RAG in Depth

RAG vs. Large Context LLM Models

The rise of the Large Context LLM Models.

- February 2024: Gemini 1.5. Pro. Up to 1M tokens.
- The entire Harry Potter series in the context window.
- Is RAG dead?

RAG vs. Large Context LLM Models

- Efficiency and Use Cases.
- Cost and Efficiency Concerns.
- Simplicity and Explainability.
- Control over Information.
- Large Memory Sizes and Easy Updates.
- Challenges with Latency and Context.

Conclusion

- RAG still holds significant value.
- RAG is still indispensable for certain applications.
- Both technologies will coexist, complementing each other to cover a broader range of AI applications and use cases.