Sample Conversation 1

Actual article link for reference.

Chatbot conversation:

You: what is rag?

Bot: In the context provided, "RAG" stands for "Retrieval Augmented Generation," which is a natural language processing (NLP) technique that combines the strengths of both retrieval- and generative-based AI models. It delivers accurate results based on pre-existing knowledge while also being able to process and consolidate that knowledge to create unique, context-aware answers, instructions, or explanations in human-like language. RAG is a superset of generative AI as it combines the strengths of both generative and retrieval AI. It's different from cognitive AI which mimics the way the human brain works to get results.

You: what are its benefits?

Bot: Retrieval Augmented Generation (RAG) is a natural language processing (NLP) technique that combines the strengths of both retrieval- and generative-based AI models. The main advantages of RAG are:

1. **Delivering accurate results:** RAG can make the most of pre-existing knowledge to deliver precise answers or instructions.
2. **Creating context-aware responses:** By processing and consolidating the retrieved knowledge, RAG can generate unique, context-aware answers in human-like language.
3. **Being a superset of generative AI:** RAG is an extension of generative AI as it uses the capability to create new content but also leverages the strengths of retrieval AI.
4. **Enhancing efficiency and innovation:** By automating the process of extracting insights from vast amounts of data, RAG can help organizations unleash new levels of efficiency, innovation, and growth.
5. **Mimicking human-like language:** RAG generates responses in a conversational manner that is easy for users to understand, making interactions with AI systems more natural and intuitive.

You: How is it used today?

Bot: Retrieval Augmented Generation (RAG) is a Natural Language Processing (NLP) technique that combines the strengths of both retrieval- and generative-based artificial intelligence models. It can deliver accurate results based on pre-existing knowledge, while also generating unique, context-aware answers, instructions, or explanations in human-like language. RAG is different from generative AI in that it is a superset of generative AI, as it combines the strengths of both generative and retrieval AI.

RAG works by first retrieving relevant information based on user queries or input, and then generating responses that are aware of the context of the conversation. This approach makes RAG more efficient than large-scale generative models, as the initial retrieval phase narrows down the context and thus the volume of data that needs to be processed in the generation phase. Additionally, RAG models require less training data compared to building a generative AI model from scratch.

In practice, RAG is used in applications such as chatbots, customer support systems, and information retrieval systems where accurate and context-aware responses are required. For example, Cohesity's RAG platform under development is designed to accept both human and machine-driven input, tokenize the input with keywords that quickly filter large enterprise backup data down to a smaller subset of relevant data, select representations from within those documents or objects that are most relevant to the query or question, package the result with the original query to an LLM like GPT4 for generation of human-sounding answers. This approach ensures that generated responses are not only knowledgeable and up-to-date but also diverse and relevant to specific business content.

You: /bye
Exiting chat mode...