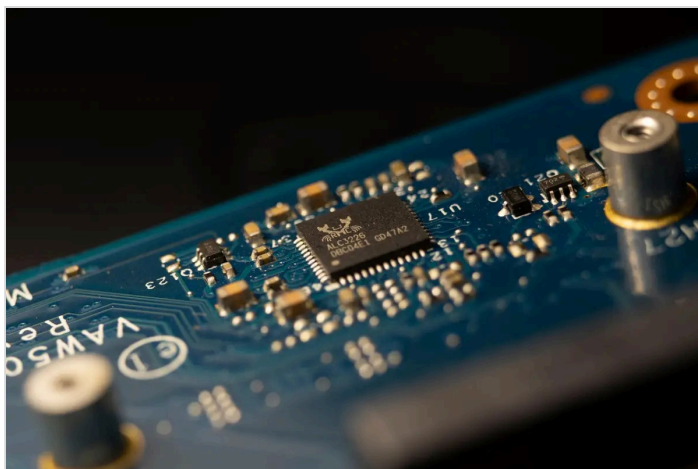


L U M B R E



Un chatbot con memoria de tus datos: RAG sin fórmulas

Materia: Desarrollo de Software con IA

07/10/2026

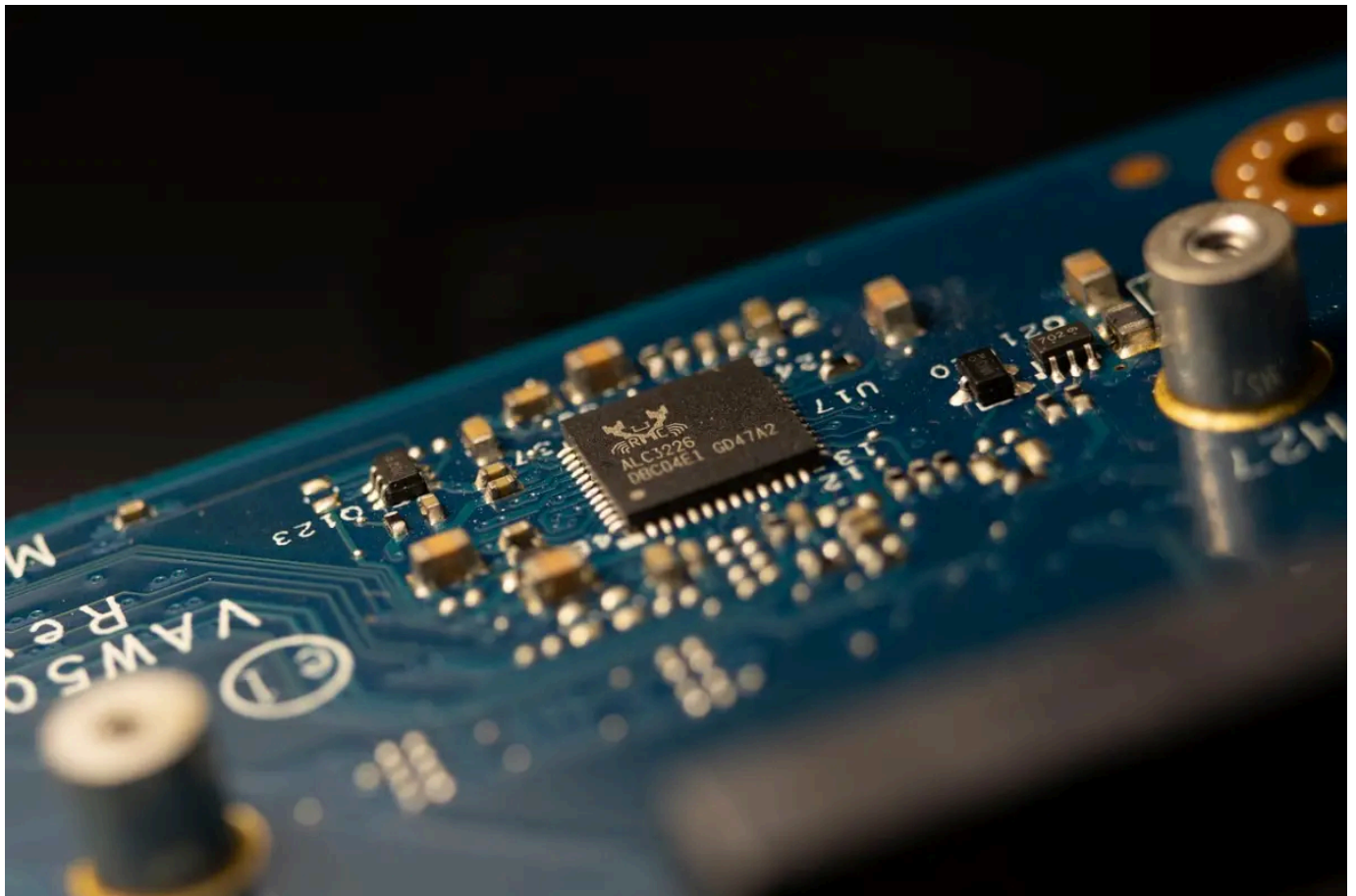
Antes de empezar

La IA que usa tu app no conoce tus datos: sabe del mundo general, no de tu reglamento interno ni del historial de tus clientes. La técnica que cierra ese hueco se llama RAG —generación aumentada por recuperación— y su idea cabe en una frase: antes de dejar que la IA responda, búscale en tus documentos el pedazo relevante y pónselo enfrente. Este OA te da el flujo completo de cuatro etapas, las fallas típicas de cada etapa y el prompt para construir tu propio chatbot con memoria.

QUÉ APRENDERÁS

- Describir el flujo documentos → índice → búsqueda → respuesta con fuentes.

La IA de tu app es un extraterrestre con modales



La app ya es tuya; ahora mételle IA dentro: asistente, clasificación, RAG con fuentes citadas. Y antes del primer prompt de código, la hoja de costos.

Un modelo de lenguaje sabe muchísimo del mundo y absolutamente nada de tu empresa: no leyó tu reglamento, no conoce tus políticas de devolución, no sabe qué acordaste con el cliente Pérez. Si le preguntas directo, improvisará con lo más parecido que tiene: generalidades convincentes. **RAG** (Retrieval-Augmented

Generation) es la solución de sentido común: no le pidas la respuesta; primero busca en TUS documentos el fragmento relevante, entrégaselo, y entonces que responda con tu material enfrente.

ACTIVIDAD INTERACTIVA DISPONIBLE EN LUMBRE

1. Las cuatro etapas del RAG



Dónde se rompe un RAG malo

La magia de exigir fuentes al final: cada respuesta se vuelve verificable en segundos. Un chatbot que cita es un chatbot que puedes auditar; uno que no cita es un generador de rumores con tu logo.

Etapas	Falla típica	Se nota en...
Indexar	trocear mal: un pedazo corta una cláusula a la mitad, o PDFs con texto basura entran crudos	respuestas que mezclan dos secciones sin relación
Buscar	la pregunta usa otras palabras que el documento; el pedazo correcto nunca se recupera	«según la información disponible...» seguido de una generalidad
Aumentar	no se instruye «usa solo este material» → la IA rellena huecos con su cultura general	datos correctos mezclados con políticas inventadas
Responder	no exige fuentes → imposible auditar qué venía del documento y qué no	el cliente pregunta «¿dónde lo dice?» y no hay respuesta

El prompt para construir tu primer RAG

El umbral y la negativa: las dos líneas que hacen honesto al chatbot

Quiero un chatbot de soporte para mi app que responda SOLO con mis documentos\n(docs/: reglamento, preguntas frecuentes, politicas).\n\nImplementa RAG en las 4 etapas y explicame cada decision:\n1. INDEXACION: como troceas los documentos (tamano y solapamiento) y porque; como\n normalizas los PDFs. Los archivos sin texto util van a un reporte, no al indice.\n2. BUSQUEDA: para cada pregunta, recupera los 4 pedazos mas relevantes y muestra\n tambien el score de relevancia. Si el mejor score esta bajo umbral, responde\n "no tengo informacion sobre eso" en lugar de improvisar.\n3. AUMENTAR: plantilla final del prompt: pedazos + pregunta + instruccion explicita\n "usa SOLO el material; si no esta ahi, di que no esta; no agregues conocimiento propio".\n4. RESPUESTA: toda afirmacion debe citar el documento y seccion de origen.\n\nEntregame ademas 10 preguntas de prueba: 4 con respuesta en los docs, 3 con respuesta\nparcial, 3 que NO estan en los docs (ahi el chatbot debe negarse). Corre las 10 y\nmuestrame las respuestas con sus citas.

Para reforzar

Prueba

2. Un cliente pregunta por la garantía de un producto que NO está en tus documentos. El chatbot responde con una política de garantía convincente y sin citas. ¿En qué etapa está la falla raíz?

- ☐ Indexar: faltó trocear mejor los documentos.
- ☐ Responder: el modelo de lenguaje es malo y hay que cambiarlo.
- ☐ Aumentar (y su instrucción): no se forzó "usa solo el material y niegate si no está"; la IA rellenó con cultura general.
- ☐ Buscar: debió recuperarse más cantidad de pedazos por pregunta.

3. Un chatbot RAG bien implementado puede responder cualquier pregunta porque busca en todos tus documentos.

- ☐ Verdadero
- ☐ Falso

Detecta el RAG de pega

ACTIVIDAD
INTERACTIVA
DISPONIBLE EN
LUMBRE

4. Un alumno presume su chatbot con IA sobre sus documentos. Señala la señal de que en realidad no hizo RAG:



5. Reúne cinco documentos reales de tu proyecto (o inventa los de una tienda ficticia) y corre el experimento mental de las 10 preguntas: ¿cuáles respondería tu material, cuáles solo en parte, cuáles no existen? Ese papel es el borrador de tu suite de pruebas RAG antes del primer prompt.

El valor de un asistente no es su elocuencia: es que lo que dice esté en algún lado y puedas señalar dónde.

Referencia curricular

Materia:	Desarrollo de Software con IA
-----------------	-------------------------------

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



<https://empresa.lumbre.lat/object/fsi-m20-0a01>

Material educativo generado desde Lumbre · lumbre.lat