RAG Best Practices

Retrieval-Augmented Generation improves LLM accuracy by injecting external knowledge into prompts. Documents should be chunked properly and embeddings should be regenerated whenever documents change.

For production systems, persist vector indexes and use Redis or databases for session storage.