MERLiN®

Five Questions to Ask if You Are Considering a Private Large Language Model (LLM) for Your Law Firm.

By John Tredennick and Dr. William Webber

Many law firms and legal departments are considering building their own AI language models instead of using commercial LLMs like OpenAI's GPT or Anthropic's Claude. At first glance, it sounds like a great idea. You can train the model on your own documents, customize it to generate responses with your firm's unique voice, and perhaps keep the data more secure.

While intuitively that seems appealing, it may not be as attractive or cost effective an option as you might think. Before investing time, money and energy in a private LLM, ask these five questions:

1. What are the training costs and efforts required?

According to McKinsey, initial setup and training costs can run to millions of dollars. This seems like a big number but it includes costs for data preparation and model development, model training, and integration with existing systems. You will also have to include staffing expenses–including the cost of at least one AI data scientist. A CIO and CTO technology guide to generative AI | McKinsey.

If you doubt that estimate, know that OpenAI recently unveiled a model-building service for a minimum of $2-3M. How Much Does It Cost to Train a Large Language Model? A Guide. You might also want to consider: Fast Company: The hidden costs of AI models. With a good advisor, you might find ways to reduce these costs but they will still add up. Model training is not a trivial undertaking. Make sure you know what to expect in costs to build your model before you begin.

2. How much are the ongoing expenses to run your own model?

Along with the upfront costs, there are significant recurring annual expenses to be considered. These include data center costs to run the model, model upkeep, and maintaining integration with existing systems.

Running a private LLM can be like operating a supercomputer. McKinsey estimates annual costs between $500,000 and $1 million for infrastructure and staffing. While your model might not cost this much to run and maintain, your team will be responsible for keeping this complex computing system running. Here is a hint: It is a lot more challenging than running your own email server (which we all abandoned in favor of the cloud years ago).

3. How will we keep the system current?

Lawyers are constantly generating new work product. If you want the model to incorporate new documents, you will have to retrain it. Retraining requires that you repeat step 1, you effectively have to start from scratch. Even quarterly updates will place a burden on the organization.

The most important thing to understand is this: There is no simple way to add new documents to an LLM. Once initial training is complete, the model is locked down–it can't learn anything new and won't remember anything shared through prompts. New documents will be invisible to it until they can be incorporated during a retraining.

4. Are you satisfied with a GPT 3.5 level LLM?

The primary alternatives to using a commercial LLM like GPT-4 or Claude Opus is to license one of the many open source models like Llama, Falcon and Mistral. *E.g.* 8 Top Open-Source Large Language Models For 2024. They are literally free and can be downloaded in minutes from the web. Cloud providers like AWS and Google will host many of these models for you but you will have to pay computing costs on a 24/7 basis to keep them running.

The rub is this. At least to date, these models can't compare in analysis and output to the top commercial models like GPT-4.0 and Claude Opus. Rather, the best compare themselves with OpenAI's GPT 3.5, which has fallen behind the newer LLMs.[1] Are you satisfied with running a system that most of us have shelved in favor of GPT-4 (which has been upgraded several times already) or the recently released Claude 3? Can the firm accept a second rate AI engine?

5. How do you plan to keep up as newer models are released?

Even if you were satisfied with a GPT 3.5 level of AI, that isn't the end of the story. Large commercial providers like OpenAI, Anthropic and Google are constantly releasing new models. Just in the weeks before this article was written, Anthropic released three versions of Claude 3 that rendered Claude 2 obsolete. Open AI released the stable version of GPT-4 Turbo that tossed GPT-4 16 K and GPT-4 32 K to the curb, let alone the GPT 3.5 models.

Guess what. Those vendors aren't stopping with the latest releases. By the time you install, train and release your LLM, we may be on GPT 5 or an even newer version of Claude. How will you feel at that point when you announce your new GPT 3.5 like model?

To be sure, open source vendors also upgrade their models in an effort to keep up. When an open source vendor releases a new version of their model (e.g., moving from Llama 2 to Llama 3), they have usually made substantial improvements and changes to the model's parameters, structure and training process. As a result, you can't simply "upgrade" your existing private LLM to the new version without retraining it on your data.

A Practical Alternative: Retrieval-Augmented Generation (RAG):

Fortunately, there is a practical and less expensive alternative. A Retrieval-Augmented Generation (RAG) system leverages powerful commercial LLMs like GPT-4 or Claude in combination with an AI-powered search of your firm's documents. This approach can provide tailored insights at a fraction of the cost and complexity of a private LLM.

How does a RAG system work?

A Rag system combines three elements: (1) users; (2) a search platform; and (3) a connected LLM. Let's walk through these three requirements.

> **1. Users:** You already have the users. They are the people in your organization who need access to your private information. They connect using a browser to your private network.

[1] Llama 3, which is just in the process of being released as of this writing, claims that its larger (and most expensive model) can match Anthropic's mid-range model, Claude 3 Sonnet.

2. Search: The search engine is a bit more complicated–but not in comparison to building and training your own LLM. You can license a search engine or connect an open source search engine like ElasticSearch to your file system. New documents can quickly be added to the index so they will be found when you search.

3. LLM: Last, you need access to an LLM. This is done through an API, which functions like a commercial gateway to GPT, Claude or other LLMs. Communications to and from the LLM can be encrypted. The provider will contractually promise not to read, share or save the communications sent to the LLM (which can't store that information either).

Does the latter requirement sound risky? It shouldn't. Most organizations have been using Office 365 or Google Docs for years, sending information to the cloud through similar methods. All rely on similar contractual promises.

Here is a simple view of the architecture to get you started.

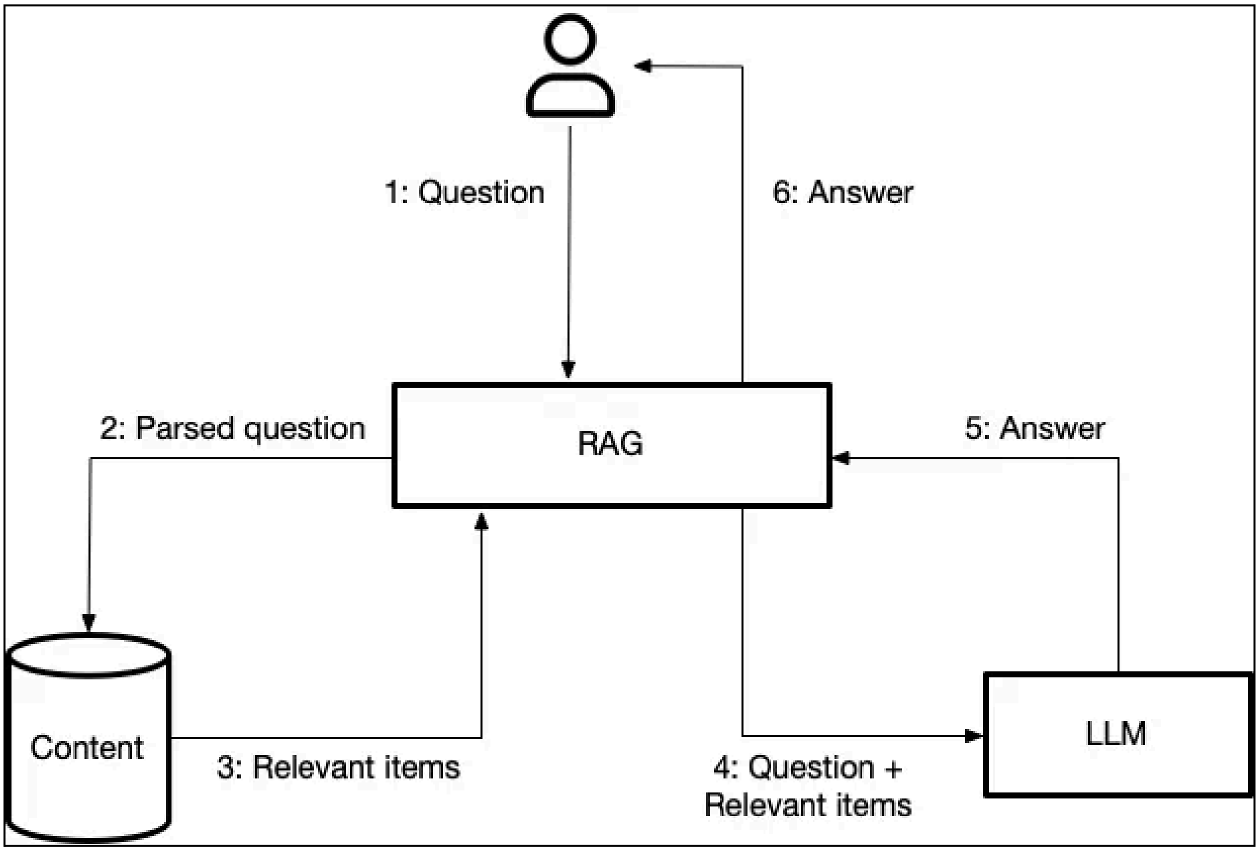

Illustration from: Question Answering through Retrieval Augmented Generation

Once you combine these elements the process is straightforward. A user asks a question, which prompts the system to run a search to find the most relevant (often recent) documents that might help answer the question.

The system retrieves those candidates and submits them to the LLM along with the initial question. The LLM reads the documents and then attempts to answer the question. From there, the user can ask further questions or broaden the inquiry by finding and submitting other documents to provide additional information.

The importance of a RAG system, even with a Private LLM.

Ironically, even if you decide to invest in a private LLM, you'll still need a RAG system to keep your firm's documents up to date. Retraining a private LLM weekly or even monthly to incorporate new documents is simply not practical.

In contrast, a RAG system can easily index new documents as they are created, making them immediately available for search and analysis. This means that regardless of whether you choose a private LLM or a commercial one, a RAG system is an essential component of any AI-powered document management strategy for your law firm. Why pay extra for the private LLM when the costs to access a system like GPT or Claude is so much less (and the commercial systems are so much smarter and more powerful)?

Is it really this easy?

No, of course not. But implementing a RAG system is a thousand times easier than managing, training and maintaining a private LLM. While I am not in that market, I feel confident you can find developers to build such a system for a fraction of the time and cost required to implement a private LLM. Going further, I am equally confident that you can find systems that are already built and just need configuration adjustments to adapt to your environment. Again, there may be work involved and certainly costs, but we are talking about a fraction of the hassle and costs of a private LLM.

Are there times a private LLM is the right choice?

Maybe. Without doubt one can conjure up times where a private LLM is the right choice. And there are plenty of vendors who will tell you exactly that. If you have a specialty practice like immigration, a private LLM may be a worthwhile investment.

Caveat Emptor (Buyer Beware)

While the idea of a private LLM may be appealing, it's important to consider the real world costs and challenges involved. For many law firms, a RAG system that leverages commercial LLMs and AI-powered document search will be a more practical and cost-effective solution. As with any major technology decision, it's crucial to thoroughly evaluate your needs, budget, and capabilities before deciding whether to invest in a private LLM or opt for an alternative approach.

About the Authors

John Tredennick (JT@Merlin.Tech) is the CEO and founder of Merlin Search Technologies, a software company leveraging generative AI and cloud technologies to make investigation and discovery workflow faster, easier, and less expensive. Prior to founding Merlin, Tredennick had a distinguished career as a trial lawyer and litigation partner at a national law firm.

With his expertise in legal technology, he founded Catalyst in 2000, an international ediscovery technology company that was acquired in 2019 by a large public company. Tredennick regularly speaks and writes on legal technology and AI topics, and has authored eight books and dozens of articles. He has also served as Chair of the ABA's Law Practice Management Section.

Dr. William Webber (wwebber@Merlin.Tech) is the Chief Data Scientist of Merlin Search Technologies. He completed his PhD in Measurement in Information Retrieval Evaluation at the University of Melbourne under Professors Alistair Moffat and Justin Zobel, and his post-doctoral research at the E-Discovery Lab of the University of Maryland under Professor Doug Oard.

With over 30 peer-reviewed scientific publications in the areas of information retrieval, statistical evaluation, and machine learning, he is a world expert in AI and statistical measurement for information retrieval and ediscovery. He has almost a decade of industry experience as a consulting data scientist to ediscovery software vendors, service providers, and law firms.

About Merlin Search Technologies

Merlin is a pioneering cloud technology company leveraging generative AI and cloud technologies to re-engineer legal investigation and discovery workflows. Our next generation platform integrates GenAI and machine learning to make the process faster, easier, and less expensive. We've also introduced Cloud Utility Pricing, an innovative software hosting model that charges by the hour instead of by the month, saving clients substantial savings on discovery costs when they turn off their sites.

With over twenty years of experience, our team has built and hosted discovery platforms for many of the largest corporations and law firms in the world. Learn more at merlin.tech.