

L U M B R E



# Recuperación híbrida y reranking

**Grado:** 8

**Materia:** Sistemas de Recuperación Aumentada · IA

11/10/2026

## Antes de empezar

Cuando el RAG falla suele ser recuperación: combinar búsqueda vectorial y léxica con fusión de rankings y reordenar candidatos con cross-encoders.

### QUÉ APRENDERÁS

- Aplicar búsqueda híbrida (densa y léxica) y reranking con cross-encoder para mejorar la recuperación de un sistema RAG.

## Recuperación híbrida y reranking

VIDEO

### Búsqueda híbrida y reranking en RAG

Video explicativo (inglés): guía completa de la búsqueda híbrida BM25 + embeddings + reranker en pipelines de RAG.



Escanea para verlo en Lumbre

<https://www.youtube.com/watch?v=XvKiTfd6Xvo>

### Encontrar no es lo mismo que recuperar algo

El chunking y el vector hacen que la recuperación exista. El buscador híbrido y el reranker hacen que encuentre el fragmento correcto.

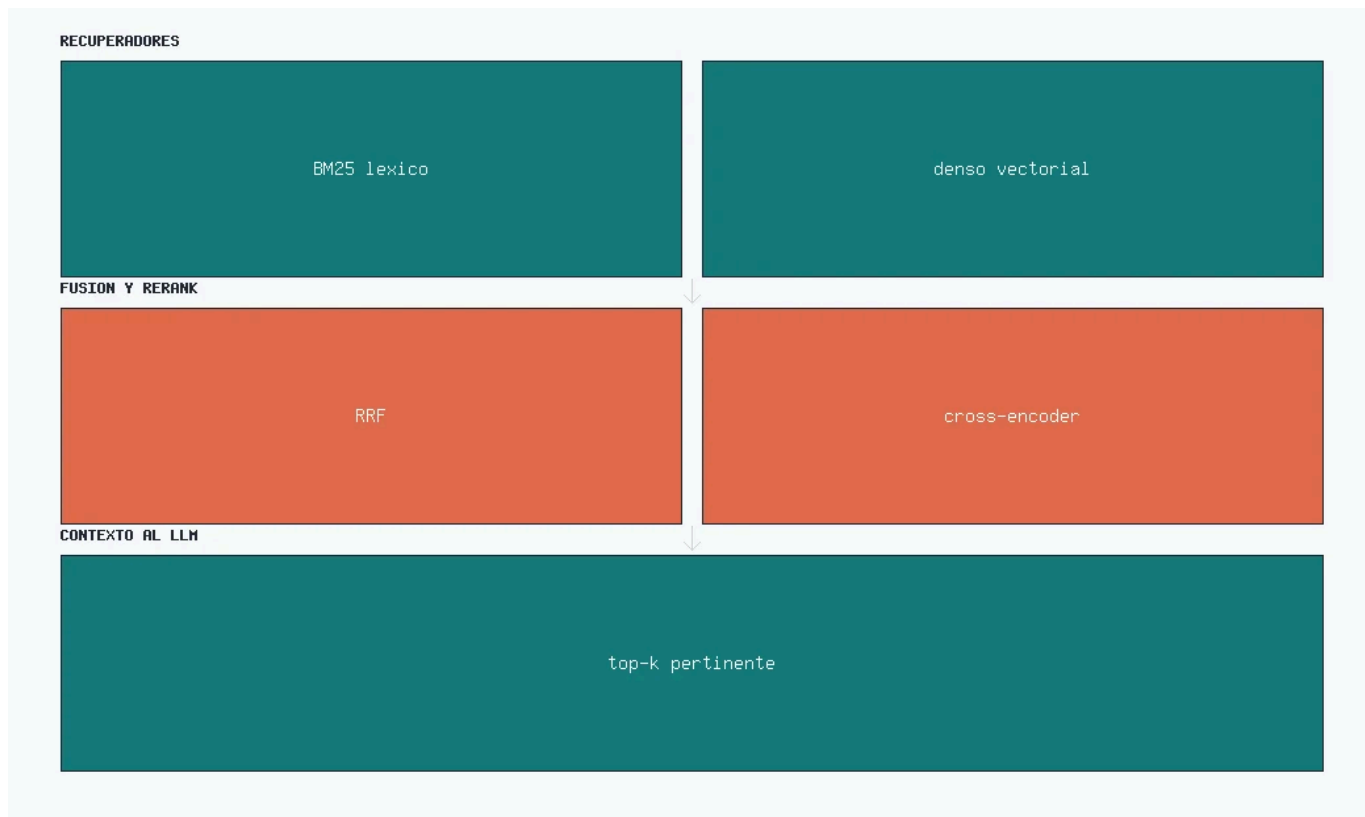
La calidad de un RAG depende sobre todo de la **recuperación**: si no traes el fragmento pertinente, el modelo no puede usarlo. La búsqueda puramente vectorial es buen punto de partida, pero falla con nombres exactos, códigos y términos raros. Combinar recuperación densa y léxica, y reordenar después, cierra esa brecha.

## Densa y léxica: dos tipos de falla

- Densa (embeddings): captura semántica y sinónimos, pero difumina términos exactos y tokens infrecuentes.
- Léxica (BM25): acierta con palabras exactas y códigos, pero ignora parafraseo y sinónimos.

## Búsqueda híbrida

Un buscador **híbrido** ejecuta ambos recuperadores y fusiona sus rankings, por ejemplo con *reciprocal rank fusion*. Lo que cada método recupera bien sube, y la debilidad de uno la compensa el otro.



*El híbrido amplía el recall; el reranking recupera precisión sobre esos candidatos.*

## Reranking con cross-encoder

El índice vectorial usa un **bi-encoder** (consulta y documento por separado) porque es rápido. Un **cross-encoder** lee consulta y documento juntos y puntúa su relevancia con mucha más precisión, pero es caro: se usa solo sobre un top-k pequeño para **reordenar**.

## Reescritura de la consulta

- Multi-query: genera variantes de la pregunta para ampliar la recuperación.
- HyDE: escribe una respuesta hipotética y busca por similitud a ella.
- Expansión: añade términos que el usuario omitió.

*Cada pieza cubre el hueco de la anterior.*

| Método                 | Ventaja            | Debilidad                       |
|------------------------|--------------------|---------------------------------|
| Vectorial (bi-encoder) | rápido y semántico | términos exactos y tokens raros |

| Método                 | Ventaja                | Debilidad                         |
|------------------------|------------------------|-----------------------------------|
| BM25 (léxico)          | términos exactos       | sinónimos y paráfrasis            |
| Híbrido + RRF          | mejor recall           | requiere afinar la fusión         |
| Cross-encoder (rerank) | precisión al reordenar | coste alto, solo sobre candidatos |

## Errores frecuentes

- Confiar solo en denso cuando el dominio usa códigos, modelos o leyes numeradas.
- Pedir top-k muy grande: diluye el contexto útil y encarece la generación.
- Reordenar con un cross-encoder mal calibrado que prioriza longitud.

## Ejemplo resuelto: el documento exacto no aparece

1. Confirmar que la consulta incluye un término exacto (un SKU o un número de norma).
2. Añadir BM25 al pipeline vectorial y fusionar con RRF.
3. Sobre el top-50 fusionado, aplicar un cross-encoder para quedarse en top-5.
4. Medir con el set de evaluación de RAG si la respuesta ahora tiene el fragmento correcto.

## ¿Para qué sirve en la realidad?

Buena parte de la "mala respuesta" de un RAG es en realidad mala recuperación. Híbrido + reranking es la palanca más rentable para subir la calidad sin cambiar de modelo ni de base vectorial.

### 1. ¿Por qué se combina BM25 con búsqueda vectorial en un sistema híbrido?

- ☐ Porque BM25 es más rápido
- ☐ Porque el vector falla en términos exactos y el léxico en semántica; se compensan
- ☐ Porque el híbrido elimina el reranking
- ☐ Para no tener que hacer chunking

### 2. Un cross-encoder se usa para reordenar un conjunto pequeño de candidatos porque es caro pero preciso.

- ☐ Verdadero
- ☐ Falso

### 3. La técnica de reescritura que genera una respuesta hipotética y busca documentos similares a ella se llama \_\_\_\_\_.

4. Clasifica cada técnica según si amplía candidatos o los reordena.

| Elementos:                                                                                                                                                                                                  |       |
|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------|
| <ul style="list-style-type: none"><li>• búsqueda vectorial densa</li><li>• cross-encoder reranker</li><li>• multi-query rewriting</li><li>• fusión híbrida RRF</li><li>• reordenar top-50 a top-5</li></ul> |       |
| Amplía candidatos                                                                                                                                                                                           | . . . |
| Reordena candidatos                                                                                                                                                                                         | . . . |

5. Une cada término con su definición.

|    |               |       |
|----|---------------|-------|
| 5. | BM25          | ----- |
| 5. | Bi-encoder    | ----- |
| 5. | Cross-encoder | ----- |
| 5. | RRF           | ----- |

**Opciones:** recuperación léxica por términos exactos · codifica consulta y doc por separado, rápido · lee consulta y doc juntos, preciso · fusión de rankings de varios recuperadores

ACTIVIDAD INTERACTIVA DISPONIBLE EN LUMBRE

6. Recuperación avanzada



Recuperación híbrida y reranking

- Recuperacion RAG
- Hibrida
- vectorial semantica
- BM25 lexica
- fusion RRF
- Reordenar
- cross-encoder caro
- top-k a top-5
- Query
- multi-query
- HyDE

El mapa mental interactivo es interactivo en la versión online.



Retrieval-Augmented Generation (Lewis et al., 2020) <https://arxiv.org/abs/2005.11401>  
Paper fundacional que acopla recuperador y generador.

Survey de RAG para grandes modelos de lenguaje <https://arxiv.org/abs/2309.10945>  
Panorama de estrategias de recuperación y reranking.

**7. Tu RAG legal falla en consultas que citan números de artículo exactos. ¿Qué pieza del pipeline (chunking, denso, léxico o reranker) te espera que más impacte, y cómo lo confirmarías con la evaluación de RAG?**

---

---

---

---

---

## Referencia curricular

|                 |                                         |
|-----------------|-----------------------------------------|
| <b>Grado:</b>   | 8                                       |
| <b>Materia:</b> | Sistemas de Recuperación Aumentada · IA |

Versión interactiva completa (con actividades, videos y autoevaluación):



<https://univ.lumbre.lat/object/cdia-rag-005-recuperacion-hibrida-y-reranking>

Material educativo generado desde Lumbre · [lumbre.lat](https://lumbre.lat)