PDF 2: Definition of RAG

Retrieval-Augmented Generation, usually called RAG, is a pattern for building AI systems that answer questions with help from an external knowledge base. Instead of relying only on what a model learned during training, a RAG system first retrieves relevant information from your documents or databases, then uses that retrieved material to generate the final answer.

A simple RAG loop has four steps. First, you ingest content and create an index, often using embeddings in a vector store, sometimes combined with keyword search. Second, the user asks a question. Third, the system retrieves the best matching passages. Finally, the model generates a response conditioned on those passages.

RAG is useful because it improves accuracy, keeps answers up to date as your data changes, and provides a path to citations and auditing. It can also help enforce access control by retrieving only what the user is allowed to see. Common uses include internal knowledge assistants, customer support bots, research tools, and document analysis workflows. In evaluation, teams measure retrieval quality, answer faithfulness, and latency to confirm the system stays reliable at scale. Good systems add guardrails such as citation requirements, answer refusal when evidence is missing, and clear logging for troubleshooting.