Engineer

Generación aumentada por recuperación

respuestas generadas con documentos recuperados al preguntar

Qué es

La generación aumentada por recuperación es una forma de responder a una pregunta con un modelo de lenguaje en dos pasos. Primero, la aplicación busca en un conjunto de documentos los pasajes que más probablemente contengan la respuesta. Después, los coloca en la entrada del modelo junto a la pregunta, y el modelo redacta su respuesta a partir de ellos, no de lo que absorbió durante el entrenamiento. El nombre, normalmente abreviado como RAG, procede de un artículo de investigación de 2020 sobre tareas lingüísticas que requieren conocimientos, como responder a preguntas abiertas.

Ese artículo describía dos tipos de memoria. Un modelo preentrenado conserva lo aprendido en sus parámetros, lo que los autores llamaban memoria paramétrica. Un índice separado de documentos, consultado por un recuperador entrenado, es la memoria no paramétrica. Su sistema unía un modelo preentrenado de secuencia a secuencia con un índice de vectores densos de Wikipedia y comparaba dos formas de utilizar los resultados: una condiciona toda la respuesta a los mismos pasajes recuperados; la otra puede usar pasajes distintos para cada token que escribe. Consiguió nuevos resultados de vanguardia en tres pruebas de referencia de respuesta a preguntas de dominio abierto, y los autores observaron que el texto generado era más específico, más variado y más fiel a los hechos que el de un modelo comparable sin recuperación.

Una aplicación puede construir el mismo patrón con piezas que ya tiene: un índice de búsqueda de sus propios documentos y un gran modelo de lenguaje al que llama mediante una API. La parte de recuperación es una búsqueda corriente, y aquí aparecen sus dos familias. La recuperación dispersa busca coincidencias entre palabras, como la búsqueda de texto completo que ordena documentos con una función de puntuación como BM25. La recuperación densa convierte la pregunta y cada pasaje en vectores numéricos y devuelve los pasajes cuyos vectores están más cerca: eso es búsqueda vectorial. Un artículo de 2020

sobre recuperación densa de pasajes concluyó que un recuperador denso entrenado con una cantidad modesta de pares de preguntas y pasajes superaba una referencia BM25 sólida entre 9 y 19 puntos porcentuales en la frecuencia con que el pasaje correcto aparecía entre los primeros veinte resultados. Un sistema puede ejecutar ambas búsquedas y combinar sus listas.

Utilice el concepto en una decisión

Aplique «¿Qué es la generación aumentada por recuperación?» a una decisión real. Podemos ayudarle a relacionar el concepto con el sistema que está construyendo o comprando. Sus registros al alcance de los agentes de IA, una única versión de cada cliente, acceso limitado a lo que corresponde y cada respuesta vinculada a su fuente.

Explore nuestros servicios Póngase en contacto →

Compruebe los supuestos detrás del término

Utilice la explicación y los trabajos enlazados para formular preguntas sobre su proyecto. Podemos examinar los supuestos con usted antes de elegir una herramienta. Flujo de datos geográficos 50 veces más rápido — 50x más rapidez en los flujos de procesamiento de datos geográficos

Flujo de datos geográficos 50 veces más rápido Póngase en contacto →

Por qué importa

Un modelo de lenguaje solo sabe lo que figuraba en sus textos de entrenamiento, y esos textos terminan en una fecha fija y nunca incluyeron sus contratos, sus tarifas ni el acta de la semana pasada. El artículo sobre RAG señala dos límites del conocimiento guardado en parámetros: cuesta mostrar de dónde procede una respuesta y cuesta actualizarlo. La recuperación aborda ambos. La respuesta puede citar los pasajes recibidos para que el lector abra la fuente y la compruebe. Además, el conocimiento nuevo llega al modelo en cuanto un documento entra en el índice, sin ninguna ejecución de entrenamiento.

También devuelve la cuestión de quién puede ver qué al lugar que le corresponde. La recuperación es una búsqueda, y una búsqueda puede ejecutarse con los derechos de quien pregunta, de modo que el modelo solo recibe pasajes que esa persona podría haber abierto por sí misma. Un modelo entrenado con esos mismos documentos no tiene esa frontera: todo lo que aprendió está disponible para cualquiera que lo utilice.

El coste es que la respuesta no puede ser mejor que lo que devuelve la búsqueda. Un modelo que recibe pasajes incorrectos redacta una respuesta fluida a partir de ellos, y uno que no recibe nada útil puede responder de memoria de todos modos. Por eso, la mayor parte del trabajo de un sistema de generación aumentada por recuperación se dedica a los datos y a la búsqueda, no al modelo.

Cómo se aplica en la práctica

La traza siguiente sigue una pregunta a través de un sistema de generación aumentada por recuperación. La búsqueda se ejecuta con los permisos del usuario que pregunta, cada pasaje llega con su archivo y página de origen, y las instrucciones indican al modelo que responda únicamente a partir de los pasajes:

```
pregunta: ¿Qué condiciones de entrega se aplican al pedido 4471?
usuario:  sales-3 (rol: sales, región: north)

recuperados, filtrados según lo que sales-3 puede leer:
  [1] contracts/framework-2025.pdf, página 4
      "Las mercancías se entregan en el almacén del comprador
       salvo que el pedido indique lo contrario."
  [2] orders/4471.txt, línea 12
      "Condiciones de entrega: según el acuerdo marco."

instrucciones:
  Responda solo a partir de los pasajes siguientes y cítelos como [n].
  Si no responden a la pregunta, indíquelo.

respuesta:
  El pedido 4471 remite al acuerdo marco [2], según el cual
  las mercancías se entregan en el almacén del comprador [1].
```

Cada parte de esa traza es un dato que el sistema puede comprobar. Una prueba puede confirmar que una pregunta conocida devuelve los pasajes correctos; un script puede verificar que cada `[n]` de la respuesta se refiere a un pasaje realmente recuperado; y quien lee la respuesta puede abrir la página 4 del contrato.

La preparación necesaria se explica en la página de este sitio sobre datos de empresa listos para agentes de IA. Describe cómo convertir PDF antiguos, documentos escaneados y archivos de correo en texto que se pueda buscar, manteniendo la fuente a su lado, y cómo conservar en cada fila su procedencia: el sistema, la carga, el archivo y la página. Así, un agente puede citar el registro que respalda una cifra y una persona puede abrirlo. La página también sitúa las reglas en los datos, no en el prompt: el agente actúa con los privilegios de su usuario, nunca con los de un administrador, y cada lectura se registra junto con la persona para la que se realizó.

Dónde surgen los problemas

La búsqueda no encuentra el pasaje y el modelo responde de todos modos. Cuando no se recupera el pasaje correcto, el modelo recibe algo igualmente y sigue redactando una respuesta. Mida la recuperación por separado: anote preguntas reales junto con los pasajes que las responden y compruebe con qué frecuencia aparecen esos pasajes entre los resultados que recibe el modelo. Indíquele que lo diga cuando los pasajes no respondan a la pregunta e incluya preguntas de ese tipo entre los casos de prueba.

Pasajes que cortan la respuesta por la mitad. Dividir los documentos en fragmentos de longitud fija separa una tabla de su encabezado y una cláusula de su excepción; cualquiera de esas mitades recuperada por separado induce a error. Divida según la estructura que ya tiene el documento, sus secciones, páginas y filas de tablas, y conserve su título y ubicación con cada fragmento.

El índice ignora quién pregunta. Un índice construido por un servicio que puede leerlo todo devolverá pasajes de documentos que la persona que pregunta no puede abrir, y el modelo se los resumirá. Filtre por los permisos del usuario dentro de la búsqueda, en el índice o en la base de datos, antes de que nada llegue al modelo. Una regla en el prompt no es control de acceso.

Instrucciones escondidas en un documento. El texto recuperado lo escribió otra persona, y el modelo no puede distinguir de forma fiable un pasaje de una instrucción. El perfil del NIST para IA generativa describe la inyección indirecta de prompts como ataques contra aplicaciones basadas en modelos de lenguaje a través de datos que probablemente se recuperen, y un artículo de 2023 demostró esos ataques contra sistemas reales. Trate cada pasaje recuperado como entrada no fiable, no permita que un pasaje active una acción sin una comprobación en el código y no dé a las herramientas que el modelo puede llamar más derechos que los del usuario.

Un índice que se queda atrás. Un documento cambia o se elimina, pero el índice conserva los pasajes antiguos, de modo que el sistema responde con seguridad a partir de un contrato que ya no está vigente. Actualice el índice cada vez que cambie una fuente, retire los pasajes cuando se elimine su documento y muestre la fecha de cada fuente junto a la respuesta.

Citas que no dicen lo que afirma la respuesta. Un modelo puede citar el pasaje [2] para una frase que ese pasaje [2] no respalda, y el NIST señala que las propias citas generadas pueden ser inventadas. Compruebe en el código que cada cita apunta a un pasaje recuperado y haga que personas comparen una muestra de respuestas con los pasajes citados como parte de las pruebas periódicas.

Lecturas adicionales

- Generación aumentada por recuperación para tareas de procesamiento del lenguaje que requieren conocimientos — El artículo de 2020 que presentó los modelos RAG y comparó sus dos formulaciones en respuesta a preguntas y generación.
- Recuperación densa de pasajes para responder a preguntas de dominio abierto — Un recuperador denso entrenado con pares de preguntas y pasajes, comparado con BM25 para encontrar el pasaje que contiene la respuesta.
- NIST AI 600-1: perfil de inteligencia artificial generativa — El perfil de riesgos del NIST para IA generativa, que aborda las invenciones, la privacidad de los datos y la inyección indirecta de prompts.
- No es lo que aceptó: comprometer aplicaciones reales integradas con modelos de lenguaje mediante inyección indirecta de prompts — Cómo las instrucciones escondidas en datos recuperados toman el control de una aplicación, con ataques a sistemas reales.

Glosario

Elija según sus restricciones reales

Traiga un conjunto de datos representativo, las preguntas que debe responder y el cuello de botella actual. Podemos acordar cómo medir la calidad, la capacidad de procesamiento y la recuperación.

Vea cuánto cuesta Póngase en contacto →

Haga práctica la terminología

Pregunte cómo se aplica «¿Qué es la generación aumentada por recuperación?» a su organización. Basta un ejemplo concreto de su sistema actual para iniciar la conversación.

Explore nuestros servicios

welcome@engineer.company +45 50 10 61 84 Copenhague, Dinamarca

Llévese esta página

Tarjeta de contacto

Creado con esmero y cariño por D.

CVR 45969703 · Tagensvej 105C, 3. th, 2200 København N

Derechos de autor © 2025 – actualidad · Engineer ApS · Todos los derechos reservados.

Impreso desde https://engineer.company/es/glossary/retrieval-augmented-generation/