

◆ BLUEPRINT · RETRIEVAL-AUGMENTED GENERATION

RAG that survives production.

The reference architecture behind a grounded, accurate assistant — ingestion, hybrid retrieval, reranking, evaluation and permission inheritance.

Ingestion

Embeddings

Hybrid retrieval

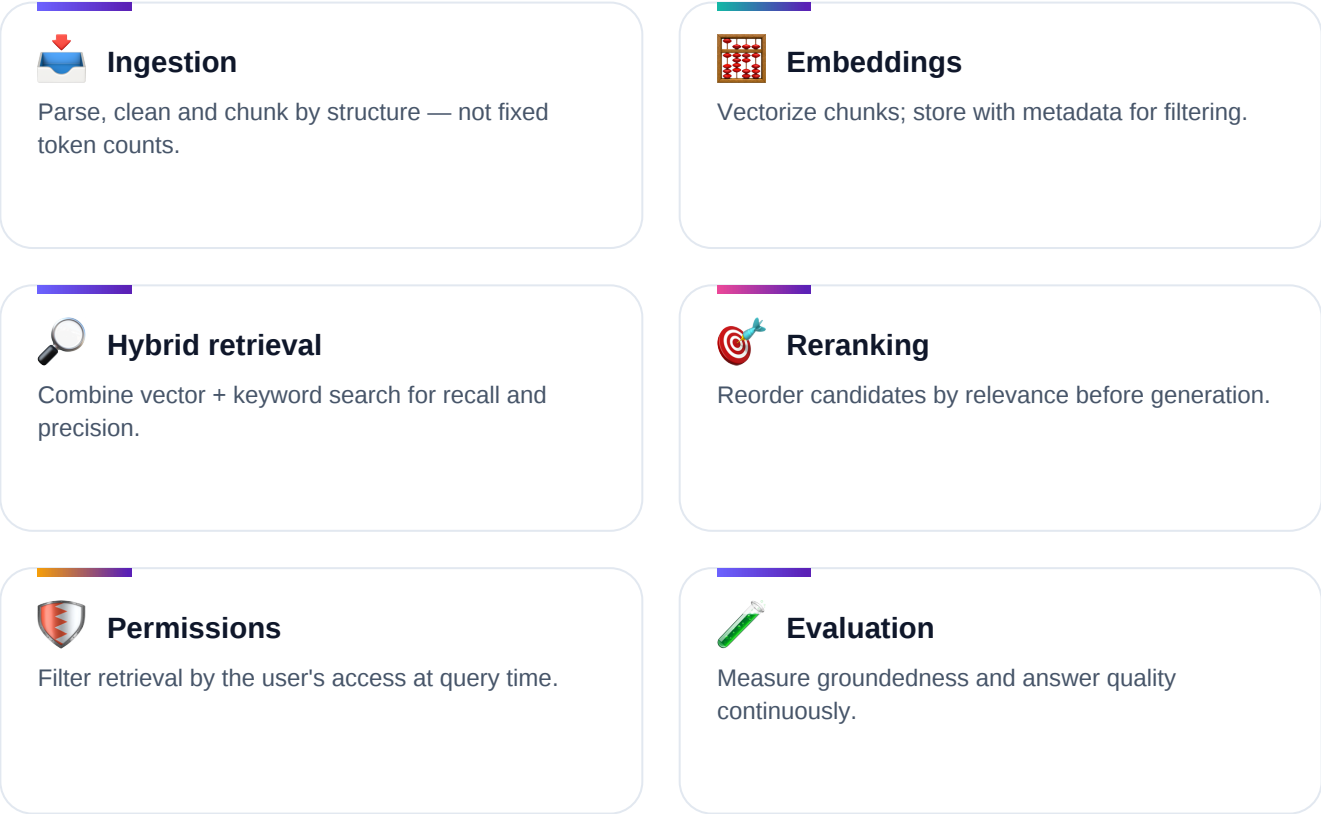
Reranking

Evals

The pipeline

From document to grounded answer.

A demo retrieves once and hopes. Production RAG layers retrieval, reranking and evaluation — and inherits the user's permissions at every step.



What separates demo from production.

- ✓ Chunk by document structure; preserve headings and context.
- ✓ Always combine semantic and keyword retrieval (hybrid).
- ✓ Add a reranking stage — it's the biggest quality lever.
- ✓ Inherit user permissions into the retrieval filter.
- ✓ Cite sources so answers are verifiable.
- ✓ Evaluate groundedness on a fixed test set before every change.