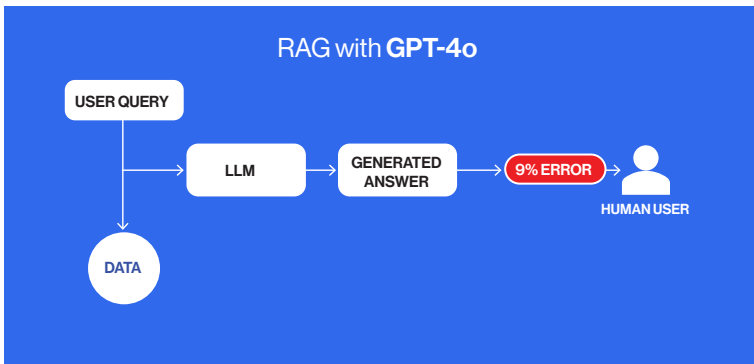


RAG-Verification

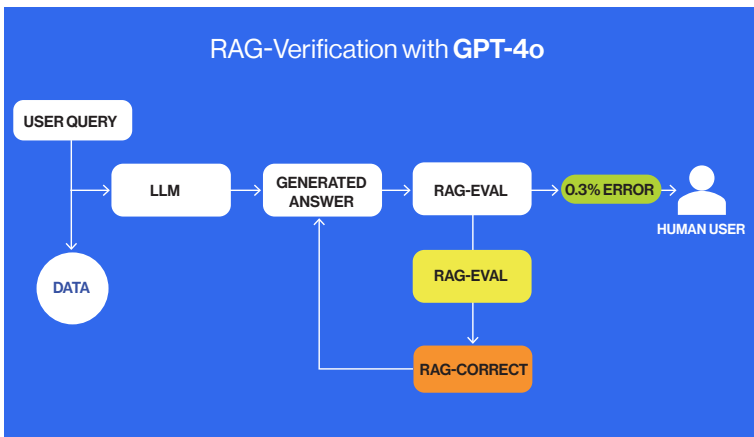
Primer's solution to LLM hallucinations

Understanding the challenge: Why RAG falls short



Large language models (LLMs) generate erroneous information, called hallucinations. The solution to the problem was supposed to be retrieval-augmented generation (RAG) where the LLM grounds its generated answers in retrieved data. But even with RAG, the state of the art LLMs have an error rate of 5% to 10%. A knowledge system that makes an error every 10 times you use it is not reliable enough to use, especially in high-stakes use cases such as defense and intelligence.

Introducing RAG-Verification: The fact-checking AI framework



To bring hallucinations under control, Primer created RAG-Verification (RAG-V), a framework that automatically corrects LLM errors before they reach human users. The approach is similar to journalism fact-checking. It detects and corrects errors by comparing each factual claim in an LLM-generated text to its cited data sources. Because this fact-checking process can be parallelized, it runs fast enough to be used in a real-time interactive AI system. RAG-V reduces the state of the art hallucination rate by as much as 100X, making AI systems reliable enough for mission-critical use.

About Primer

Primer provides trusted decision-ready AI to the world's most critical organizations. Our software is deployed by the U.S. government, strategic allies, and Fortune 100 companies to extract timely insight and decision advantage from massive datasets. Primer has offices in Arlington, Virginia, and in Pasadena and San Francisco, California. For more information, please visit www.primer.ai.