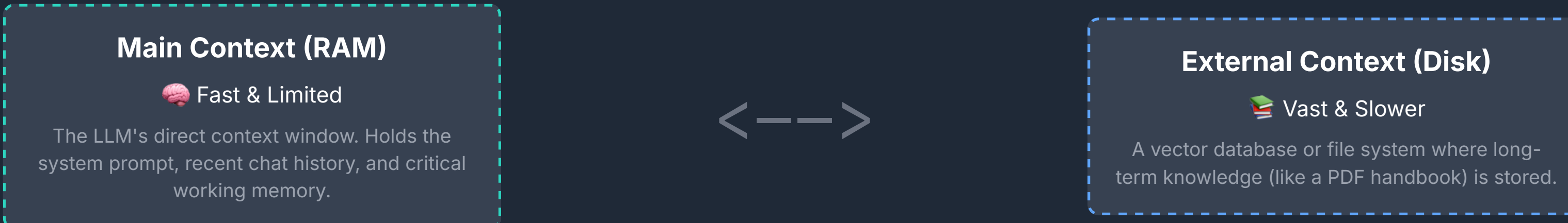


From LLMs to Operating Systems

Inspired by the [MemGPT paper](#) and the [DeepLearning.AI course](#), these notebooks explore how to overcome the memory limitations of LLMs by treating them like the CPU of a traditional operating system

Concept 1: The Memory Hierarchy

Just like a computer has fast RAM and slow disk storage, an agent has a tiered memory system. The key is to intelligently manage what information is loaded into the LLM's limited "working memory" (the context window).



Concept 2: Self-Editing Memory

The core idea of MemGPT: give the LLM the tools to manage its own memory. Instead of us deciding what's important, the agent learns to save critical information itself.

Concept 3: Agentic RAG

This memory system enables a powerful, agent-driven RAG process. The agent actively decides when it needs more information and goes to get it from its external memory.

Conclusion

By giving LLMs tools to manage a memory hierarchy, we transform them from simple chatbots into more capable, autonomous agents that can learn, reason, and solve complex, multi-step problems. This is the core idea of treating LLMs like an **operating system** for intelligence.