

Retrieval-Augmented Generation Architecture

1. User Query

Question or prompt

2. Vector Database

Retrieve relevant docs

3. Retrieved Docs

Top-k documents

4. LLM Generation

Query + Context

5. Response

Answer with citations

Benefits

- Access to current and external knowledge
- Reduced hallucination (grounded in sources)
- Source attribution and citations
- No model retraining required
- Update knowledge by updating database
- Enables smaller LLMs with external knowledge

Cost Considerations

- Retrieval latency: +20-50ms per query
- Vector DB costs: \$0.0001-0.001 per query
- Embedding computation for queries
- Storage costs for vector database
- Net cost often neutral vs larger models
- Enables use of smaller, cheaper LLMs