Lesson Downloads

Show TranscriptSummarize Video

Retrieval-Augmented Generation (RAG) is a powerful approach for keeping Generative AI models informed with the most recent data, particularly when dealing with domain-specific questions. It cleverly combines the comprehensive understanding capacity of a large language model (LLM) with the most up-to-date information pulled from a database of relevant text snippets. The beauty of this system is in its ability to ensure that responses remain accurate and reflective of the latest developments.

Technical Terms:

Semantic-embedding: A representation of text in a high-dimensional space where distances between points correspond to semantic similarity. Phrases with similar meanings are closer together.
Cosine similarity: A metric used to measure how similar two vectors are, typically used in the context of semantic embeddings to assess similarity of meanings.
Vector databases: Specialized databases designed to store and handle vector data, often employed for facilitating fast and efficient similarity searches.

Quiz Question
What role does semantic embedding play in RAG?

- It serves as a security feature to protect the model
- It helps in retrieving relevant text snippets based on similarity of meaning.
- It translates questions into different languages.

Submit