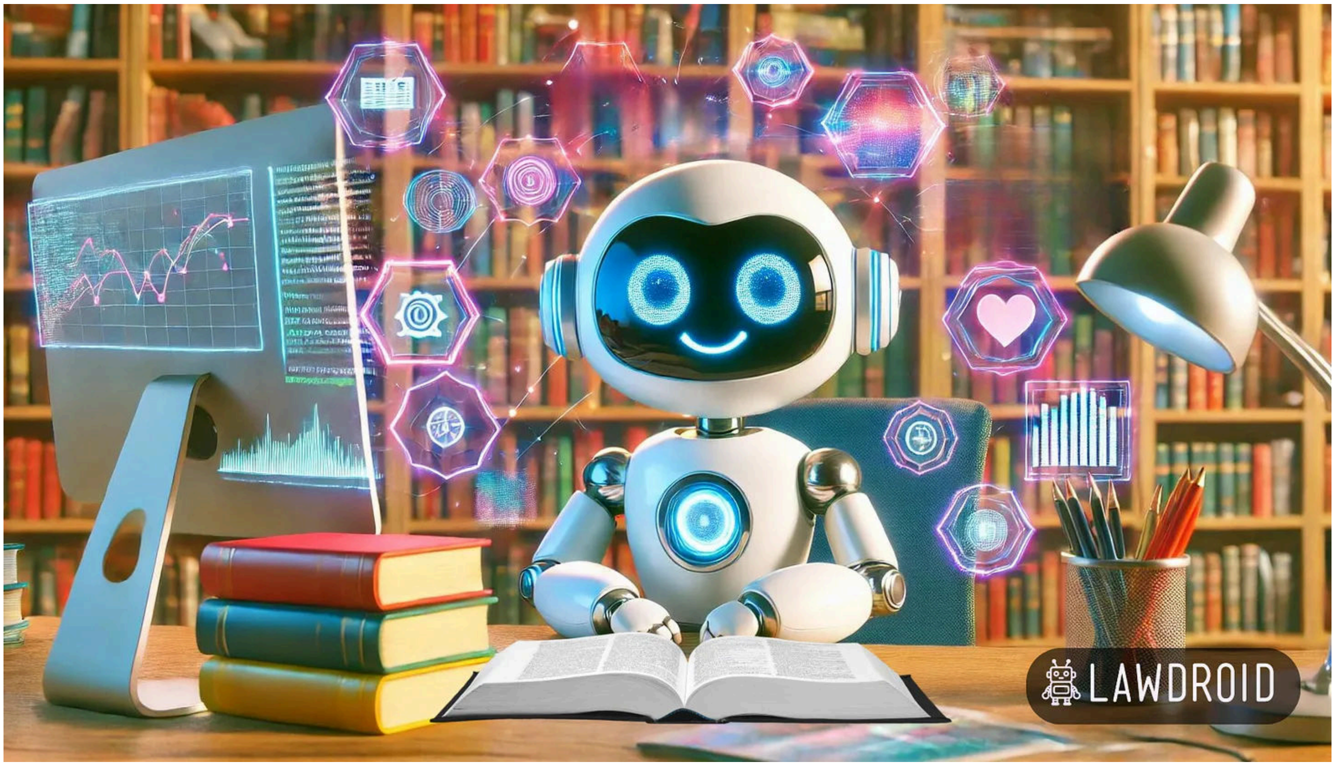
RAG: Why Does It Matter, What Is It, and Does It Guarantee Accuracy?

Where I explain what RAG is and why it's reliable enough for most applications

TOM MARTIN
JUN 29, 2024

❤ 5 💬 🔁 2 Share ⋯



Thanks again to you dear reader! You're dear to my heart. ❤️

As for you noobs, make yourself comfortable, the beer is in the fridge. 🍺

LawDroid Manifesto is a reader-supported publication. To receive new posts and support my work, consider becoming a free or paid subscriber.

This substack, LawDroid Manifesto, is here to keep you in the loop about the intersection of AI and the law. Please share this article with your friends and colleagues and remember to tell me what you think in the comments below.

If you thought my articles about prompt engineering and hallucinations were nerdy, well then strap yourself in! I'm going to take you into the world of "retrieval augmented generation" or "RAG" for short. If you've been following the discussion about the use of AI in the law, you probably heard the term and may have wondered what it means. I thought I'd perform a public service of explaining it in detail because it's something that sometimes get glossed over. And, if you're talking to AI vendors, you want to understand it. The reality is that it may sound overly technical, but it's really like giving AI the ability to take an open book exam. Who would score better, someone taking an open book exam or a closed book exam? Makes sense, right?

If this sounds interesting to you (then you get a pocket protector my friend as you are now nerd-certified), please read on...

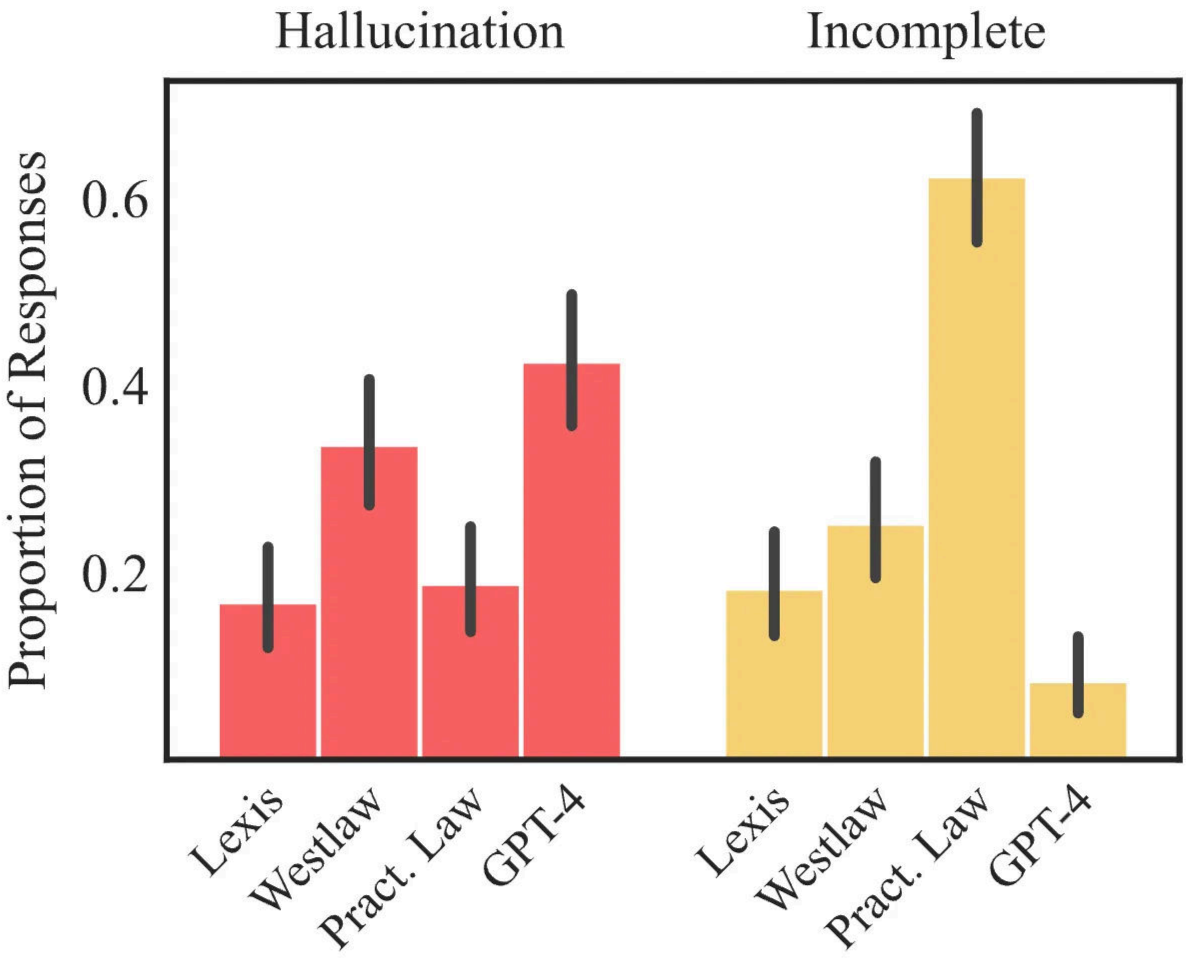
Why Does RAG Matter?

Let's start with answering the "Why?" question first instead of the what.

"Why does retrieval augmented generation matter to me?" is likely the first question that comes to mind, especially for what appears to be a somewhat technical subject. When addressing lawyers' concerns about hallucinations, retrieval augmented generation (RAG) has been cited by legal research companies as the solution.

Casetext claimed that its use of retrieval augmented generation "eliminat[ed]" hallucinations and that its flagship product CoCounsel was the "world's first reliable AI legal assistant" (Casetext, 2023). Thomson Reuters, which acquired Casetext, stated that RAG "avoid[ed]" hallucinations (Thomson Reuters, 2023). LexisNexis claimed that RAG guaranteed "hallucination-free" legal citations (LexisNexis, 2023).

However, doubt has recently been cast on these claims by a study, “Hallucination-Free? Assessing the Reliability of Leading AI Legal Research Tools,”[1] written by researchers from Stanford and Yale intent on assessing the accuracy of AI-powered legal research tools. “[W]e find that the AI research tools made by LexisNexis (Lexis+ AI) and Thomson Reuters (Westlaw AI-Assisted Research and Ask Practical Law AI) each hallucinate between 17% and 33% of the time.”[2] By comparison, GPT-4’s hallucination rate is 43%.



AI Legal Research Hallucination Rates

By the way, for a whirlwind tour of hallucinations, read my other article: “Hallucinations: What Are They, Why Do They Happen, How to Fix Them?”

Given the considerable disparity between legal research providers' RAG-powered claims to be hallucination-free, on the one hand, and testing of those self-same tools revealing hallucination rates between 17 to 33%, on the other hand, it begs the question for the legal practitioner: What is RAG and does it work?

What is Retrieval Augmented Generation?

Retrieval Augmented Generation (RAG) is a technique that enhances the output of a large language model (LLM) by incorporating relevant information from external knowledge sources, such as a web pages, documents or a database. [3]

RAG is Like an Open Book Exam

RAG is like giving an LLM an open book exam. The LLM can "read" relevant material before answering the question. The opposite, directly asking an LLM a question without the advantage of RAG or any further information, is like a closed book exam. The LLM cannot reference any external materials when it develops its answers.

To continue with the analogy for a moment (because it is weirdly appropriate and descriptive of how LLMs work), let's ask this question: What does a student rely on when writing a closed book exam?

In a closed book exam, a student would answer exam questions based on what they remember being taught in the class and generally from their education up to that moment in time. Likewise, an LLM, like GPT-4 was trained on 1.76 trillion parameters of information. [4] The training data likely includes a diverse range of web pages, books, articles, and other texts, but the specifics have not been made public by OpenAI. This forms the model's "parametric memory" (because those facts are stored in its parameters or weights created from its training) and this is why you may have heard of different models having knowledge cut off dates. For example, the most recent knowledge cut off for GPT-4 is April 2023 and October 2023 for GPT-4o.

Who Invented RAG?

Like many innovations in AI, RAG was not invented by a single individual, but rather emerged through collaborative efforts of a number of researchers. The key paper that introduced and formalized the RAG technique was: "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks." [5]

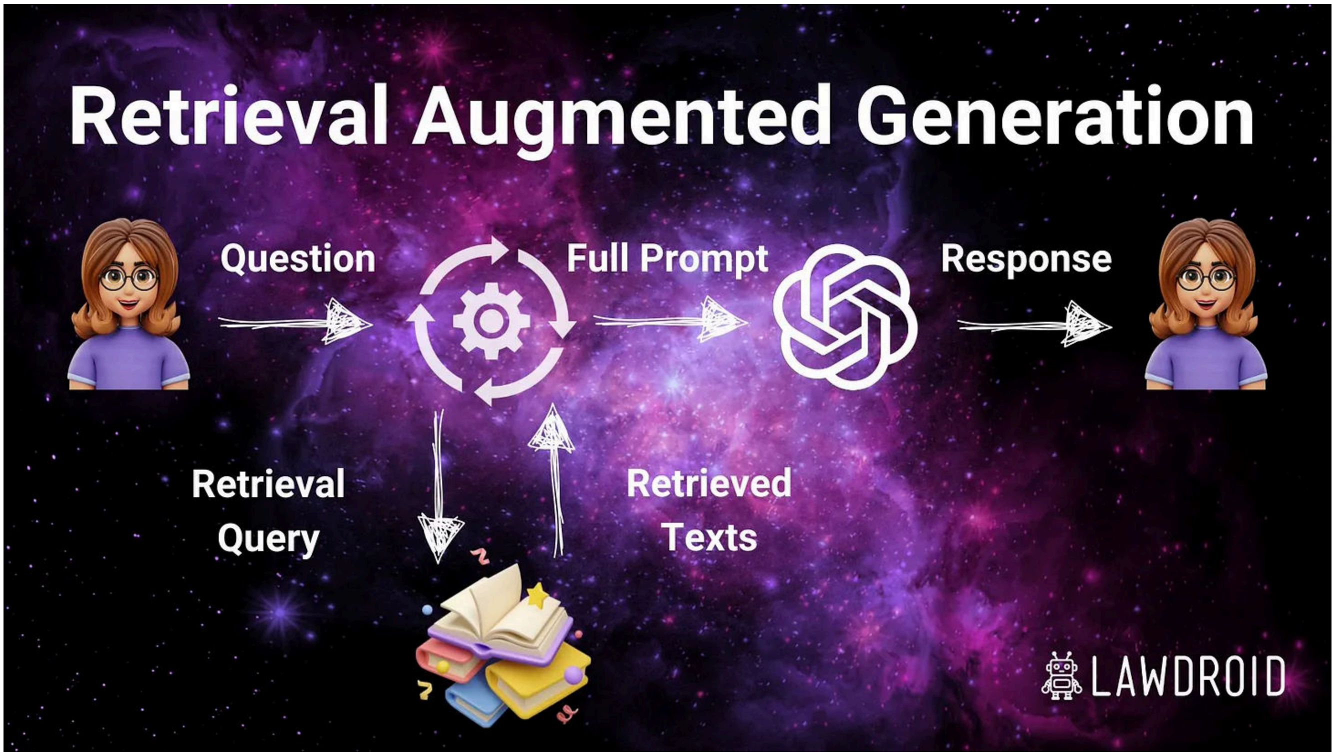
Patrick Lewis, the lead author, who now leads a RAG team at Cohere, acknowledged that while his team coined the term "RAG," the underlying concepts built on previous work in information retrieval, question answering, and language modeling. Lewis also acknowledged that Retrieval Augmented Generation was a horrible name choice: “We definitely would have put more thought into the name had we known our work would become so widespread… We always planned to have a nicer sounding name, but when it came time to write the paper, no one had a better idea.” [6] 😂

From its origin, RAG was conceptualized as a technique to minimize hallucinations and make LLMs more reliable. With the use of RAG, an LLM “is more strongly grounded in real factual knowledge [and that] makes it ‘hallucinate’ less with generations that are more factual, and offers more control and interpretability.” [7]

How Does RAG Work?

Now that we have a notion of what RAG is, and why it’s important to know something about it, let’s see how it works.

RAG Step-by-Step



Retrieval Augmented Generation Illustrated

- **Question / retrieval query**: When a user submits a query, it is converted to a vector representation and matched against the vector database to retrieve relevant information. This retrieves semantically similar information, not just exact keyword matches.
- **Retrieved texts**: The retrieved relevant data is added to the user's input to create an augmented prompt for the LLM. Retrieved documents are ranked by their relevance score (similarity to the query), allowing the system to prioritize the most pertinent information. A relevance threshold can be set to exclude documents that fall below a certain similarity score, ensuring only sufficiently relevant information is retrieved.
- **Full prompt / response**: The LLM uses the augmented prompt, combining the retrieved information (non-parametric memory) with its training data (parametric memory), to generate a more accurate and informed response.

This technique is particularly useful for applications like customer support chatbots, internal Q&A systems, and other scenarios where access to current, domain-specific information is crucial.

RAG Benefits

Key benefits of RAG include:

1. Improving accuracy and reliability of LLM outputs by grounding them in external sources of facts.
2. Providing access to current information beyond the LLM's original training data.
3. Reducing hallucinations and false information from LLMs.
4. Enabling citation of sources, increasing transparency and trustworthiness.
5. Lowering costs by reducing the need for frequent model retraining.

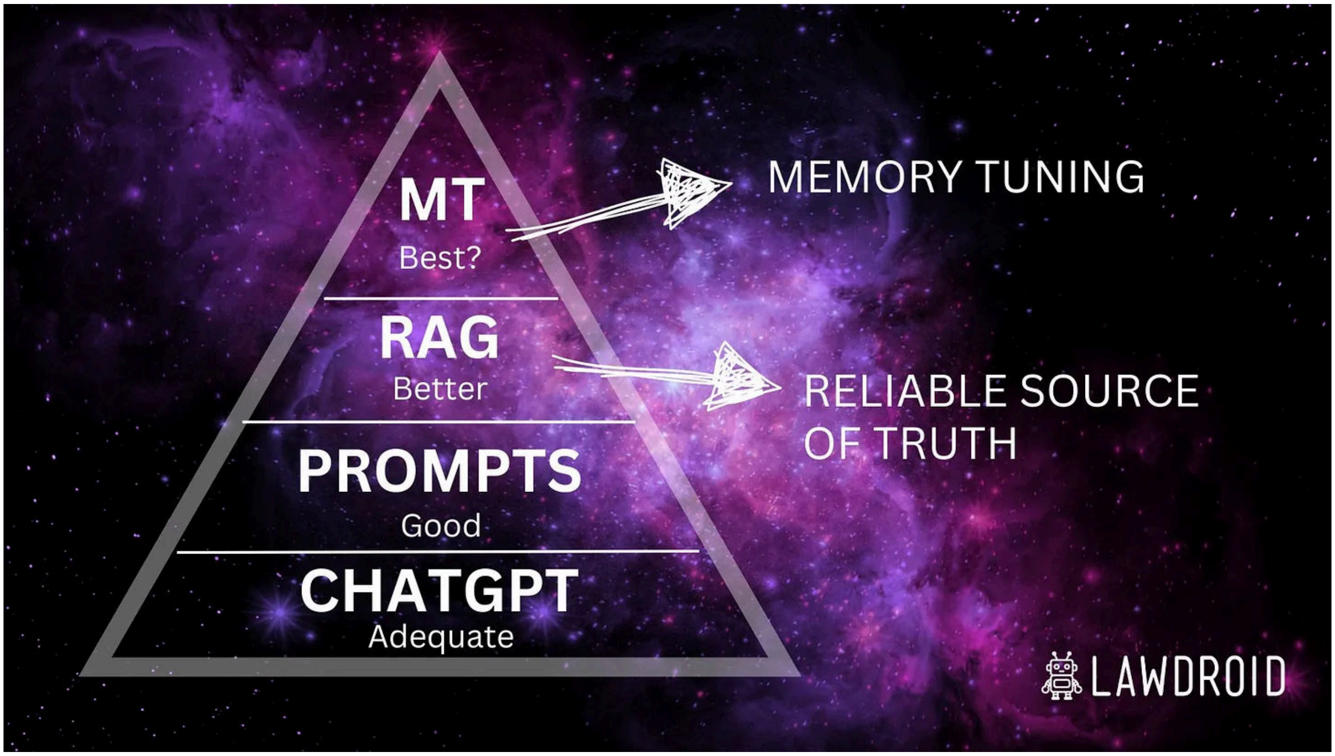
RAG addresses several challenges of LLMs by grounding responses in authoritative, up-to-date sources. It allows organizations to leverage their own data without retraining the entire model, making it a cost-effective approach to improving LLM outputs for specific domains or internal knowledge bases.

RAG Challenges

The main challenges of RAG include:

1. Loss of context when working with large datasets. Breaking documents into smaller pieces can result in losing important connections and nuances.
2. Retrieving outdated content from the database. If documents are not updated, the RAG system may still use old information, leading to inaccurate responses.
3. Slow response generation. Latency can occur at different steps of the RAG pipeline, requiring optimization.
4. Handling perplexity. A number of equally-likely next-word alternatives available during text generation, causing the LLM to be perplexed by its choices.
5. Hallucinations. While RAG helps reduce hallucinations, it doesn't eliminate them entirely, especially when dealing with larger, more complex document sets, like primary and secondary legal sources.

To address these challenges, solutions are being explored, including document hierarchies, knowledge graphs, recursive retrieval methods, self-criticism [8], and combining RAG with other techniques like prompt engineering and fine-tuning [9].



LLM Accuracy Pyramid

For a comprehensive backgrounder on prompt engineering, read my other article: "Prompt Engineering: What Is It, Why It's Important, and Is It Obsolete?"

Does RAG Guarantee Accuracy?

No, retrieval augmented generation does **not** guarantee accuracy and it does **not** eliminate hallucinations. However, RAG does a very good job of minimizing hallucinations and making LLMs reliable enough for many applications.

RAG Improves Reliability

RAG helps minimize hallucinations in LLM responses through several key mechanisms:

- Grounding responses in external knowledge (non-parametric memory): RAG retrieves relevant information from a curated knowledge base before generating a response. This grounds the LLM's output in factual, up-to-date information rather than relying solely on its pre-trained knowledge (parametric memory).
- Reducing reliance on pre-trained knowledge: By augmenting the LLM with external information, this reduces the need for the model to "fill in the gaps" with potentially inaccurate information (hallucinations).
- Providing citations: Many RAG solutions allow for citing the sources of retrieved information, increasing transparency and allowing users to verify the information.
- Constraining generation: The retrieved context acts as a constraint on the LLM's generation, guiding it towards more factual and relevant responses.

By leveraging these mechanisms, RAG significantly reduces the likelihood of hallucinations in LLM responses, improving the overall reliability and trustworthiness of AI-generated content.

Reliable Enough for the Risk?

Any analysis of LLM accuracy and reliability must ask: *Reliable enough for what?*

The framing of the debate in legal to date has focused almost exclusively on “bet the farm” and “cut the blue wire or red wire”-type uses of LLMs. While requiring 100% accuracy may make sense when asking an LLM to moot questions relating to the calculation of the statute of limitations or sentencing for capital murder, it is not reasonable for low or medium risk applications that also exist within the legal context.

There are many applications of RAG that are perfectly suited for the level of risk presented by LLMs’ current state of the art. To assess if an LLM is fit to purpose, we must classify the risk of the use case: Is it low risk, medium risk or high risk?

Here are some working definitions [10] I have developed:

- **Low Risk**: Use cases where the consequences of errors are minimal, affecting user convenience rather than critical outcomes. Errors can be easily corrected without significant negative impact. This includes applications such as AI-enabled video games or spam filters.

- **Medium Risk**: Use cases where errors can lead to moderate negative consequences, such as financial loss or misinformation, but typically do not involve life-threatening or severe legal ramifications. For instance, when using AI chatbots, humans should be made aware that they are interacting with a machine so they can take an informed decision to continue or not.
- **High Risk**: Use cases where errors can result in severe negative consequences, including health and safety issues, significant legal problems, or substantial financial loss. Some examples, affecting people's fundamental rights (evaluation of the reliability of evidence, automated examination of visa applications).

LLMs can be used in legal for much more than just narrow, deep legal research the consequence of which is catastrophe if mishandled.

Low risk LLM use in a legal practice may include translating text from one language to another, checking for grammar errors, helping to draft an email or letter, drafting an outline for a blog post, formatting legal citations, enabling natural language search of case law, or generating an image. There are a myriad of other low risk uses.

Medium risk LLM use in the legal context may include creating a draft of a contract or pleading, redlining a contract, providing general legal information and answering FAQs (such as, what are the steps to get divorced?, how do I get a restraining order?, how much does it cost to get a will?), summarizing documents (such as deposition transcripts, financial statements, emails and texts), or briefing cases. There are of course many other possibilities.

Sometimes good enough is good enough.

A couple of considerations make LLM usage more palatable for medium and even high risk applications within the legal context:

- Using human beings for quality control. In fact, legal ethics and federal rules requires it. [11] Lawyers are required to exercise their own independent judgment and to supervise the work of unlicensed individuals. This includes reviewing AI-generated work product to ensure it comports with professional standards.

- Shared AI accuracy benchmarking. Recently, efforts have begun to establish industry benchmarks for AI accuracy. [12] This is a critical need. Rather than AI vendors relying solely on their internal benchmarks for accuracy, a shared industry standard would foster comparison, transparency and trust.

Closing Thoughts

While retrieval augmented generation is not a panacea that guarantees 100% accuracy and eliminates all hallucinations, it represents a significant step forward in making LLMs more reliable and grounded in factual, up-to-date information. By augmenting LLMs with external knowledge, RAG helps to guide generation, reduce reliance on outdated pre-trained knowledge, and provide transparency through source citations.

Crucially, the question we must ask is not simply whether RAG delivers perfect accuracy, but rather whether it is reliable enough for the level of risk presented by specific use cases. In the legal context, there is a wide range of potential applications for LLMs, from low-risk tasks like language translation and grammar checking to medium-risk activities such as contract drafting and document summarization. By classifying the risk level of each use case and implementing appropriate human oversight, law firms can safely harness the power of LLMs to streamline workflows and enhance productivity.

Moreover, the legal profession's ethical obligations and federal rules mandating lawyer supervision of AI-generated work product provide an additional layer of quality control. By exercising independent judgment and carefully reviewing the output of RAG-powered LLMs, legal professionals can catch and correct errors, mitigating the impact of any remaining hallucinations.

The development of shared industry benchmarks for AI accuracy represents a critical step towards fostering trust and transparency in the use of LLMs within the legal sector. By establishing objective standards for evaluating the reliability of RAG-powered systems, legal professionals can make informed decisions about when and how to deploy these powerful tools.

As the technology continues to advance, retrieval augmented generation, coupled with human expertise and robust accuracy benchmarks, holds immense promise for transforming the practice of law.

By embracing these innovations while remaining mindful of their limitations, the legal community can harness the power of AI to deliver more efficient, effective, and accessible legal services to clients. Let's go!

POLL

Would you use a RAG-powered legal research tool?

Yes	56%
No	11%
It depends ;)	33%

9 VOTES · POLL CLOSED

By the way, if you'd like to learn more about how how AI works and how it will impact the legal profession, you should apply to LawDroid University!

My NEW 5-part webinar series, Generative AI for Lawyers: Empowering Solos and Small Law Firms, is now available at LawDroid University.

LawDroid University is available for free for everyone to use.

- **Free to use** - It's 100% free educational content for everyone, just sign up below.

- **Insightful** - Get educated about the intersection of artificial intelligence and the law as taught by experts.
- **Value Packed** - Filled with videos, summaries, key takeaways, quotable quotes, transcripts and more! Find sessions on AI and the State of the Art, Ethics, Access to Justice, Practice of Law, Education, and the Business of Law.
- **AI Q&A** - Ask a chatbot questions about the content and get fully informed answers immediately.

👉 To immerse yourself in this enriching educational voyage, learn more, or sign up, please visit https://lawdroid.com/subscriptions/lawdroid-university/.

LAWDROID
5-PART WEBINAR SERIES
Generative AI for Lawyers
Empowering Solos and Small Law Firms
lawdroid.com

[1] Magesh et al., Hallucination-Free? Assessing the Reliability of Leading AI Legal Research Tools, https://dho.stanford.edu/wp-content/uploads/Legal_RAG_Hallucinations.pdf

[2] Id.

[3] What is RAG (Retrieval-Augmented Generation)?, https://aws.amazon.com/what-is/retrieval-augmented-generation; What Is Retrieval-Augmented Generation, aka RAG?, https://blogs.nvidia.com/blog/what-is-retrieval-augmented-generation.

4 Roberts et al., How Much Knowledge Can You Pack Into the Parameters of a Language Model?, arXiv:2002.08910 [cs.CL] 10 Feb 2020, https://arxiv.org/pdf/2002.08910. A parameter refers to a numerical value that the model learns during training. These parameters are the components that make up the model's internal representation of language and knowledge.

5 Lewis et al., Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks, arXiv:2005.11401v4 [cs.CL] 12 Apr 2021, https://arxiv.org/pdf/2005.11401. This paper was published in 2020 as part of the Advances in Neural Information Processing Systems (NeurIPS) conference. The authors were researchers from Facebook AI Research, University College London, and New York University.

6 What Is Retrieval-Augmented Generation, aka RAG?, https://blogs.nvidia.com/blog/what-is-retrieval-augmented-generation.

7 Lewis et al., Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks, https://arxiv.org/pdf/2005.11401.

8 OpenAI today released CriticGPT, a model based on GPT-4, writes critiques of ChatGPT responses to help human trainers spot mistakes during RLHF, https://openai.com/index/finding-gpt4s-mistakes-with-gpt-4; McAleese, et al., LLM Critics Help Catch LLM Bugs, https://cdn.openai.com/llm-critics-help-catch-llm-bugs-paper.pdf

9 Memory tuning is a new method whereby LLMs augmented with a massive Mixture of Memory Experts (MoME) can easily memorize large datasets and increase factual accuracy and reliability. Banishing LLM Hallucinations Requires Rethinking Generalization, https://github.com/lamini-ai/Lamini-Memory-Tuning/blob/main/research-paper.pdf

10 Turns out the EU AI Act has taken a similar approach. The EU AI Act categorizes AI systems into four risk levels: 1) unacceptable risk, 2) high risk, 3) limited risk, and 4) minimal or no risk. The AI Act focuses most of its regulatory efforts on the unacceptable and high-risk categories, aiming to prevent harm while still fostering innovation in AI. https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai.

11 ABA Model Rules 1.7, 2.1 and 5.4. Rule 11 also requires lawyers to aver to the truthfulness, non-frivolousness and reliability of pleadings filed in federal court.

12 Artificial Lawyer, LITIG Forms Legal Industry AI Benchmarking Initiative, https://www.artificiallawyer.com/2024/06/28/litig-forms-legal-industry-ai-benchmarking-

initiative

5 Likes · 2 Restacks

Comments

Write a comment...