

L U M B R E



Evaluación de RAG

Grado: 8

Materia: Sistemas de Recuperación Aumentada (RAG) · Inteligencia Artificial

11/10/2026

Antes de empezar

Un RAG falla de dos maneras: no trayendo el documento correcto o trayéndolo y mal respondiéndolo. Evaluarlo exige métricas para la recuperación y para la respuesta, y un bucle que conecte ambas.

QUÉ APRENDERÁS

- Evaluar un sistema de recuperación aumentada midiendo por separado la calidad de la recuperación y la de la generación.

Evaluación de RAG

Dos fases, dos diagnósticos

Evaluar un RAG es diagnosticar dos órganos: ¿recuperó la evidencia buena? ¿generó una respuesta fiel a ella? Sin separar las dos fases, un mal resultado no te dice ni dónde corregir.

VIDEO

Evaluar un RAG con RAGAs

Vídeo en español que mide en la práctica fidelidad, relevancia y cobertura de un sistema RAG.



Escanea para verlo en Lumbre

<https://www.youtube.com/watch?v=oELZcpNCU9c>

Dos etapas, dos métricas

- La recuperación se mide por si trajo los documentos correctos.
- La generación se mide por fidelidad al contexto y utilidad.
- Confundirlas oculta la causa real del fallo.

Métricas de recuperación

Recall y precisión a k, MRR o nDCG valoran si los fragmentos relevantes aparecieron entre los recuperados y bien ordenados. Son la base sobre la que se sostiene todo lo demás.

Pertinencia y ruido

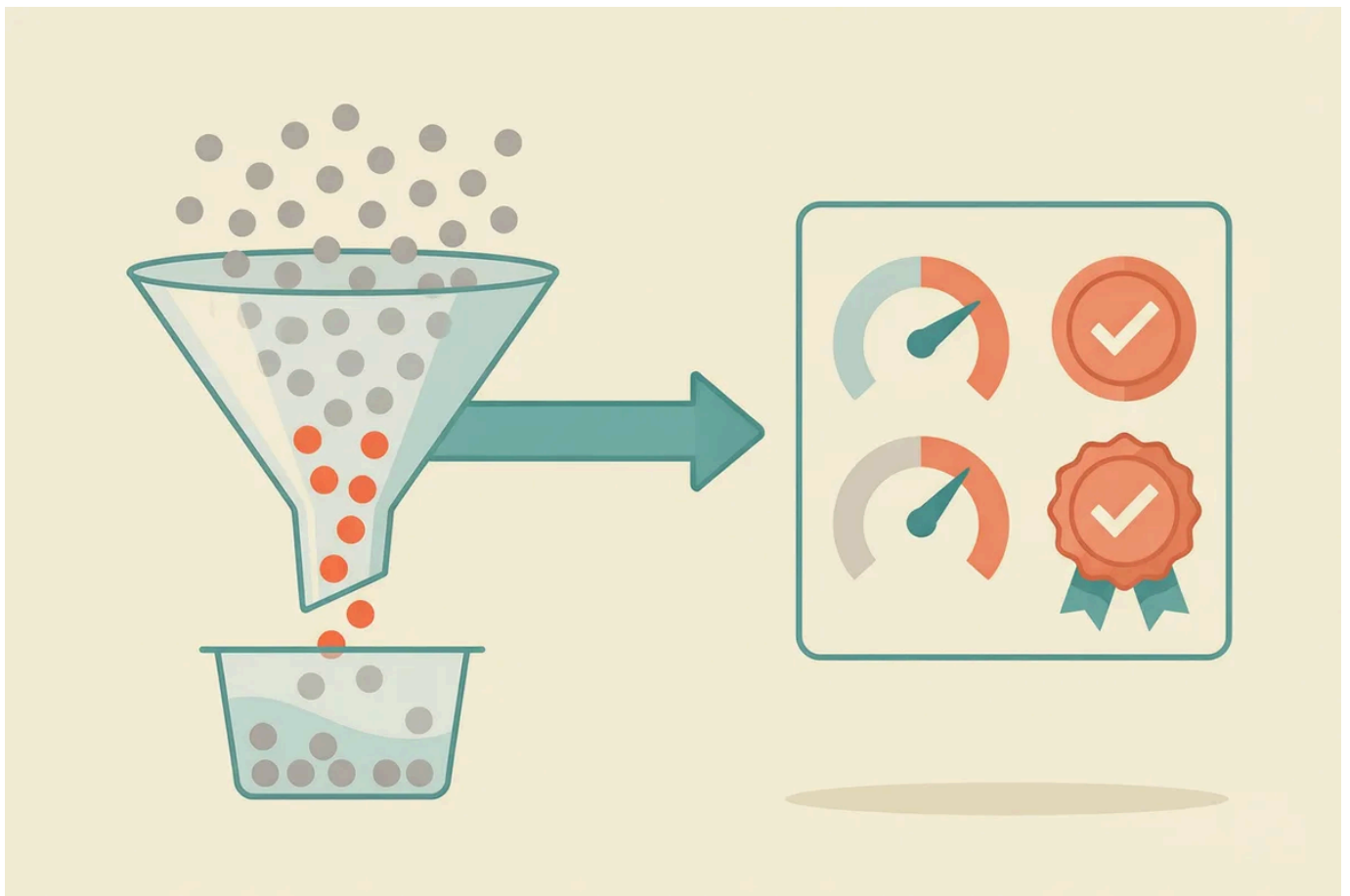
Recuperar muchos fragmentos irrelevantes contamina el contexto y confunde al generador. La precisión mide cuánto de lo traído aporta de verdad.

Métricas de generación

- Fidelidad: ¿la respuesta se apoya solo en el contexto?
- Relevancia: ¿responde a la pregunta planteada?
- Cobertura: ¿usa toda la evidencia necesaria?

El marco RAGAS y jueces automáticos

Existen bibliotecas que puntúan fidelidad y relevancia usando un modelo juez. Escalan la evaluación sin humanos, aunque heredando sus sesgos.



Primero se mide si el embudo trajo la evidencia buena; después, si la respuesta le fue fiel: dos órganos, dos diagnósticos.

Conjuntos de prueba de referencia

Reunir preguntas con sus respuestas y documentos oro permite medir de forma reproducible. Sin un buen conjunto de prueba, toda métrica flota.

Evaluar de extremo a extremo

Además de las fases sueltas, importa el resultado final: ¿el usuario obtiene una respuesta correcta y citada? Una prueba integral detecta fallos de integración.

Atribuir el fallo

Si la respuesta es mala, ¿fue por mala recuperación o mala generación? Comprobar si el documento oro estaba entre los recuperados separa las culpas.

Iterar componentes

- Cambiar el troceado y reevaluar la recuperación.
- Sustituir la incrustación y medir de nuevo.
- Ajustar el prompt y verificar fidelidad.

Regresiones en producción

Un modelo nuevo o un cambio de índice pueden empeorar sin que se note. Congelar un conjunto de prueba y reejecutarlo protege contra regresiones silenciosas.

Coste, latencia y calidad

No solo de precisión vive el sistema: medir también coste por consulta y retardo. Un RAG carísimo por un punto de recall puede no merecer la pena.

Errores frecuentes

- Medir solo de extremo a extremo: un porcentaje global no dice si falló el buscador o el redactor.
- Ajustar el sistema contra el conjunto de prueba hasta saturarlo: las métricas suben y el mundo real sigue igual (Goodhart).
- Olvidar coste y latencia: un RAG perfectísimo pero carísimo por consulta se muere en producción.

Ejemplo resuelto: diagnosticar respuestas malas tras un cambio

Tras modificar el troceado, los usuarios del asistente se quejan: las respuestas siguen bien redactadas pero citan datos viejos.

1. Reejecutas el conjunto de prueba congelado antes y después del cambio.
2. El recall de los documentos oro cayó: la evidencia buena ya no llega al contexto.
3. La fidelidad de generación se mantuvo: el redactor hace bien su papel con lo que le dan.
4. Atribuyes la culpa a la recuperación y restauras el troceado por secciones.
5. Automatizas esta doble medición en el despliegue para que la próxima regresión salte sola.

¿Para qué sirve en la realidad?

Los equipos que mantienen RAG en producción tratan la evaluación como tests unitarios: sin un protocolo fijo de métricas, cada cambio es una apuesta a ciegas sobre usuarios reales.

1. En la evaluación de un RAG, ¿por qué conviene medir por separado recuperación y generación?

- ☐ porque usan el mismo modelo
- ☐ para atribuir el fallo a la fase correcta y saber qué corregir
- ☐ porque la recuperación no tiene métricas
- ☐ para duplicar el coste

2. La métrica $\text{recall}@k$ en un RAG mide cuántos de los fragmentos relevantes se recuperaron entre los k primeros.

- ☐ Verdadero
- ☐ Falso

ACTIVIDAD INTERACTIVA DISPONIBLE EN LUMBRE

3. Evaluar RAG



4. Ordena el diagnóstico de una respuesta incorrecta de un RAG:

Elementos:

- comparar la salida con la respuesta de referencia
- verificar si el documento oro se recuperó
- si se recuperó, revisar cómo se usó al generar
- atacar la fase que falló y reevaluar

Prueba	. . .
Recuperacion	. . .
Generacion	. . .
Corregir	. . .

Método de evaluación <https://en.wikipedia.org/wiki/Evaluation> Principio general de medir un sistema contra un criterio.

5. Une cada métrica de RAG con la pregunta que responde.

5.	¿llegó la evidencia correcta entre los k fragmentos?	-----
5.	¿cuánto de lo recuperado aportaba de verdad?	-----
5.	¿la respuesta se apoya solo en el contexto?	-----
5.	¿respondimos a lo que preguntaron?	-----

Opciones: recall · precisión · fidelidad · relevancia

Dos fases, dos diagnósticos

Evaluar un RAG

Recuperación

recall y precisión

orden: mrr y ndcg

Generación

fidelidad

relevancia

cobertura

Protocolo

conjunto de prueba fijo

jueces automáticos

Producción

regresiones

coste y latencia

El mapa mental interactivo es interactivo en la versión online.



Precisión y exhaustividad https://en.wikipedia.org/wiki/Precision_and_recall Las dos mitades de la medición clásica de recuperación de información.

6. Tu RAG baja en calidad tras un cambio de troceado, pero la fidelidad de generación se mantuvo. ¿Qué fase señalas como culpable y qué métrica de recuperación revisarías?

Referencia curricular

Grado:	8
Materia:	Sistemas de Recuperación Aumentada (RAG) · Inteligencia Artificial

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



<https://univ.lumbre.lat/object/cdia-rag-004-evaluacion-de-rag>

Material educativo generado desde Lumbre · lumbre.lat