Ingeniería de Sistemas de Información Empresarial 67(4):551–
561 (2025) https://doi.org/10.1007/s12599-025-00945-3

LEMA

Generación aumentada por recuperación (RAG)

Michael Klesel·H. Felix Wittmann

Recibido: 22 de julio de 2024 / Aceptado: 7 de abril de 2025 / Publicado en línea: 1 de junio de 2025

Palabras claveGeneración aumentada por recuperación - Inteligencia artificial - Grandes modelos de lenguaje - Recuperación de información

1 Introducción

La necesidad de información es un aspecto fundamental de la naturaleza humana y, como tal, existen esfuerzos continuos para mejorar la recuperación de información mediante sistemas de información (Alavi y Leidner).2001; Alavi y otros.2024Las empresas se ven particularmente afectadas por esto, ya que disponen de una gran cantidad de datos y sus empleados necesitan acceder a ellos. Desafortunadamente, los sistemas actuales no logran satisfacer adecuadamente las expectativas de los empleados. De hecho, algunos estudios han demostrado que el 79 % de los empleados están insatisfechos con las interfaces de usuario de los sistemas de búsqueda empresarial (Cleverley y Burnett).2019Esto ha generado la necesidad de nuevos enfoques que puedan abordar mejor las necesidades de información de las organizaciones.

Agentes conversacionales (AC) impulsados por inteligencia artificial (IA) y modelos de lenguaje grandes (MLG) basados en transformadores en particular (Vaswani et al.).2017Las consultas automáticas (CA) han revolucionado la forma en que se accede a la información hoy en día. En comparación con los sistemas empresariales tradicionales, las CA ofrecen dos ventajas clave: en primer lugar, permiten a los usuarios formular preguntas de forma natural e intuitiva utilizando lenguaje natural.

Aceptado tras dos revisiones por Christine Legner

M. Klesel (&) - HF Wittmann
Universidad de Ciencias Aplicadas de Frankfurt, Nibelungenplatz 1, Frankfurt, Alemania
correo electrónico: michael.klesel@fra-uas.de

M. Klesel
Centro Hessiano de Inteligencia Artificial (hessian.AI), Darmstadt, Alemania

lenguaje, recibiendo respuestas igualmente conversacionales. En segundo lugar, son cada vez más capaces de abordar tareas de búsqueda complejas, facilitando la resolución de problemas y la toma de decisiones en diversos ámbitos (White).2024). Por ejemplo, las personas pueden usar las CA para acceder a información de recetas de cocina (Jaber et al.2024) y para obtener ayuda con tareas complejas relacionadas con la química (Bran et al.2024) y problemas geométricos (Trinh et al.2024).

En las organizaciones, la necesidad de información suele estar relacionada con datos que no se encuentran habitualmente en internet. Por ejemplo, un empleado puede necesitar un resumen de un análisis exhaustivo de requisitos o los detalles de un contrato. Dado que es improbable que los sistemas de aprendizaje automático (LLM) estándar hayan utilizado los datos necesarios, como documentos contractuales, durante su formación, las respuestas que generan probablemente no sean fiables. En general, los LLM pueden generar ocasionalmente respuestas sin fundamento. Este tipo de respuesta se denomina... alucinación (Maynez y otros.2020; Ji y otros. 2023), que se define como " contenido que es inconsistente con los hechos del mundo real o las aportaciones de los usuarios" (Ji y otros.2023, pág. 1).[1]
Las alucinaciones son particularmente críticas, porque socavan la confiabilidad de los resultados y se han observado en diversos escenarios, como el uso multilingüe de LLM (Guerreiro et al.).2023), o en situaciones específicas del contexto, como la medicina (Pal et al.2023).

La generación aumentada por recuperación (RAG, por sus siglas en inglés) se ha propuesto como un nuevo marco para la IA que busca integrar conocimiento adicional, como datos organizacionales, y generar resultados que puedan vincularse a ese conocimiento (Lewis et al.).2020 Esto permite a los usuarios acceder a la información

[1]Recientemente, la literatura ha sugeridomierdacomo un término más apropiado, ya que no existe el concepto de veracidad en la formación de los LLM (Hicks et al.).2024Si bien coincidimos en que tiene mérito proponer un término nuevo y posiblemente más apropiado, en este manuscrito utilizamos alucinaciones para garantizar la coherencia con la bibliografía pertinente.

Desde el interior de una organización, reduce el riesgo de alucinaciones. Esta nueva arquitectura ofrece avances importantes con respecto a las anteriores y plantea nuevos retos para la investigación y el ámbito académico.

Artículos anteriores sobre palabras clave ya han cubierto aspectos importantes de la IA, a saber, la IA justa (Feuerriegel et al.). 2020), IA como servicio (Lins et al.2021), modelos de fundamentos (Schneider et al.2024), y la IA generativa (Feuerriegel et al.2024Contribuimos a este compromiso continuo con los desarrollos actuales de la IA centrándonos en RAG. En concreto, revisamos la arquitectura fundamental de RAG y destacamos algunas extensiones que pueden mejorar una arquitectura RAG estándar. Mostramos cómo se puede utilizar RAG en diferentes escenarios de uso y resumimos las ventajas y los desafíos más importantes que deben tenerse en cuenta al utilizar RAG y sus extensiones específicas. Finalmente, analizamos importantes líneas de investigación para el futuro de la IA.BISEcomunidad al resaltar las implicaciones que surgen como consecuencia del uso de arquitecturas RAG.

2 Generación Aumentada por Recuperación (RAG)

2.1 Marco fundamental

La idea central de RAG es combinar las capacidades generativas de los LLM con el conocimiento externo recuperado de una base de datos separada (por ejemplo, una base de datos organizacional) (Lewis et al.). 2020). Mientras que Lewis et al. (2020) reconocer trabajos previos sobre la integración de datos externos (Guu et al.2020; Karpukhin y otros.2020 Pérez y otros.2019), ellos acuñaron el término ''Generación aumentada por recuperación (RAG)'' y propuso un marco general que aprovecha la fortaleza de la memoria paramétrica preentrenada (es decir, el LLM) con la memoria no paramétrica (es decir, una base de datos separada) como una nueva forma de mejorar el rendimiento para tareas intensivas en conocimiento.

La memoria paramétrica se refiere a la información almacenada en los parámetros de un modelo. En lugar de almacenar directamente datos fácilmente comprensibles, la memoria paramétrica almacena parámetros del modelo que pueden utilizarse posteriormente para regenerar información. Cuantos más parámetros tenga un modelo, más información podrá representar fielmente (Brown et al.).2020Los modelos actuales suelen incluir el número de parámetros en su nombre. Por ejemplo, Mistral 7B (Jiang et al.2023) es un modelo con siete mil millones ($7 \cdot 10^9$) parámetros. Si bien la cantidad de parámetros es un detalle técnico, tiene consecuencias importantes. Por ejemplo, la evaluación de modelos suele involucrar esta cantidad, ya que los modelos más grandes requieren más recursos para ejecutarse. Por ello, los modelos generalmente se comparan con otros modelos que tienen la misma cantidad de parámetros. Por otro lado, si un modelo pequeño funciona bien en comparación con un modelo grande, generalmente se considera que el modelo pequeño es el mejor. Preferible. En el contexto de RAG, es importante señalar que la memoria paramétrica solo contiene información que se ha proporcionado como parte del entrenamiento.

En contraste, la memoria no paramétrica, o memoria externa, se refiere a la información que se encuentra fuera del modelo (por ejemplo, información de una base de datos). Por lo tanto, esta información es independiente de las restricciones del modelo. Ejemplos de recursos de memoria no paramétrica incluyen sitios web como Wikipedia o datos específicos de un dominio (por ejemplo, datos organizacionales). En otras palabras, la memoria no paramétrica permite la integración de conocimiento no utilizado previamente en el proceso de entrenamiento. Por ejemplo, Veturi et al. (2024) utilizar datos organizacionales (por ejemplo, documentos de políticas) para mejorar el rendimiento de un sistema de preguntas frecuentes (FAQ).

En principio, los datos utilizados para entrenar un modelo lineal de aprendizaje (por ejemplo, un corpus de texto de Wikipedia) también podrían usarse para la memoria no paramétrica. De hecho, la mayoría de las interfaces de usuario actuales utilizan algún tipo de arquitectura RAG para obtener los detalles fácticos de los datos originales. Por ejemplo, un sistema de preguntas frecuentes (Veturi et al.)2024No solo proporcionará una respuesta a una pregunta específica, sino también un enlace al documento correspondiente (por ejemplo, un documento de política). Dado que la mayoría de las organizaciones utilizan modelos de aprendizaje automático (LLM) de un proveedor (por ejemplo, de Microsoft), la arquitectura RAG puede utilizarse para añadir datos internos y, por lo tanto, conocimiento contextual. Al igual que el entrenamiento de modelos con LLM, una arquitectura RAG requiere una fase de recopilación de datos donde los datos externos (es decir, no paramétricos) se almacenan en una base de datos dedicada. Es importante destacar que esta base de datos es distinta de la memoria paramétrica (es decir, el LLM).

Cifra1Proporciona una descripción general de las diferencias entre un modelo base, el ajuste fino de LLM y RAG. En el primer escenario (Modelo de Fundación),Todos los datos de entrenamiento dan como resultado un LLM y forman parte del modelo paramétrico. Lo más habitual es que los modelos se entrenen con grandes corpus de texto de la World Wide Web (Touvron et al.).2023), incluyendo el conjunto de datos CommonCrawl y el Pile (Gao et al. 2020La creación y el entrenamiento del modelo base requieren enormes recursos e infraestructura. Por lo tanto, su creación solo es factible para organizaciones muy grandes. Muchos de estos modelos están disponibles como servicio y, por consiguiente, pueden ser utilizados por cualquier persona. Los modelos base están bien equipados para una amplia gama de aplicaciones que no dependen de datos organizacionales ni privados (Schneider et al.). 2024).

En el segundo escenario (Ajuste fino del LLM),Se utilizan datos adicionales específicos del dominio, como documentos internos, para actualizar los parámetros del LLM.[2]y así mejorar

[2]Esto se suele hacer utilizando retropropagación (Rumelhart et al.). 1986El ajuste fino de los LLM se ha visto enormemente facilitado por técnicas como LoRA (Hu et al.).2022) que se basan en la retropropagación. Para una visión general actual, véase, por ejemplo, (Ding et al.2023).

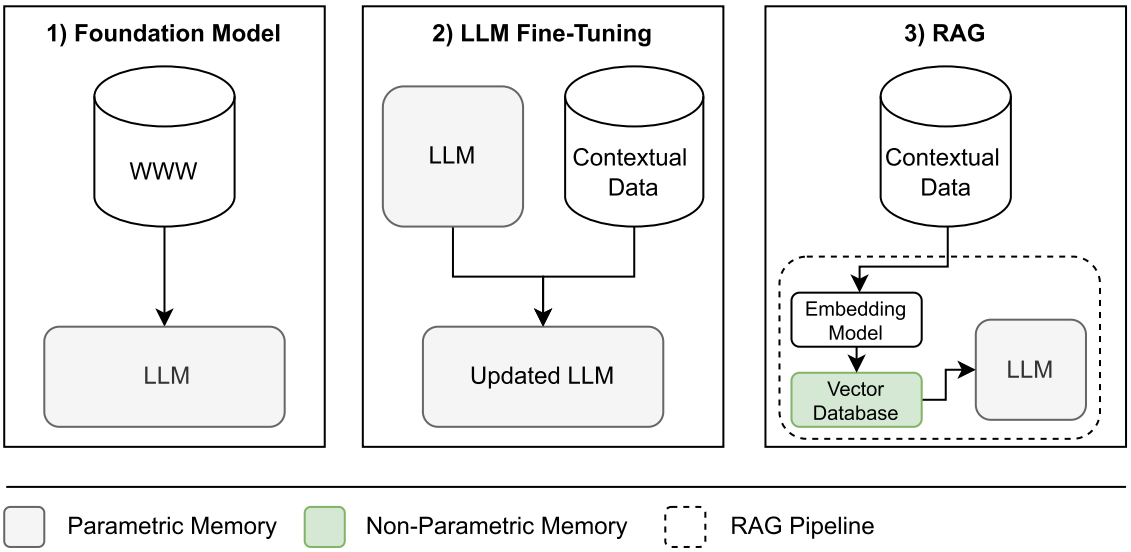

Figura 1Comparación de enfoques LLM

el desempeño de un LLM con respecto a los requisitos específicos y las tareas del dominio en cuestión.

El ajuste fino es menos costoso y requiere menos recursos que la creación de un modelo base. Esto permite que organizaciones pequeñas e individuos ajusten modelos base para un contexto específico. Esto resulta particularmente interesante para aplicaciones que requieren conocimiento específico del dominio que no se encuentra en los corpus de texto de uso común disponibles en Internet. Por ejemplo, el ajuste fino se puede utilizar para entrenar un modelo capaz de responder preguntas específicas del dominio de la agricultura (Balaguer et al.).2024).

En el tercer escenario (TRAPO),Se genera una base de datos independiente con información vectorial (es decir, incrustaciones) utilizando un modelo de incrustación y datos contextuales. Esta parte independiente se denomina memoria no paramétrica. Como explicaremos más adelante, la base de datos vectorial se utiliza para el aumento de datos con el LLM para mejorar sus resultados. La creación de una base de datos vectorial requiere incluso menos recursos que el ajuste fino (Balaguer et al.).2024Por lo tanto, con las capacidades tecnológicas suficientes, esto es en principio factible para muchas organizaciones.

El uso de una arquitectura RAG da como resultado una canalización, que se muestra en la figura.2El pipeline RAG comienza con una consulta y termina con un resultado. Entre medias, hay tres partes fundamentales:recuperación, aumento,ygeneración. Tenga en cuenta que la ampliación es la salida del recuperador y sirve como entrada para el generador.

Modelo de incrustaciónEl modelo de incrustación traduce datos de diferentes modalidades, como texto, audio, imágenes o vídeo, en un vector. Además, y esto es importante, se utilizó el mismo modelo de incrustación que se empleó para la creación de Se debe utilizar una base de datos vectorial para traducir la consulta de entrada en un vector, ya que la similitud entre la consulta y los fragmentos (o documentos) en la base de datos se mide utilizando estos vectores (Steck et al.).2024).

PerdigueroEl recuperador busca la información más relevante en la base de datos vectorial calculando la similitud entre una consulta de entrada y los documentos de dicha base de datos. Para ello, utiliza los vectores calculados con el modelo de incrustación y el vector calculado para la consulta. El recuperador debe usar el mismo modelo de incrustación para la consulta que el que se usó previamente para los documentos en la base de datos vectorial. Como resultado, el recuperador sugiere un contexto, que suele ser una lista de fragmentos o documentos recuperados.[3]Lo más habitual es que esto se haga seleccionando los k mejores resultados (por ejemplo, los 5 mejores), ordenados según su puntuación de similitud. Este tipo de ordenación se conoce a veces como principio de ordenación probabilística (Robertson).1977Por ejemplo, si la consulta se refiere a un cliente específico, los documentos más relevantes (por ejemplo, los 5 documentos principales) relacionados con este cliente se incluyen en el contexto.

AumentoEl contexto se utiliza para generar una consulta extendida. Por ello, los modelos de lenguaje utilizan un tipo de plantilla de aumento que define cómo se amplía una consulta de usuario. Lo más importante es que el contexto se incluye explícitamente en esta consulta. Por ejemplo, una plantilla de aumento básica indica al modelo de lenguaje que utilice información específica, especificando: «Utiliza el siguiente contexto: [contexto]». A continuación, le pide al modelo que responda a una pregunta basándose en esa información. «Dada la información de contexto, responde a la pregunta...»

[3]Debido a que las ventanas de contexto siguen aumentando, las soluciones más recientes pueden usar documentos completos en lugar de fragmentos de texto más pequeños.

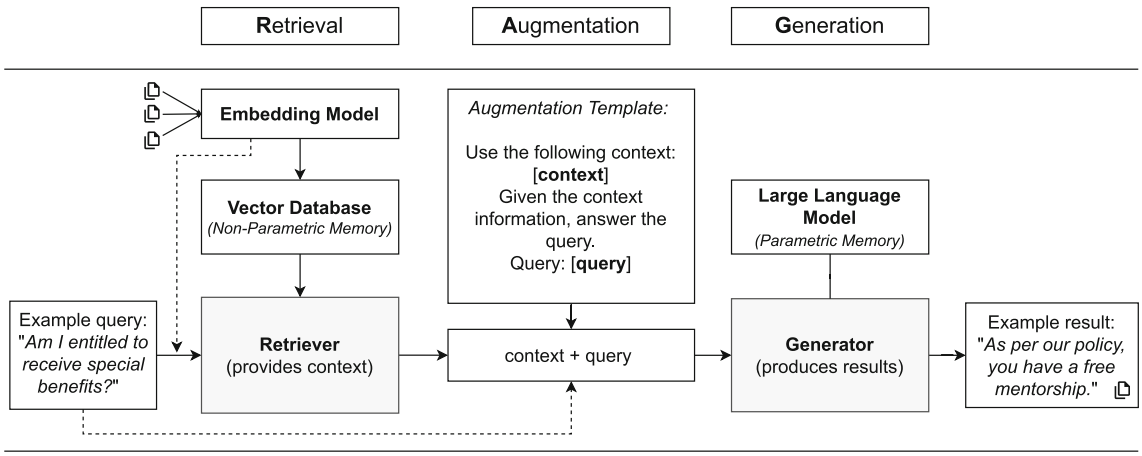

Figura 2Una arquitectura RAG básica basada en Lewis et al.2020)

consulta: [consulta].' La parte [contexto] puede incluir enlaces a los documentos más relevantes que se encuentren en la base de datos vectorial.

GeneradorEl generador toma la consulta, la amplía con información de la base de datos y genera un nuevo resultado. De nuevo, el proceso de generación puede usar la información del contexto y proporcionar el enlace al documento original (por ejemplo, un hipervínculo a un documento).

2.2 RAG mejorado

Cifra2Proporciona una visión general de lo que puede considerarse una arquitectura RAG básica. Además, la literatura reciente sugiere varias maneras de mejorar el rendimiento de una arquitectura RAG básica. Por ejemplo, los enfoques de recuperación de información jerárquica permiten una comprensión e integración más profunda de la información entre documentos, lo que mejora el rendimiento en tareas de razonamiento complejas de varios pasos. En particular,RAPTOR (El procesamiento abstractivo recursivo para la recuperación organizada en árbol) incrusta, agrupa y resume recursivamente el texto en múltiples niveles de abstracción (Sarthi et al.2024).

Además, los enfoques basados en grafos, como GraphRAG, ofrecen mejoras significativas al extraer grafos de conocimiento y estructurarlos jerárquicamente para mejorar las tareas basadas en RAG (Edge et al.). 2024GraphRAG se puede utilizar para extraer un grafo de conocimiento a partir de un texto, construyendo una jerarquía que luego se utiliza para aprovechar estas estructuras basadas en grafos para realizar una tarea basada en RAG.

Además, los pensamientos aumentados mediante recuperación (RAT) mejoran el aumento a través de una Cadena de Pensamiento (CoT) de cero ejemplos, refinándola iterativamente con la información recuperada. Este método proporciona una salida más contextual y coherente, apoyando tareas como la generación de código y el razonamiento matemático (Wang et al.).2024cEl mecanismo subyacente, CoT, que se ha vuelto fundamental para mejorar las capacidades de razonamiento LLM, se desarrolló por primera vez. por Google en 2022 (Wei et al.2022) y ahora se ha integrado en modelos como GPT-4o.

Además, el ajuste fino aumentado por recuperación (RAFT) combina las ventajas de RAG y el ajuste fino, creando conjuntos de datos sintéticos para ajustar modelos a dominios específicos (Zhang et al.).2024RAFT supera a RAG tradicional en dominios especializados como la medicina. Consiste en la creación de un conjunto de datos sintéticos con consultas, documentos relevantes y respuestas objetivo. Un modelo puede ajustarse con este conjunto de datos para adaptarlo al conocimiento y estilo del dominio. RAFT permite que el modelo «estudie» el conocimiento del dominio con antelación, lo que se traduce en un mejor rendimiento que RAG tradicional.

Métodos innovadores como RA-ISF (retroalimentación iterativa aumentada por recuperación) descomponen las tareas en submódulos, mejorando el razonamiento fáctico y reduciendo las alucinaciones (Liu et al.).2024).

Estos ejemplos pretenden reflejar el potencial que ofrece RAG. Para una descripción más completa de las mejoras recientes, remitimos a revisiones exhaustivas de RAG (Zhao et al.).2024; Yu y otros.2024; Gao y otros.2023).

3. Oportunidades y desafíos de RAG

3.1 Casos de uso de RAG

En esta sección, ofrecemos una descripción general de ejemplos de casos de uso donde RAG puede utilizarse para mejorar sustancialmente el rendimiento de tareas específicas (véase la tabla).1).

La RAG ha impulsado el desarrollo de sofisticados agentes de respuesta a preguntas, destacando su aplicación en el ámbito de las preguntas frecuentes (FAQ). Mediante la integración de conocimiento específico del dominio, los sistemas de FAQ basados en RAG pueden generar respuestas precisas y fiables a las preguntas más comunes.

Tabla 1Ejemplos de casos de uso con RAG

Tarea	Ejemplo de caso de uso	Referencias
Preguntas frecuentes	RAG puede utilizarse para desarrollar un agente sofisticado de respuesta a preguntas. Esto puede implementarse, por ejemplo, para preguntas frecuentes (FAQ). Al incorporar conocimiento específico, un sistema de FAQ basado en RAG es capaz de responder con precisión a preguntas específicas formuladas en lenguaje natural.	Veturi y otros (2024)
En tiempo real información recuperación	RAG puede utilizarse para integrar fuentes de datos adicionales y proporcionar información en tiempo real. Por ejemplo, la aplicación ChatGPT para Android recupera información de la web en tiempo real para mejorar la puntualidad y la precisión de los resultados. También se puede incorporar material externo y actualizado permitiendo a los usuarios cargar documentos sobre la marcha y utilizar esta información en el proceso de generación.	Amri y otros (2024) y Khan et al. (2024)
Contexto-específico respuestas	RAG puede utilizarse para mejorar el desarrollo de aplicaciones que requieren información específica. Por ejemplo, los asistentes de programación con inteligencia artificial, adaptados al contenido específico de cada curso, pueden usar fuentes de datos adicionales (como los materiales del curso) para mejorar la precisión de los resultados e incluir una referencia al material pertinente.	Wei y otros (2024), Kazemitabaar y otros (2024), Rai et al. (2024) y Strobel y Banh (2024)
Mejorado contenido generación	RAG puede utilizarse para guiar el proceso de generación de contenido mediante la incorporación de datos adicionales. De esta forma, la calidad y relevancia del contenido generado reflejan el conocimiento actual y las tendencias relevantes para la generación de contenido como los comentarios.	Wu y otros (2024) y Wang et al. (2024b)
Federado buscar	RAG puede utilizarse para mejorar las capacidades de obtención de información relevante a través de fuentes de datos heterogéneas en combinación con un LLM.	Wang y otros (2024a)

en lenguaje natural por el usuario, reemplazando el método tradicional de manejo de preguntas frecuentes (FAQ) – mediante el uso de un conjunto de preguntas y respuestas estáticas "predefinidas" – y mejorando así la experiencia general del usuario (Veturi et al.).2024).

RAG también puede mejorar la recuperación de información en tiempo real, lo que aumenta la puntualidad y la precisión de la información proporcionada. Por ejemplo, la aplicación ChatGPT para Android obtiene información actualizada de la web en tiempo real. Además, RAG facilita la incorporación de materiales externos y actuales al permitir que los usuarios carguen documentos cuando lo necesiten. Esta capacidad resulta especialmente útil en situaciones donde la información más reciente es crucial (Amri et al.).2024; Khan y otros.2024).

En aplicaciones que requieren respuestas contextualizadas, RAG es una herramienta valiosa. Por ejemplo, los asistentes de programación con inteligencia artificial adaptados al contenido específico de cada curso pueden recurrir a fuentes de datos adicionales, como los materiales del curso, para mejorar la precisión de sus resultados. Este enfoque no solo aumenta la precisión de los resultados, sino que también garantiza que las respuestas generadas estén directamente vinculadas a los materiales de referencia pertinentes (Wei et al.).2024 Kazemitabaar y otros.2024Rai y otros.2024Strobel y Banh2024).

RAG mejora significativamente el proceso de creación de contenido al incorporar los conocimientos y tendencias más recientes. Esta capacidad es especialmente relevante para la creación de contenido como comentarios, donde la inclusión de información actualizada es fundamental. Al guiar el proceso de generación de contenido con datos adicionales, RAG garantiza que La calidad y relevancia de los resultados se mantienen a un nivel alto (Wu et al.).2024; Wang y otros.2024b).

Finalmente, RAG mejora la búsqueda federada al optimizar la capacidad de recuperar información relevante de fuentes de datos heterogéneas en conjunto con LLM. Esta combinación permite un proceso de recuperación más completo y eficiente, garantizando que los usuarios puedan acceder sin problemas a información relevante de una amplia gama de fuentes. Esto resulta particularmente beneficioso en entornos de información complejos donde los datos se distribuyen en múltiples plataformas (Wang et al.).2024a).

3.2 Oportunidades de RAG

Al incorporar datos externos, la implementación de RAG proporciona un comprensión contextual mejorada (Lewis y otros.2020Como resultado, las consultas que requieren conocimientos específicos no presentes en el entrenamiento del modelo lineal de aprendizaje (LLM, por sus siglas en inglés) y, por lo tanto, no reflejados en sus parámetros, pueden procesarse de forma eficaz. El uso de modelos base suele ser problemático, ya que los datos de entrenamiento están desactualizados. Por ejemplo, un modelo con datos de entrenamiento de 2023 o anteriores no puede responder preguntas relacionadas con las elecciones europeas de 2024. Con RAG, se pueden añadir datos más recientes, como los del sitio web oficial de las elecciones, que pueden denominarse memoria no paramétrica. De este modo, un sistema basado en RAG tiene el potencial de recuperar información precisa sobre las elecciones.

Cuando los modelos lineales generalizados (LLM) intentan responder preguntas sobre un dominio específico que no forma parte de los datos de entrenamiento,alucinaciones son probables. Una forma de abordar este problema es ajustar el LLM, lo cual es menos costoso que construir un modelo base, pero requiere, sin embargo, recursos considerables. Los estudios han demostrado que el ajuste fino también puede provocar alucinaciones (Gekhman et al.).2024Por otro lado, RAG es una forma eficaz de incorporar este conocimiento. Al añadir información adicional, se pueden responder preguntas utilizando estos datos, lo que reduce la probabilidad de respuestas inexactas. Por lo tanto, RAG es una medida eficaz para mejorarexactitud de los hechos.

Una arquitectura RAG permite proporcionar referencias a los datos contextuales almacenados en la base de datos vectorial. Proporcionar referencias válidas al resultado generado se ha denominado puesta a tierra (Magesh y otros.2024La referencia a fuentes es una ventaja significativa, ya que proporciona al usuario información adicional sobre el origen de la información. Por lo tanto, un usuario que busca información en un área específica puede consultar esta referencia para verificar la respuesta y obtener información adicional.[4]

Además, una arquitectura RAG puede utilizarse para limitar el espectro de respuesta a un dominio de conocimiento deseado. (límites de conocimiento para las maestrías en derecho),lo cual puede definirse implícitamente aportando conocimiento adicional. Los modelos base a menudo carecen de precisión en las respuestas, ya que no se les proporcionan las condiciones límite y las restricciones adecuadas que delimitan el espacio de soluciones de un modelo de aprendizaje automático. Un agente conversacional implementado en un sitio web empresarial no podría proporcionar respuestas irrelevantes para los intereses de la empresa. Por ejemplo, una consulta como «Cuéntame un chiste.Un agente de una aplicación (por ejemplo, una plataforma educativa) no debería responder a esta pregunta, ya que no redundaría en beneficio de la plataforma (por ejemplo, debido a consideraciones de costes). Cuéntame un chiste. El ejemplo es bastante sencillo, y una respuesta de CA podría evitarse mediante una restricción explícita en la solicitud. Los límites de lo que un chatbot empresarial debería o no debería responder pueden ser más complejos y difíciles de definir explícitamente en una solicitud simple. RAG ofrece la posibilidad de hacerlo implícitamente al restringir las respuestas del chatbot al dominio cubierto por los documentos utilizados para crear la base de datos vectorial, junto con una solicitud adecuada. Considere la plantilla de aumento que se muestra en la figura.2Esta plantilla se puede usar para restringir la respuesta de un modelo lineal de aprendizaje (LLM) a un contexto particular especificado en la plantilla. En otras palabras, la plantilla de aumento define los límites en referencia a la base de datos y puede Por lo tanto, permanecen iguales cuando la base de datos cambia (por ejemplo, cuando una organización agrega datos adicionales a la base de datos vectorial). En nuestra plantilla de aumento de ejemplo en la Fig.2 «Proporcione la información de contexto y responda a la consulta». La respuesta del LLM debe generarse únicamente dentro del contexto especificado. Suponiendo que el contexto sea una base de datos con información de clientes, el LLM generará respuestas basadas en dichos documentos.

Finalmente, una arquitectura RAG también conlleva una reducción. costo inicial de propiedad,Dado que la creación de una base de datos vectorial requiere menos recursos computacionales que el ajuste fino de un modelo base, RAG se presenta como una alternativa potencial al ajuste fino de LLM (Tabla 1).2).

3.3 Desafíos amplificados por RAG

Junto con estas oportunidades, RAG también presenta nuevos desafíos para las organizaciones. En su nivel más fundamental, la mayoría de los desafíos de RAG se relacionan congestión de datosy Capacidades de operaciones de aprendizaje automático (MLOps).La gestión de datos se ha identificado como un desafío importante en la investigación de SI (Abbasi et al.).2016En general, dado que los sistemas basados en RAG requieren esfuerzos adicionales para fusionar datos de fuentes heterogéneas, estos desafíos se ven agravados. Por lo tanto, las organizaciones necesitan desarrollar capacidades adicionales para abordar esta necesidad. Conceptos modernos como las estructuras de malla de datos (Dehghani)2022; Blohm y otros.2024También puede considerarse un enfoque útil para desarrollar sistemas basados en RAG. Además, se requiere un gran esfuerzo para garantizar la alta calidad de los datos adicionales. En la práctica, puede haber casos en los que las fuentes de datos contengan información contrafactual o incluso falsa que deba eliminarse. Por lo tanto, las organizaciones necesitan asignar más recursos y desarrollar nuevas capacidades, como las de MLOps, para implementar sistemas basados en RAG.

Además de los problemas de gestión de datos, los datos subyacentes inevitablemente presentan nuevos desafíos en términos de datos no deseados.inclinaciónefectos. Esto se debe a que las organizaciones cuentan con una nueva forma de incorporar datos a la infraestructura de IA mediante una base de datos vectorial. Podría decirse que esto es similar al ajuste fino de modelos lineales de aprendizaje (LLM), donde la organización también debe ser consciente de los nuevos sesgos no deseados. Sin embargo, difiere del uso de modelos base o soluciones de software como servicio (SaaS), donde la organización no puede influir en los datos utilizados para entrenar o ajustar el modelo. Un ejemplo conocido es el uso de documentos occidentales, que probablemente reflejen únicamente una perspectiva occidental y pueden resultar indeseables en un contexto internacional. Esto forma parte de un área más amplia de investigación en curso relacionada con la prevención y corrección de sesgos (p. ej., Mehrabi et al.).2022Gallegos y otros. 2024), un área que también está regulada por la Unión Europea (véase, por ejemplo, la Ley de IA de la UE).

[4]Los modelos de cimentación en sí mismos no proporcionan tales referencias. Los resultados se denominansin conexión a tierra (Magesh y otros.2024Cabe destacar que también existen referencias que no admiten la generación de una salida. Estas se denominanmal fundamentado.

Tabla 2Resumen de las oportunidades que ofrecen las arquitecturas basadas en RAG

Oportunidades	Descripción
Comprensión contextualizada mejorada	Las arquitecturas RAG utilizan la secuencia de entrada.incógnitapara recuperar documentos de textozy utilizarlos como contexto adicional'' (Lewis et al.2020, pág. 2). Por lo tanto, un sistema basado en RAG tiene una comprensión contextual más amplia en comparación con los modelos de base.
Reducción de las alucinaciones y mejora de la precisión de los hechos	Al incluir un contexto ampliado, un sistema basado en RAG puede generar resultados basados en el contexto, lo que reduce las alucinaciones y aumenta la precisión fáctica (Lewis et al.).2020Shuster y otros.2021)
Toma de tierra	La salida generada se combina con referencias a los datos contextuales (es decir, la fundamentación (Magesh et al.). 2024Esto permite a los usuarios verificar la salida y obtener información adicional siguiendo la referencia.
Límites del dominio del conocimiento para los LLM	Los sistemas basados en RAG pueden utilizarse para especificar el dominio que resulta de interés para los datos proporcionados. De este modo, se pueden excluir dominios irrelevantes o indeseables.
Costo inicial de propiedad	Desarrollar una base de datos vectorial es mucho más rentable en comparación con el ajuste fino en lo que respecta al costo total de propiedad (Balaguer et al.).2024)

Una arquitectura RAG básica (ver Fig.2) también sufre de lo que nos gustaría llamar el ''''Efecto de trozos con anteojera'' (Supongamos que extraemos un párrafo (fragmento) de un documento de texto extenso, como un libro de Harry Potter. ¿Hasta qué punto podría una persona comprender ese párrafo sin haber leído la novela completa? Es probable que se note una falta de comprensión, sobre todo en lo que respecta a términos propios del contexto de las novelas y la trama principal. Si bien este puede ser un ejemplo extremo debido a la magnitud del universo imaginario, que incluye magia y personajes ficticios, el principio de comprensión integral (BCE) también se aplica a documentos contextualizados en aplicaciones empresariales. Por lo tanto, el uso de RAG con datos enriquecidos aún presenta limitaciones en cuanto a la comprensión integral. En estos casos, los avances recientes, incluyendoRAPTOR (Sarthi et al.2024) y GraphRAG (Edge et al.2024), debería tenerse en cuenta.

El rendimiento del recuperador depende de un sistema de clasificación eficaz, es decir, un mecanismo capaz de identificar los documentos más relevantes. Generalmente, este sistema se basa en el principio de clasificación probabilística (Robertson).1977), lo cual no siempre es ideal. Por esa razón, se pueden considerar nuevos enfoques para mejorar los resultados de la clasificación. Los enfoques actuales incluyen modelos de clasificación invariantes a la permutación (Pang et al.2020), reordenamientos teniendo en cuenta las listas y búsquedas híbridas (Bruch et al. 2023) (Mesa3).

4 Implicaciones para los investigadores de BISE

Este artículo introductorio busca proporcionar una visión general fundamental de RAG, resaltar las características de las arquitecturas RAG y describir sus implicaciones. Dado que anteriormente El trabajo ya ha identificado importantes vías de investigación con modelos de fundamentos (Schneider et al.).2024; Feuerriegel y otros.2024), ilustramos algunas preguntas de investigación matizadas que surgen en combinación con RAG desde tres perspectivas diferentes: (1)organizativo, (2) individual,y (3) económico (ver figura.3).

En primer lugar, el uso de arquitecturas basadas en RAG tiene implicaciones para las organizaciones. Al igual que con las arquitecturas y paradigmas de software anteriores, las organizaciones necesitan nuevas habilidades para implementar y aprovechar las nuevas tecnologías de IA (Berente et al.).2021Esto es particularmente cierto para la arquitectura de TI y las capacidades de gestión de datos, que representan un gran desafío para las organizaciones (Abbasi et al.).2016; Blohm y otros.2024). Además de deficiencias bien conocidas como la centralización de la gestión de datos (Velu et al.2013Una arquitectura basada en RAG plantea nuevos retos que las organizaciones deben abordar. En particular, deben determinar cómo utilizarán sus datos. Por ejemplo, el conjunto de datos de una organización puede utilizarse para el ajuste fino de modelos (es decir, el ajuste fino de LLM), el desarrollo de bases de datos vectoriales (es decir, RAG) o ambos (por ejemplo, mediante RAFT). Desde una perspectiva teórica, deben identificarse las configuraciones adecuadas (Park y Mithas).2020) que guían a las organizaciones sobre cómo organizar sus datos para un ajuste fino, RAG o ambos. Además, identificar las ventajas y desventajas y las configuraciones preferidas es un desafío, ya que depende en gran medida de factores contextuales y ambientales como el tamaño de la organización o el sector. Por esta razón, también plantea interrogantes sobre el desarrollo de capacidades internas frente a externas (Nevo et al.).2007Las siguientes preguntas de investigación son ejemplos de investigaciones realizadas por académicos de BISE a nivel organizacional:¿Los altos niveles de capacidad de gestión de datos, por ejemplo, Data Mesh incluyendo RAG, conducen a altos niveles de rendimiento organizacional? ¿Cuál es el equilibrio óptimo entre los datos que se envían a RAG y los datos finos?

Tabla 3Resumen de los desafíos importantes exacerbados por RAG

desafíos	Descripción
Se requieren capacidades adicionales de gestión de datos/MLOps	Los sistemas basados en RAG requieren la inclusión de fuentes de datos heterogéneas y potencialmente dinámicas. Por lo tanto, se necesitan nuevas capacidades de gestión de datos para satisfacer esta necesidad. Nuevos conceptos, como las estructuras de malla de datos, resultan potencialmente útiles para los sistemas basados en RAG (Dehghani).2022; Blohm y otros.2024).
Nuevos sesgos potenciales	Con una comprensión contextual más amplia mediante nuevos datos, también existe el riesgo de introducir nuevos sesgos que requieren esfuerzos adicionales para evitar la propagación de dichos sesgos en los sistemas basados en LLM y RAG. Para una revisión actualizada, véase, por ejemplo, Mehrabi et al.2022) o Gallegos et al. (2024).
Efecto de fragmento cegado (BCE)	Una implementación estándar de RAG tiene limitaciones para comprender de forma integral grandes volúmenes de datos. Nuevos enfoques comoRAPTOR (Sarthi y otros.2024) son necesarias para reducir este problema.
Eficacia de la recuperación	La eficacia de una arquitectura RAG depende de la eficacia de su mecanismo de recuperación. Esto incluye la eficacia de la clasificación de documentos (es decir, ¿se clasifican primero los documentos más relevantes?) y el rendimiento del proceso de recuperación.

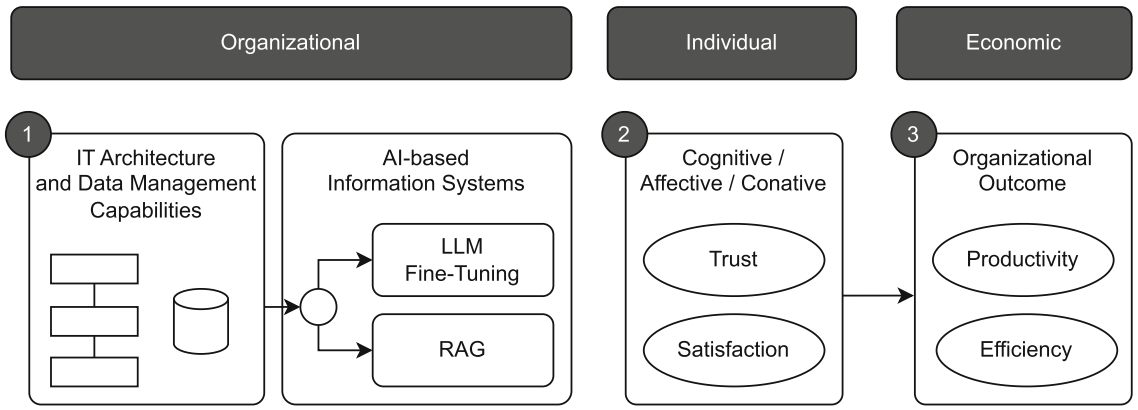

Figura 3Preguntas de investigación relacionadas con RAG

¿Un ajuste que conduzca a un rendimiento organizacional superior? o¿En qué medida una arquitectura basada en RAG contribuye a una mejor alineación de las TI con el negocio?

En segundo lugar, las personas que interactúan con sistemas basados en RAG (por ejemplo, mediante un CA) experimentarán cambios en la forma en que se presentan los resultados. Lo más importante es que RAG ofrece la posibilidad de añadir referencias a datos contextuales, lo que se ha denominado «fundamentación» (Magesh et al.).2024Proporcionar referencias a los datos contextuales está estrechamente relacionado con el concepto de IA explicable (XAI) (Schneider).2024; Longo y otros.2024), porque los usuarios obtienen información adicional sobre los resultados de un LLM. Hasta ahora, la literatura previa ha utilizado diferentes enfoques para proporcionar explicaciones post hoc, incluido LIME (Ribeiro et al.2016), valores de Shapley (SHAP) (Lundberg y Lee2017), o mapeo de activación de clase ponderado por gradiente (Grad-CAM) (Selvaraju et al. 2017Estos enfoques añaden una capa adicional que proporciona explicaciones visuales o textuales. Por ejemplo, SHAP se puede usar para resaltar partes de una imagen que influyen de manera importante en la predicción de la IA. Proporcionar referencias, como

Las citas en un libro, al relacionarse con los datos contextuales, pueden considerarse un enfoque alternativo y complementario para brindar una capa explicativa adicional a los usuarios. Diversos determinantes del comportamiento, incluidos los constructos cognitivos, afectivos y conativos (CAC) (Bagozzi),1992), puede verse influenciado. Esto es similar a investigaciones previas que han investigado la relación entre XAI y constructos latentes como la confianza (Hamm et al.2023) o intención de usar (Meske y Bunde2022Sostenemos que el impacto del enraizamiento aún no se ha investigado a fondo y se requieren más datos empíricos para determinar si constituye una valiosa adición a la IA explicable (XAI). Además, es necesario explorar en qué medida el enraizamiento influye en constructos percibidos, como la explicabilidad percibida o las intenciones de confianza percibidas. Además de explorar constructos latentes, los sistemas basados en RAG también tienen el potencial de reducir el tiempo real de recuperación, lo que a su vez puede mejorar el rendimiento del usuario. Dado que los sistemas basados en RAG se utilizan comúnmente como base para un análisis de contingencia (CA), representan una alternativa de vanguardia a los sistemas basados en conocimiento más tradicionales, como una intranet o

Wikipedia puede ser más eficaz para encontrar información relevante. Algunos ejemplos de preguntas de investigación que surgen con los sistemas basados en RAG relevantes para la investigación BISE a nivel individual son:¿Pueden las explicaciones basadas en la fundamentación superar los enfoques tradicionales de XAI a la hora de mejorar la confianza del usuario? o¿En qué medida RAG mejora el rendimiento laboral de los individuos?

Finalmente, RAG también invita a realizar más investigaciones que analicen el valor económico de las nuevas arquitecturas de sistemas de información. En términos más generales, se necesita investigación que analice qué resultados se pueden esperar de la evolución y los avances de las nuevas arquitecturas (Haki et al.). 2020En última instancia, las organizaciones buscan oportunidades para aumentar la eficiencia y la productividad. La metodología RAG tiene el potencial de mejorar los procesos de negocio y la toma de decisiones organizacionales. Sin embargo, aún no está claro hasta qué punto las organizaciones pueden beneficiarse del uso de RAG. Además, RAG también puede ayudar a las organizaciones a cumplir con los requisitos regulatorios. Por ejemplo, la Ley Europea de IA exige mayor transparencia en el uso de la IA. De nuevo, el concepto de fundamentación tiene el potencial de contribuir a este requisito y ofrecer una posible vía para las organizaciones. Por estas razones, las siguientes preguntas de investigación son ejemplos orientados a los negocios.BISEinvestigadores:¿Cómo pueden las organizaciones lograr una ventaja competitiva con sistemas basados en RAG?y¿En qué medida pueden los sistemas basados en RAG mejorar el cumplimiento de los requisitos reglamentarios?

Expresiones de gratitudLos autores desean expresar su sincero agradecimiento a la editora del departamento, Christine Legner, por su orientación constructiva a lo largo del proceso de revisión, y a los dos revisores anónimos por sus reflexivos comentarios que fortalecieron significativamente este trabajo.

FondosLa financiación de acceso abierto fue facilitada y organizada por Projekt DEAL.

Referencias

Bran M, Cox AS, Schilter O, Baldassari C, White AD, Schwaller P (2024) Ampliación de grandes modelos de lenguaje con herramientas químicas. Nat Mach Intell 6:525–535.https://doi.org/10.1038/s42256-024-00832-8

Ding N, Qin Y, Yang G, Wei F, Yang Z, Su Y, Sun M (2023) Ajuste fino y eficiente de parámetros de modelos de lenguaje preentrenados a gran escala. Nat Mach Intell 5(3):220–235.https://doi.org/10.1038/s42256-023-00626-4

Abbasi A, Sarker S, Chiang R (2016) Investigación sobre macrodatos en información sistemas: Hacia una agenda de investigación inclusiva. J Assoc Inf Syst 17(2):3.https://doi.org/10.17705/1jais.00423

Alavi M, Leidner DE (2001) Gestión del conocimiento y conocimiento Sistemas de gestión: Fundamentos conceptuales y cuestiones de investigación. MIS Q 25(1):107–136.https://doi.org/10.2307/3250961

Alavi M, Leidner DE, Mousavi R (2024) Gestión del conocimiento Perspectiva de la inteligencia artificial generativa. J Assoc Inf Syst 25(1):1–12.https://doi.org/10.17705/1jais.00859

Amri S, Bani R, Bani S (2024) Un enfoque para el análisis de Documentos financieros utilizando IA generativa. En: Actas de la 7ª conferencia internacional sobre redes, sistemas inteligentes y seguridad, ACM, Meknes, pp 1–5.https://doi.org/10.1145/3659677.3659736

Bagozzi RP (1992) La autorregulación de actitudes, intenciones y comportamiento. Soc Psychol Q 55(2):178.https://doi.org/10.2307/2786945

Balaguer A, Benara V, Cunha RLdF, Filho RdME, Hendry T, Holstein D, Marsman J, Mecklenburg N, Malvar S, Nunes LO, Padilha R, Sharp M, Silva B, Sharma S, Aski V, Chandra R (2024) RAG vs ajuste fino: oleoductos, compensaciones y un estudio de caso sobre agricultura.https://doi.org/10.48550/ARXIV.2401.08406 Berente N, Gu B, Recker J, Santhanam R (2021) Gestión de artificiales inteligencia. MIS Q 45(3):1433–1450.https://doi.org/10.25300/MISQ/2021/16274

Blohm I, Wortmann F, Legner C, Köbler F (2024) Productos de datos, datos Malla y tejido de datos: ¿Nuevo(s) paradigma(s) para datos y análisis? Bus Inf Syst Eng 66:643–652.https://doi.org/10.1007/s12599-024-00876-5

Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A, Agarwal S, Herbert-Voss A, Krueger G, Henighan T, Child R, Ramesh A, Ziegler DM, Wu J, Winter C, Hesse C, Chen M, Sigler E, Litwin M, Gray S, Chess B, Clark J, Berner C, McCandlish S, Radford A, Sutskever I, Amodei D (2020) Los modelos de lenguaje aprenden con pocos ejemplos. En: Actas de la 34.ª conferencia internacional sobre sistemas de procesamiento de información neuronal. Curran Associates Inc., Red Hook, NY, EE. UU., NIPS '20, págs. 1877-1901.https://doi.org/10.5555/3495724.3495883

Bruch S, Gai S, Ingber A (2023) Un análisis de las funciones de fusión para Recuperación híbrida. ACM Trans Inf Syst 42(1):1–35.https://doi.org/10.1145/3596512

Cleverley PH, Burnett S (2019) Búsqueda y descubrimiento empresarial capacidad: Los factores y mecanismos generativos para la satisfacción del usuario. J Inf Sci 45(1):29–52.https://doi.org/10.1177/ 0165551518770969

Dehghani Z (2022) Malla de datos que ofrece valor basado en datos a escala. O'Reilly, Sebastopol

Edge D, Trinh H, Cheng N, Bradley J, Chao A, Mody A, Truitt S, Larson J (2024) De lo local a lo global: un enfoque RAG de grafos para la generación de resúmenes centrados en consultas.https://doi.org/10.48550/arXiv.2404.16130

Feuerriegel S, Dolata M, Schwabe G (2020) Feria de IA: desafíos y oportunidades. Bus Inf Syst Eng 62(4):379–384.https://doi.org/10.1007/s12599-020-00650-3

Feuerriegel S, Hartmann J, Janiesch C, Zschech P (2024) Generativo IA. Bus Inf Syst Eng 66(1):111–126.https://doi.org/10.1007/s12599-023-00834-7

Gallegos IO, Rossi RA, Barrow J, Tanjim MM, Kim S, Dernoncourt F, Yu T, Zhang R, Ahmed NK (2024) Sesgo y equidad en modelos de lenguaje grandes: una revisión. Comp Linguist 50(3):1–79.https://doi.org/10.1162/coli_a_00524

Gao L, Biderman S, Black S, Golding L, Hoppe T, Foster C, Phang J, He H, Thite A, Nabeshima N, Presser S, Leahy C (2020) The pile: Un conjunto de datos de 800 GB de texto diverso para el modelado del lenguaje. https://doi.org/10.48550/arXiv.2101.00027

Gao Y, Xiong Y, Gao X, Jia K, Pan J, Bi Y, Dai Y, Sun J, Wang M, Wang H (2023) Generación aumentada por recuperación para modelos de lenguaje grandes: una revisión.https://doi.org/10.48550/ARXIV.2312.10997

Gekhman Z, Yona G, Aharoni R, Eyal M, Feder A, Reichart R, Herzig J (2024) ¿El ajuste fino de los LLM en base a nuevos conocimientos fomenta las alucinaciones?https://doi.org/10.48550/arXiv.2405.05904

Guerreiro NM, Alves DM, Waldendorf J, Haddow B, Birch A, Colombo P, Martins AFT (2023) Alucinaciones en modelos de traducción multilingües de gran tamaño. Trans Assoc Comput Linguist 11:1500–1517.https://doi.org/10.1162/tacl_a_00615

Guu K, Lee K, Tung Z, Pasupat P, Chang MW (2020) REINO: Preentrenamiento de modelos de lenguaje con recuperación de información. En: Actas de la 37.ª conferencia internacional sobre aprendizaje automático, JMLR.org, ICML'20.https://doi.org/10.5555/3524938.3525306 Haki K, Beese J, Aier S, Winter R (2020) La evolución de Arquitectura de sistemas de información: un modelo de simulación basado en agentes. MIS Q 44(1):155–184.https://doi.org/10.25300/MISQ/2020/14494

Hamm P, Klesel M, Coberger P, Wittmann HF (2023) Explicación asuntos: Un estudio experimental sobre IA explicable. Electron Mark 33(1):17.https://doi.org/10.1007/s12525-023-00640-9 Hicks MT, Humphries J, Slater J (2024) ChatGPT es una tontería. Ética Inf Technol 26(2):38.https://doi.org/10.1007/s10676-024-09775-5

Hu EJ, Shen Y, Wallis P, Allen-Zhu Z, Li Y, Wang S, Wang L, Chen W (2022) LoRA: Adaptación de bajo rango de modelos de lenguaje grandes. En: Conferencia internacional sobre representaciones de aprendizaje, conferencia virtual

Jaber R, Zhong S, Kuoppamäki S, Hosseini A, Gessinger I, Brumby DP, Cowan BR, Mcmillan D (2024) Cocinando con agentes: Diseño de interacción de voz sensible al contexto. En: Actas de la conferencia CHI sobre factores humanos en sistemas informáticos, ACM, Honolulu, págs. 1-13.https://doi.org/10.1145/3613904.3642183

Ji Z, Lee N, Frieske R, Yu T, Su D, Xu Y, Ishii E, Bang YJ, Madotto A, Fung P (2023) Estudio de la alucinación en la generación del lenguaje natural. ACM Comput Surv 55(12):1–38.https://doi.org/10.1145/3571730

Jiang AQ, Sablayrolles A, Mensch A, Bamford C, Chaplot DS, Casas Ddl, Bressand F, Lengyel G, Lample G, Saulnier L, Lavaud LR, Lachaux MA, Stock P, Scao TL, Lavril T, Wang T, Lacroix T, Sayed WE (2023) Mistral 7B.https://doi.org/10.48550/arXiv.2310.06825

Karpukhin V, Oguz B, Min S, Lewis P, Wu L, Edunov S, Chen D, Yih Wt (2020) Recuperación de pasajes densos para preguntas y respuestas de dominio abierto. En: Webber B (ed) Actas de la conferencia de 2020 sobre métodos empíricos en procesamiento del lenguaje natural (EMNLP), Asociación de Lingüística Computacional, en línea, págs. 6769–6781.https://doi.org/10.18653/v1/2020.emnlp-main.550

Kazemitabaar M, Ye R, Wang X, Henley AZ, Denny P, Craig M, Grossman T (2024) CodeAid: Evaluación de la implementación en el aula de un asistente de programación basado en LLM que equilibra las necesidades de estudiantes y docentes. En: Actas de la conferencia CHI sobre factores humanos en sistemas informáticos, ACM, Honolulu, págs. 1-20.https://doi.org/10.1145/3613904.3642773 Khan AA, Hasan MT, Kemell KK, Rasku J, Abrahamsson P (2024) Desarrollo de sistemas LLM basados en generación aumentada por recuperación (RAG) a partir de PDF: un informe de experiencia.https://doi.org/10.48550/ARXIV.2410.15944

Lewis P, Pérez E, Piktus A, Petroni F, Karpukhin V, Goyal N, Küttler H, Lewis M, Yih Wt, Rocktäschel T, Riedel S, Kiela D (2020) Generación aumentada por recuperación para tareas de PLN intensivas en conocimiento. En: Actas de la 34.ª conferencia internacional sobre sistemas de procesamiento de información neuronal, Curran, Red Hook, NIPS '20, págs. 9459-9474.https://doi.org/10.5555/3495724.3496517 Lins S, Pandl KD, Teigeler H, Thiebes S, Bayer C, Sunyaev A (2021) Inteligencia artificial como servicio. Bus Inf Syst Eng 63(4):441–456.https://doi.org/10.1007/s12599-021-00708-w

Liu Y, Peng X, Zhang X, Liu W, Yin J, Cao J, Du T (2024) RA-ISF: Aprender a responder y comprender a partir del aumento de la recuperación mediante la retroalimentación iterativa autoadministrada.https://doi.org/10.18653/v1/2024.findings-acl.281

Longo L, Brcic M, Cabitza F, Choi J, Confalonieri R, Ser JD, Guidotti R, Hayashi Y, Herrera F, Holzinger A, Jiang R, Khosravi H, Lecue F, Malgieri G, Páez A, Samek W, Schneider J, Speith T, Stumpf S (2024) Inteligencia artificial explicable (XAI) 2.0: un manifiesto de desafíos abiertos y direcciones de investigación interdisciplinarias. Inf Fusión 106:102301.https://doi.org/10.1016/j. infusor.2024.102301

Lundberg SM, Lee SI (2017) Un enfoque unificado para la interpretación Predicciones del modelo. En: Actas de la 31ª conferencia internacional sobre sistemas de procesamiento de información neuronal, Curran, Red Hook, NIPS'17, págs. 4768–4777.https://doi.org/10.5555/3295222.3295230

Magesh V, Surani F, Dahl M, Suzgun M, Manning CD, Ho DE (2024) ¿Libre de alucinaciones? Evaluación de la fiabilidad de las principales herramientas de investigación jurídica (IA).https://doi.org/10.48550/arXiv.2405.20362 Maynez J, Narayan S, Bohnet B, McDonald R (2020) Sobre la fidelidad y la factualidad en la síntesis abstractiva. En: Actas de la 58.ª reunión anual de la Asociación de Lingüística Computacional, Asociación de Lingüística Computacional, en línea, págs. 1906-1919.https://doi.org/10.18653/v1/2020.acl-main.173

Mehrabi N, Morstatter F, Saxena N, Lerman K, Galstyan A (2022) A Estudio sobre sesgo y equidad en el aprendizaje automático. ACM Comput Surv 54(6):1–35.https://doi.org/10.1145/3457607

Meske C, Bunde E (2022) Principios de diseño para interfaces de usuario en IA-Sistemas de apoyo a la decisión basados en datos: el caso de la detección explicable del discurso de odio. Inf Syst Front.https://doi.org/10.1007/s10796-021-10234-5

Nevo S, Wade MR, Cook WD (2007) Un examen de la compensación entre capacidades de TI internas y externas. J Strat Inf Syst 16(1):5–23.https://doi.org/10.1016/j.jsis.2006.10.002

Pal A, Umapathi LK, Sankarasubbu M (2023) Med-HALT: médico Prueba de alucinaciones de dominio para modelos de lenguaje grandes. https://doi.org/10.48550/arXiv.2307.15343

Pang L, Xu J, Ai Q, Lan Y, Cheng X, Wen J (2020) SetRank: aprendizaje Un modelo de clasificación invariante a la permutación para la recuperación de información. En: Actas de la 43.ª conferencia internacional ACM SIGIR sobre investigación y desarrollo en recuperación de información, ACM, evento virtual China, págs. 499-508. https://doi.org/10.1145/3397271.3401104

Park Y, Mithas S (2020) Complejidad organizada de los negocios digitales estrategia: una perspectiva configuracional. MIS Q 44(1):85–127. https://doi.org/10.25300/MISQ/2020/14477

Pérez E, Karamcheti S, Fergus R, Weston J, Kiela D, Cho K (2019) Búsqueda de evidencia generalizable mediante el aprendizaje de modelos de preguntas y respuestas convincentes. En: Inui K (ed.) Conferencia sobre métodos empíricos en procesamiento del lenguaje natural y 9.ª conferencia internacional conjunta sobre procesamiento del lenguaje natural, Hong Kong. https://doi.org/10.18653/v1/D19-1244

Rai A, Chen L, Breazeal C, Ramesh B, Long Y, Aria A (2024) Diseño y atributos de evaluación para la personalización escalable y rentable de tutores de máster en derecho (LLM) en la enseñanza de programación. En: Actas de ICIS 2024

Ribeiro MT, Singh S, Guestrin C (2016) "¿Por qué debería confiar en ti?": Explicación de las predicciones de cualquier clasificador. En: Actas de la 22.ª Conferencia Internacional ACM SIGKDD sobre Descubrimiento de Conocimiento y Minería de Datos. Association for Computing Machinery, Nueva York, KDD '16, págs. 1135-1144.https://doi.org/10.1145/2939672.2939778

Robertson S (1977) El principio de ordenación probabilística en RI. J Doc 33(4):294–304.https://doi.org/10.1108/eb026647 Rumelhart DE, Hinton GE, Williams RJ (1986) Representación del aprendizaje estaciones mediante la retropropagación de errores. Nature 323(6088):533–536. https://doi.org/10.1038/323533a0

Sarthi P, Abdullah S, Tuli A, Khanna S, Goldie A, Manning CD (2024) RAPTOR: Procesamiento abstractivo recursivo para la recuperación organizada en árbol.https://doi.org/10.48550/arXiv.2401.18059 Schneider J (2024) Inteligencia artificial generativa explicable (GenXAI): una agenda de encuesta, conceptualización e investigación. Artif Intell Rev. 57(11):289.https://doi.org/10.1007/s10462-024-10916-x

Schneider J, Meske C, Kuss P (2024) Modelos de base: un nuevo paradigma para la inteligencia artificial. Bus Inf Syst Eng 66(2):221–231.https://doi.org/10.1007/s12599-024-00851-0

Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D (2017) Grad-CAM: Explicaciones visuales a partir de redes profundas mediante localización basada en gradientes. En: 2017 IEEE international conference on computer vision (ICCV), pp 618–626,https://doi.org/10.1109/ICCV.2017.74

Shuster K, Poff S, Chen M, Kiela D, Weston J (2021) Recuperación La amplificación reduce las alucinaciones en la conversación.https://doi.org/10.48550/arXiv.2104.07567

Steck H, Ekanadham C, Kallus N (2024) ¿Es la similitud coseno de ¿Las incrustaciones realmente se basan en la similitud? En: Actas complementarias de la conferencia web de la ACM 2024, ACM, Singapur, págs. 887–890.https://doi.org/10.1145/3589335.3651526

Strobel G, Banh L (2024) ¿Qué dijo el médico? Empoderamiento Comprensión del paciente mediante inteligencia artificial generativa. En: Actas de ECIS 2024, Pafos.

Touvron H, Martin L, Stone K, Albert P, Almahairi A, Babaei Y, Bashlykov N, Batra S, Bhargava P, Bhosale S, Bikel D, Blecher L, Ferrer CC, Chen M, Cucurull G, Esiobu D, Fernandes J, Fu J, Fu W, Fuller B, Gao C, Goswami V, Goyal N, Hartshorn A, Hosseini S, Hou R, Inan H, Kardas M, Kerkez V, Khabsa M, Kloumann I, Korenev A, Koura PS, Lachaux MA, Lavril T, Lee J, Liskovich D, Lu Y, Mao Y, Martinet X, Mihaylov T, Mishra P, Molybog I, Nie Y, Poulton A, Reizenstein J, Rungta R, Saladi K, Schelten A, Silva R, Smith EM, Subramanian R, Tan XE, Tang B, Taylor R, Williams A, Kuan JX, Xu P, Yan Z, Zarov I, Zhang Y, Fan A, Kambadur M, Narang S, Rodriguez A, Stojnic R, Edunov S, Scialom T (2023) Llama 2: base abierta y modelos de chat optimizados.https://doi.org/10.48550/arXiv.2307.09288

Trinh TH, Wu Y, Le QV, He H, Luong T (2024) Resolviendo la olimpiada geometría sin demostraciones humanas. Nature 625(7995):476–482.https://doi.org/10.1038/s41586-023-06747-5

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I (2017) La atención es todo lo que necesitas. En: Actas de la 31ª conferencia internacional sobre sistemas de procesamiento de información neuronal, Curran, Red Hook, NIPS'17, págs. 6000-6010.https://doi.org/10.5555/3295222.3295349

Velu CK, Madnick SE, Van Alstyne MW (2013) Centralización de datos Gestión con consideraciones de incertidumbre y flexibilidad basada en la información. J Manag Inf Syst 30(3):179–212.https://doi.org/10.2753/MIS0742-1222300307

Veturi S, Vaichal S, Jagadheesh RL, Tripto NI, Yan N (2024) RAG Sistema de predicción de respuestas contextuales basado en preguntas y respuestas.https://doi.org/10.48550/ARXIV.2409.03708

Wang S, Khramtsova E, Zhuang S, Zuccon G (2024a) FeB4RAG: Evaluación de la búsqueda federada en el contexto de la generación aumentada de recuperación. En: Actas de la 47.ª conferencia internacional ACM SIGIR sobre investigación y desarrollo en recuperación de información, ACM, Washington DC, EE. UU., págs. 763-773.https://doi.org/10.1145/3626772.3657853

Wang Y, Lipka N, Zhang R, Siu A, Zhao Y, Ni B, Wang X, Rossi R, Derr T (2024b) Aumento de la recuperación con reconocimiento de topología para la generación de texto. En: Actas de la 33.ª conferencia internacional de la ACM sobre gestión de la información y el conocimiento, ACM, Boise, págs. 2442-2452.https://doi.org/10.1145/3627673.3679746 Wang Z, Liu A, Lin H, Li J, Ma X, Liang Y (2024c) RAT: Recuperación Los pensamientos aumentados propician un razonamiento sensible al contexto en la generación de horizontes largos. https://doi.org/10.48550/ARXIV.2403.05313

Wei J, Wang X, Schuurmans D, Bosma M, Ichter B, Xia F, Chi EH, Le QV, Zhou D (2022) La inducción de la cadena de pensamiento provoca razonamiento en modelos de lenguaje grandes. En: Actas de la 36.ª conferencia internacional sobre sistemas de procesamiento de información neuronal, Curran, Red Hook, NY, EE. UU., NIPS '22, págs. 24824–24837. https://doi.org/10.5555/3600270.3602070

Wei Z, Huang D, Zhang J, Song C, Zhang S, Zhang J, Li Z, Jiang K, Li R, Duan Q (2024) GARAG: Un sistema general adaptativo de preguntas y respuestas basado en RAG. En: Actas de la conferencia internacional sobre computación en la nube y big data de 2024, Association for Computing Machinery, Nueva York, ICCBD '24, págs. 442–447.https://doi.org/10.1145/3695080.3695156 White RW (2024) Avanzando la frontera de la búsqueda con agentes de IA. Commun ACM 67(9):54–65.https://doi.org/10.1145/3655615

Wu Y, Tang B, Xi C, Yu Y, Wang P, Liu Y, Kuang K, Deng H, Li Z, Xiong F, Hu J, Cheng P, Wang Z, Wang Y, Luo Y, Yang M (2024) Xinyu: Un sistema eficiente basado en LLM para la generación de comentarios. En: Actas de la 30.ª conferencia ACM SIGKDD sobre descubrimiento de conocimiento y minería de datos. Association for Computing Machinery, Nueva York, KDD '24, págs. 6003-6014. https://doi.org/10.1145/3637528.3671537

Yu H, Gan A, Zhang K, Tong S, Liu Q, Liu Z (2024) Evaluación de Generación aumentada por recuperación: una revisión. En: Zhu W (ed) Actas de la conferencia internacional sobre computación en la nube y big data de 2024, Nueva York, ICCBD '24, págs. 442–447. https://doi.org/10.1145/3695080.3695156

Zhang T, Patil SG, Jain N, Shen S, Zaharia M, Stoica I, González JE (2024) RAFT: Adaptación del modelo de lenguaje a RAG específico del dominio.https://doi.org/10.48550/arXiv.2403.10131

Zhao P, Zhang H, Yu Q, Wang Z, Geng Y, Fu F, Yang L, Zhang W, Jiang J, Cui B (2024) Generación aumentada por recuperación para contenido generado por IA: una revisión.https://doi.org/10.48550/ARXIV.2402.19473