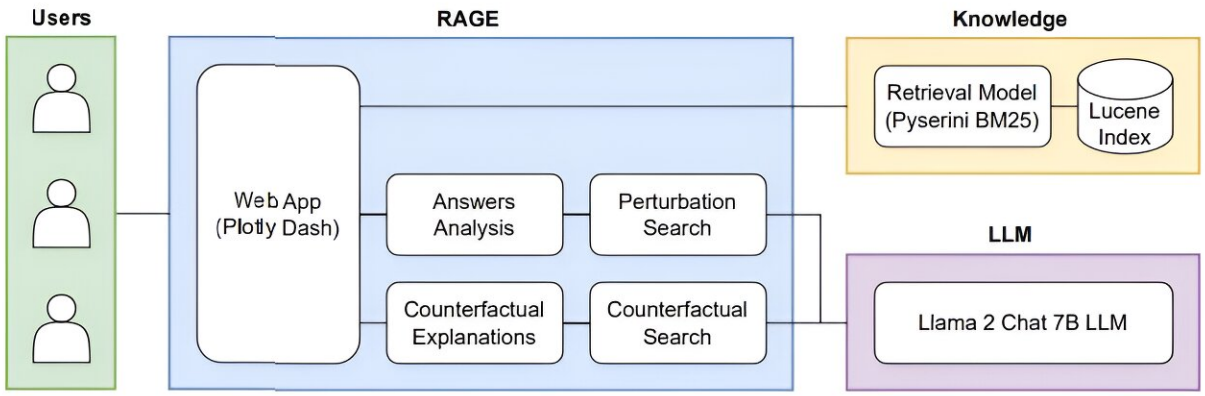
Know your source: RAGE tool unveils ChatGPT's sources

June 4 2024



The architecture of RAGE. Credit: *arXiv* (2024). DOI: 10.48550/arxiv.2405.13000

A team of researchers based at the University of Waterloo have created a new tool—nicknamed "RAGE"—that reveals where large language models (LLMs) like ChatGPT are getting their information and whether that information can be trusted.

LLMs like ChatGPT rely on "unsupervised deep learning," making connections and absorbing information from across the internet in ways that can be difficult for their programmers and users to decipher. Furthermore, LLMs are prone to "hallucination"—that is, they write convincingly about concepts and sources that are either incorrect or

nonexistent.

"You can't necessarily trust an LLM to explain itself," said Joel Rorseth, a Waterloo computer science Ph.D. student and lead author on the study. "It might provide explanations or citations that it has also made up."

Rorseth's team's new tool employs a recently developed strategy, called "retrieval-augmented generation" (RAG), to understand the context for LLMs' answers to a given prompt.

"RAG allows users to provide their own sources to an LLM for context. Our tool illustrates how different sources lead to different answers when using RAG, helping to assess whether that information is trustworthy," Rorseth said.

Because their tool focuses on retrieval-augmented generation explainability, they nicknamed it "'RAGE' against the machine."

Understanding where LLMs like ChatGPT are getting their information—and ensuring they're not repeating misinformation—will only become more important as highly sensitive, human-centered industries like the medical and legal sectors adopt these tools, Rorseth said.

"We're in a place right now where innovation has outpaced regulation," he said. "People are using these technologies without understanding their potential risks, so we need to make sure these products are safe, trustworthy, and reliable."

The research, "RAGE Against the Machine: Retrieval-Augmented LLM Explanations," will be published in the Proceedings of the 40th IEEE International Conference on Data Engineering. It is currently available on the *arXiv* preprint server.

More information: Joel Rorseth et al, RAGE Against the Machine: Retrieval-Augmented LLM Explanations, *arXiv* (2024). DOI: 10.48550/arxiv.2405.13000

Provided by University of Waterloo