Why RAG is the Backbone of Legal AI

By Laina Chan

Artificial intelligence has entered legal practice in ways that few of us imagined even five years ago. Many practitioners now experiment with ChatGPT, Gemini or similar platforms. These systems rely on **retrieval-augmented generation (RAG)**: instead of drawing only on what the model has been trained on, the tool searches external sources to prepare an answer.

In principle, RAG should reduce hallucinations by grounding outputs in real material. In practice, however, the benefit depends entirely on *what* is retrieved and *how* it is analysed. Not all RAG is equal.

1. The limits of generic RAG

Public AI tools retrieve widely from the internet. The difficulty is that much of what is returned is **secondary material**: blog posts, law firm newsletters, commentary, and opinion. While these may be useful for background, they do not meet the standards of legal authority.

Equally concerning is the absence of legal method. The reasoning processes embedded in generic LLMs are the same whether the query concerns a recipe, a holiday destination, or a claim in negligence. Without legal structure, the retrieval is undirected and the results are inconsistent.

This is why lawyers report seeing outputs that cite the wrong jurisdiction, omit controlling authorities, or fabricate sources altogether.

2. RAG with legal method

At MiAI Law, we also use RAG — but in a different way. Our platform does not send a language model into the entire internet. Instead, it retrieves only from a **curated database of primary materials**: legislation and judgments.

More importantly, the retrieval is not left to the language model's generic algorithms. We have embedded a series of **legal reasoning steps**, reduced into code, that govern how material is retrieved, filtered, and ranked.

For every query, MiAI generates a reasoning plan with multiple stages. At each stage, cases are retrieved, analysed, and sorted. Irrelevant authorities are discarded. What remains is then subjected to structured legal analysis before the report is generated.

The final output is not simply an answer but an **audit trail**: the questions the system asked itself, the material retrieved, the subset relied upon, and footnoted propositions with pinpoint citations and hyperlinks. In other words, **evidence-grade research.**

3. Why this matters

The difference is more than technical. Generic RAG may produce a plausible summary. RAG combined with legal reasoning produces something a lawyer can **defend**: transparent, verifiable, and jurisdiction-specific.

This is what distinguishes a research tool from a chatbot. One generates text that sounds right; the other delivers analysis that can be checked against the authorities themselves.

4. Beyond inputs: the role of method

It is sometimes said that the quality of an AI's output depends on the quality of its inputs. That is only partly true. Databases like LexisNexis and Thomson Reuters already contain world-class material. The issue is not the input but the **method**. Without legal methodology embedded into the RAG process, even the best database will not yield reliable answers.

5. Conclusion

RAG is indeed the backbone of legal AI, but it is not enough on its own. What matters is the integration of **structured legal reasoning** into the retrieval process. Only then can AI deliver research that is not just fast, but reliable, transparent, and fit for legal practice.

As practitioners, we must remember that AI does not reason. It predicts. Embedding legal method into that process is what turns plausible language into **legal reliability**.