RAG corporativo: conocimiento privado con citas

Recuperar primero, redactar después.

Conceptos clave

- RAG separa conocimiento *dinámico* (documentos) del modelo *estático*. Documento nuevo ⇒ se indexa, no se reentrena.
- Pipeline canónico: ingesta → **chunking** → **embeddings** → vector DB → **retrieval** → LLM con cita.
- Patrones que suman: *hybrid search* (BM25 + dense), reranking (cross-encoder), query rewriting.
- Métricas: Recall@k, MRR, **faithfulness** (clave en sectores regulados).

Ejemplo a mano

Dimensiones reales de un RAG en BBVA.

Documentos	120 000 PDFs (circulares + contratos modelo).
Chunks tras chunking	1,45 M (700 tokens/chunk).
Dimensión embedding	1024 (E5-multilingual).
Tamaño Pinecone	11 GB en producción.
Ingesta inicial	890 $ (re-procesar embeddings).
Coste mensual operativo	3 200 $ (Pinecone) + 2 150 $ (Sonnet).
Faithfulness antes / después	0,57 → **0,93**.

Concepto → aplicación: cinco bases vectoriales

Vector DB	Caso	Coste / Operación
Pinecone	RAG empresarial gestionado.	Cloud, 100–3 000 $/mes.
Qdrant	RAG self-hosted en producción.	Servidor propio, Apache 2.0.
Chroma	RAG dev y prototipos.	Local, en memoria, gratis.
Weaviate	RAG híbrido texto + vector.	Cloud o self-hosted.
pgvector	RAG pequeño dentro de Postgres.	Reusa la BBDD existente.

Idea clave del Día 15. Para conocimiento corporativo cambiante, **RAG es la respuesta correcta el 90 % de las veces**. Empieza con Chroma local; pasa a Qdrant o Pinecone cuando crezca el corpus. Hybrid search + reranking + query rewriting suben Recall@5 de 0,65 a 0,90 sin tocar el LLM.

Para llevarte: lectura Lewis et al., *Retrieval-Augmented Generation*; práctica: RAG mínimo sobre 30 PDFs propios.