# Ubicación del Capítulo dentro de la Memoria

Se recomienda situar este capítulo tras la descripción de la arquitectura general (ingesta, almacenamiento, UI) y antes del capítulo de evaluación end-to-end. Así, el lector entiende primero el pipeline, a continuación el subsistema de recuperación semántica (vector stores) y finalmente su impacto en el RAG.

# Almacenamiento Vectorial y Comparativa FAISS vs ChromaDB — Ampliación

Este anexo amplía la memoria con los experimentos realizados sobre la colección 'onda\_docs' (source\_id=1), comparando FAISS y ChromaDB con dos modelos de embedding: all-MiniLM-L6-v2 (384d) e intfloat/multilingual-e5-base (768d). Se incluyen resultados reales, tablas y figuras, así como recomendaciones para ampliaciones (bge-m3 y overlap top-k).

## Metodología resumida

Los índices se construyen con scripts propios (index\_chunks.py) generando artefactos index\_meta.json e index\_manifest.json para trazabilidad. La evaluación emplea scripts de comparativa y diagnóstico que calculan doc@k, chunk@k, text@k y latencias (p50/p95/mean) por store y k.

## Construcción de Índices — E5

|  |  |  |  |
| --- | --- | --- | --- |
| Vector Store | Tiempo Construcción (s) | Dimensión | N\_Chunks |
| FAISS | 3844.190 | 768 | 29834 |
| ChromaDB | 3901.709 | 768 | 29834 |

## Resultados — all-MiniLM-L6-v2

Con MiniLM, doc@k alcanza 60% al subir k a 25–30; chunk@k≈0%. La mediana de latencia se mantiene baja (15–21 ms), con mayor variabilidad en Chroma p95.

Figura 1. doc@k por store (MiniLM).

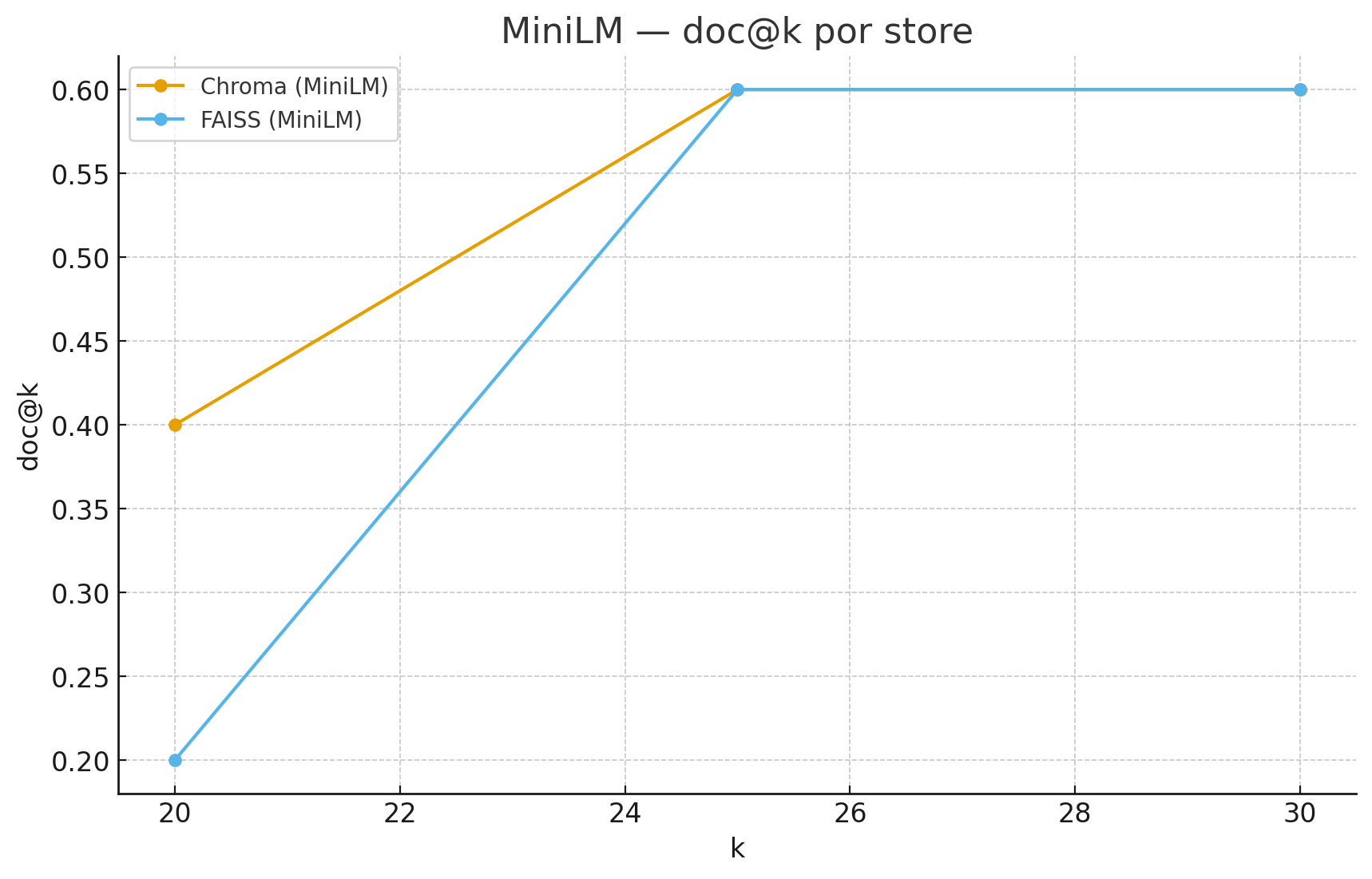


Tabla 1. MiniLM — ChromaDB por k (valores reales).

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| k | doc@k | chunk@k | text@k | p50 (ms) | p95 (ms) |
| 20 | 0.40 | 0.00 | 0.091 | 18.5 | 122.6 |
| 25 | 0.60 | 0.00 | 0.091 | 15.9 | 135.1 |
| 30 | 0.60 | 0.00 | 0.364 | 21.1 | 148.4 |

Tabla 2. MiniLM — FAISS por k (valores reales).

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| k | doc@k | chunk@k | text@k | p50 (ms) | p95 (ms) |
| 20 | 0.20 | 0.00 | 0.091 | 15.9 | 50.0 |
| 25 | 0.60 | 0.00 | 0.091 | 17.9 | 28.7 |
| 30 | 0.60 | 0.00 | 0.091 | 15.5 | 36.0 |

## Resultados — intfloat/multilingual-e5-base

Con E5, doc@k=40% en k=20 y 60% en k=40 en ambos stores; chunk@k=20%. El p50 se sitúa en torno a 40–46 ms y el p95 se estabiliza tras el warm-up.

Figura 2. doc@k por store (E5).

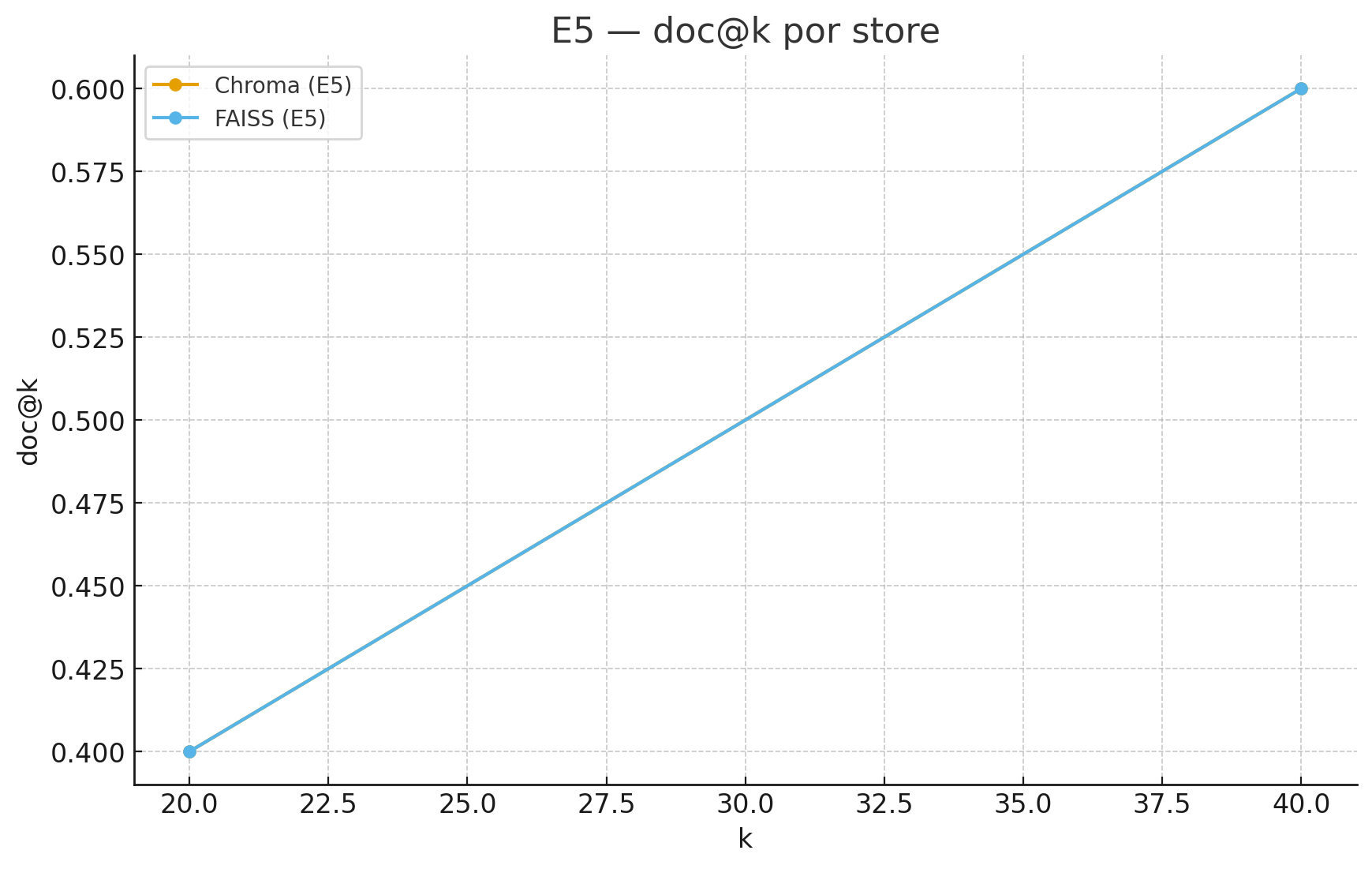


Figura 3. Latencia p50 por k (E5).

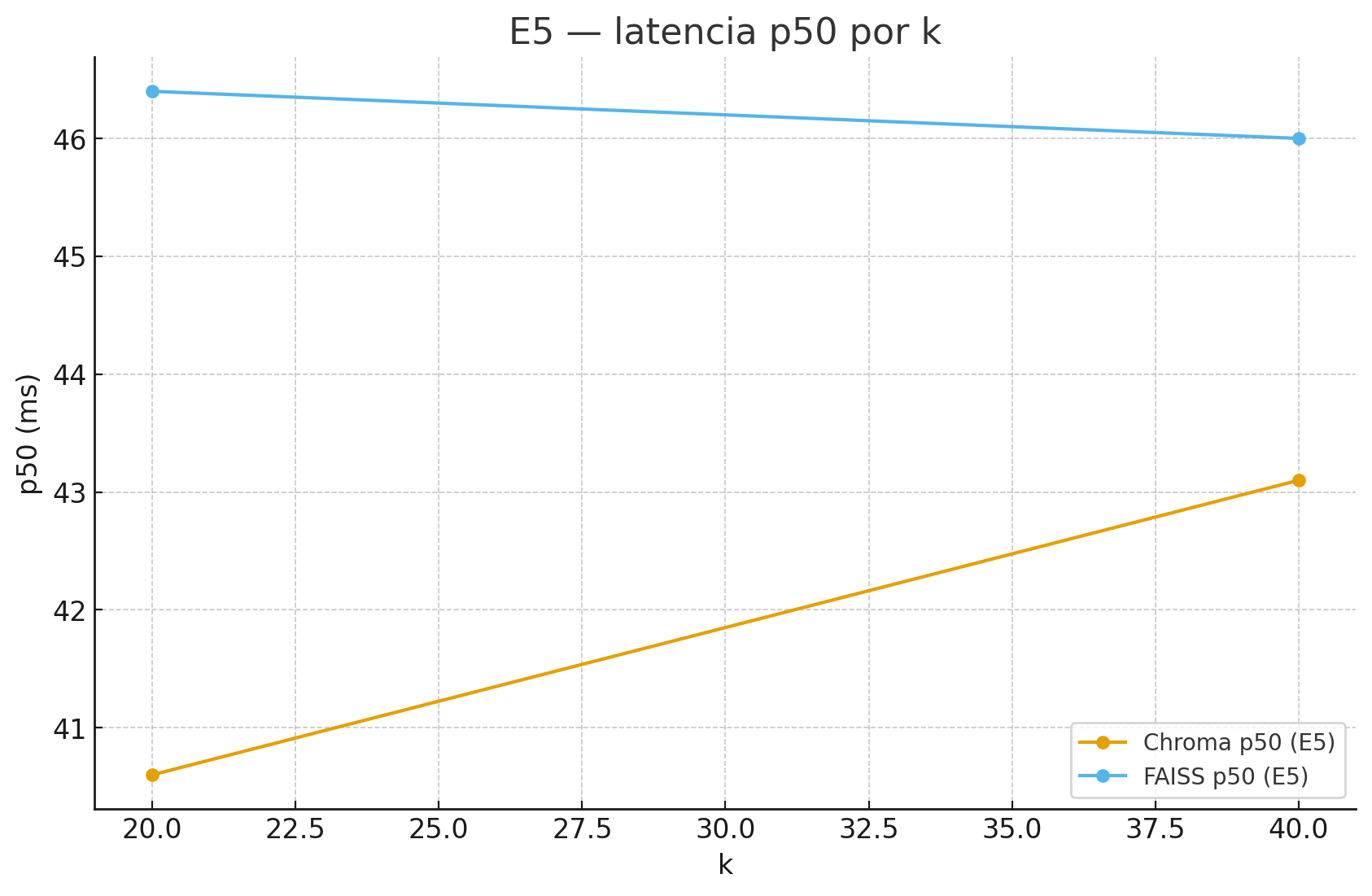


Tabla 3. E5 — ChromaDB por k (valores reales).

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| k | doc@k | chunk@k | text@k | p50 (ms) | p95 (ms) |
| 20 | 0.40 | 0.20 | 0.091 | 40.6 | 365.0 |
| 40 | 0.60 | 0.20 | 0.182 | 43.1 | 109.1 |

Tabla 4. E5 — FAISS por k (valores reales).

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| k | doc@k | chunk@k | text@k | p50 (ms) | p95 (ms) |
| 20 | 0.40 | 0.20 | 0.091 | 46.4 | 121.7 |
| 40 | 0.60 | 0.20 | 0.182 | 46.0 | 95.7 |

## Diagnóstico por consulta — E5 (docid\_presence)

k=20; probe\_k=200. “gap” = rank\_probe - k cuando está fuera de k.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| idx | docid | query | FAISS rank@k | FAISS rank@probe | gap | Chroma rank@k | Chroma rank@probe | gap |
| 1 | 390 | empadronamiento | 4 | 4 | - | 3 | 3 | - |
| 2 | 390 | licencia de obra menor | 1 | 1 | - | 1 | 1 | - |
| 6 | 118 | plazos de presentación | - | - | - | - | - | - |
| 8 | 51 | tasas municipales | - | 39 | 19 | - | 37 | 17 |
| 11 | 3 | presentación telemática | - | - | - | - | - | - |

# Recomendaciones de ampliación

## BAAI/bge-m3 (propuesta de evaluación)

Se recomienda evaluar bge-m3 como embedding multilingüe de alta calidad. Impacto esperado: mejora de doc@k en consultas en español y términos administrativos; coste: mayor tamaño de índice y tiempo de codificación. Mantener la misma colección y comandos de reindexación que con E5, cambiando únicamente el nombre del modelo en index\_chunks.py.

## Overlap top-k FAISS↔Chroma (propuesta)

Para cuantificar la divergencia entre recuperadores se propone calcular, por consulta, la similitud de Jaccard entre los conjuntos de chunk\_id recuperados en top-k por FAISS y por Chroma, y reportar media/mediana por k. Esto complementa doc@k al mostrar estabilidad entre almacenes.

# Limitaciones del ciclo evaluado (precisión)

Limitaciones del ciclo evaluado. Las pruebas se realizaron sobre el snapshot onda\_docs (corpus municipal) con 11 consultas de validación. En este ciclo no se exploraron los hiperparámetros de HNSW (p. ej., ef\_search) y no se midió el overlap top-k entre FAISS y ChromaDB; ambos quedan como trabajo futuro.