AUTO-REFLEXIÓN Y RAG EN MODELOS DE LENGUAJE PEQUEÑOS PARA CONOCIMIENTO EMPRESARIAL: UN ESTUDIO DE MAPEO SISTEMÁTICO

M.Sc. Alcides Yohacin Leaños Rodríguez
Posgrado SOE – UAGRM
https://orcid.org/0009-0008-3208-3898
Santa Cruz, Bolivia | alcides@dualbiz.net

https://doi.org/10.23670 FT.2026.1.42

Recibido 28/04/2026 - Aceptado 13/05/2026

RESUMEN

Las organizaciones requieren sistemas de inteligencia artificial capaces de razonar sobre bases de conocimiento internas sin comprometer la seguridad. Sin embargo, los modelos de lenguaje grandes (LLM) no acceden de forma nativa a información confidencial, y su despliegue local suele ser inviable debido a los costos de hardware. En este contexto, los modelos de lenguaje pequeños (SLM, 1B–13B parámetros) emergen como una alternativa viable, aunque su capacidad para soportar pipelines de generación aumentada por recuperación (RAG) con mecanismos de auto-reflexión en entornos empresariales aún no está plenamente establecida. Este estudio analiza mediante un estudio de mapeo sistemático (SMS) de la literatura en trabajos que emplean o integran componentes como RAG y auto-reflexión aplicados al soporte de conocimiento. Siguiendo las directrices de Kitchenham y Charters (2007) y Petersen et al. (2008), se examinaron publicaciones de ar Xiv, NeurIPS, IEEE Xplore y ACM entre 2020 y 2025. De un total de 510 resultados iniciales, se seleccionaron 40 estudios primarios; los resultados evidencian un creciente interés en arquitecturas que combinan recuperación densa con estrategias adaptativas basadas en incertidumbre, donde un alto porcentaje de los trabajos se construyen sobre la integración de componentes existentes. Asimismo, se identifican limitaciones en la evaluación, la eficiencia y la aplicabilidad en escenarios empresariales reales. Se concluye que estas técnicas muestran potencial para mejorar el razonamiento y la adaptación de los modelos, aunque persisten desafíos para su implementación en SLM bajo restricciones de privacidad y hardware. El estudio organiza el conocimiento existente y establece una base estructurada para futuras investigaciones en el contexto mencionado.

Palabras clave: RAG, Auto-Reflexión, Auto-correccion, Auto-induccion, LLM, SLM, Agentes autónomos, Agentes Inteligentes

ABSTRACT

Organizations require artificial intelligence systems capable of reasoning over internal knowledge bases without compromising security. However, large language models (LLMs) do not natively access confidential information, and their local deployment is often unfeasible due to hardware costs. In this context, small language models (SLMs, 1B–13B parameters) emerge as a viable alternative, although their capacity to support retrieval-augmented generation (RAG) pipelines with self-reflection mechanisms in enterprise environments has not yet been fully established. This study analyzes, through a systematic mapping study (SMS), the literature on works that employ or integrate components such as RAG and self-reflection. Following the guidelines of Kitchenham and Charters (2007) and Petersen et al. (2008), publications from arXiv, NeurIPS, IEEE Xplore, and ACM between 2020 and 2025 were examined. From an initial pool of 510 results, 40 primary studies were selected; the findings reveal a growing interest in architectures that combine dense retrieval with uncertainty-based adaptive strategies, where a high percentage of works are built upon the integration of existing components. Likewise, limitations are identified in evaluation, efficiency, and applicability in real enterprise scenarios. It is concluded that these techniques show potential for improving model reasoning and adaptation, although challenges persist for their implementation in SLMs under privacy and hardware constraints. The study organizes existing knowledge and establishes a structured foundation for future research.

Keywords: RAG, Self-Reflection, self-induction, LLM, SLM, autonomous agent, intelligent agent

INTRODUCCIÓN

La rápida adopción de la Inteligencia Artificial (IA) en entornos organizacionales ha generado una demanda creciente de sistemas inteligentes capaces de operar eficazmente sobre bases de conocimiento internas y específicas del dominio. En contextos industriales donde los datos operativos confidenciales, los procesos propietarios, la documentación regulatoria y el conocimiento institucional no están disponibles públicamente. Los Modelos de Lenguaje Grandes (LLMs, Large Language Models) de propósito general enfrentan una limitación estructural crítica: son entrenados sobre entornos de la nube y en algunos casos se ha visto el uso de manera pública información privada, por tanto existen brechas de seguridad o se torna complicado el acceso inherente a la información privada para una fácil retroalimentación, misma que se califica como sensible para la organización y que es la que impulsa la toma de decisiones cotidiana.

La Generación Aumentada por Recuperación conocida como RAG (Retreival Augmented Generation) ha ganado atención como mecanismo para anclar las respuestas de los modelos de lenguaje en documentos relevantes y actualizados recuperados de repositorios internos. Las arquitecturas RAG disocian el conocimiento paramétrico incrustado en los pesos del modelo del conocimiento fáctico almacenado en fuentes externas, permitiendo que los modelos generen respuestas informadas por contenido específico de la organización sin requerir re-entrenamiento o ajuste fino costosos. Sin embargo, si bien RAG extiende considerablemente el alcance funcional de los modelos de IA hacia dominios de conocimiento privados, introduce una nueva clase de problemas: el modelo puede recuperar información irrelevante o incompleta, no reconocer los límites de su propio conocimiento, o producir respuestas confiadas pero incorrectas, fenómeno conocido comúnmente como alucinación.

Complementario al RAG el mecanismo de auto-reflexión, la capacidad de un modelo de lenguaje para introspeccionar sobre su propio razonamiento, identificar inconsistencias y refinar iterativamente sus respuestas ha emergido como una dirección prometedora para mejorar la confiabilidad de las respuestas y la calibración epistémica. Los agentes auto-reflexivos, como los descritos en el marco Reflexión (Shinn et al., 2023) o mediante las arquitecturas Self-RAG (Asai, Wu, et al., 2024), son capaces de evaluar la calidad de la evidencia recuperada, detectar brechas lógicas en las respuestas generadas y desencadenar ciclos correctivos de recuperación o razonamiento. Estas capacidades son particularmente valiosas en entornos empresariales donde las respuestas incorrectas o mal fundamentadas pueden tener consecuencias operativas, legales o financieras.

Un desafío significativo y poco explorado concierne al despliegue de estos mecanismos en modelos de lenguaje pequeños con recursos limitados. Si bien los LLM's de frontera (ej. GPT-4, Claude 3, Gemini Ultra) poseen capacidad paramétrica suficiente para implementar comportamientos complejos de razonamiento y reflexión, las organizaciones frecuentemente operan bajo restricciones de recursos (hardware) que impiden el despliegue de modelos muy grandes. Los Modelos de Lenguaje Pequeños (SLM, Small languaje Models) modelos en el rango de 1B a 13B (1.000 a 13.000 millones de parámetros); son más viables para el despliegue en infraestructuras propias, con preservación de la privacidad y eficiencia de costos. No obstante, si estos modelos más pequeños pueden inducir eficazmente comportamientos auto-reflexivos y aprovechar pipelines RAG para el soporte de conocimiento interno sigue siendo una pregunta abierta y crítica. A pesar del creciente número de trabajos sobre arquitecturas RAG y agentes LLM auto-reflexivos, la intersección de estos mecanismos en el contexto de modelos de lenguaje pequeños o entornos con recursos limitados, orientados al conocimiento empresarial interno, sigue siendo un campo insuficientemente explorado. Los estudios de revisión existentes se centran principalmente en modelos de gran escala en contextos generales o en RAG como problema de recuperación de información, sin considerar las restricciones específicas de los despliegues industriales.

Las organizaciones de sectores como manufactura, finanzas, salud, logística y servicios profesionales gestionan grandes volúmenes de conocimiento interno, incluyendo documentación operativa, procesos, registros regulatorios e información histórica. En este contexto, el desarrollo de sistemas de inteligencia artificial capaces de acceder, interpretar y razonar sobre esta información de forma segura representa una necesidad crítica para este sector. Desde una perspectiva teórica, la integración de RAG con mecanismos de auto-reflexión en modelos pequeños plantea desafíos relevantes en términos de razonamiento, adaptación y eficiencia bajo restricciones de recursos. Desde el punto de vista práctico, esta combinación podría habilitar asistentes inteligentes desplegados localmente, preservando la privacidad de los datos y reduciendo la dependencia de infraestructuras externas. Asimismo, su pertinencia radica en la creciente demanda de soluciones de IA accesibles y seguras en entornos empresariales con limitaciones de hardware.

En este marco, la ausencia de evidencia consolidada justifica el desarrollo del presente estudio de mapeo sistemático SMS (Systematic Mapping Study), con el propósito de analizar la literatura existente sobre trabajos que emplean o integren componente de RAG y mecanismos de auto-reflexión en modelos de lenguaje aplicados al soporte de conocimiento. En particular, el estudio revisa propuestas, enfoques y tendencias relacionadas con el aprendizaje inductivo, la auto-reflexión y la generación aumentada, con el fin de identificar patrones, limitaciones y oportunidades de investigación en este campo; para posteriormente extrapolar esa información y establecer una línea sobre la aplicabilidad con los SML's y las bases de conocimiento.

Para lo cual se pretende:

(OE1) Caracterizar el panorama de métodos, técnicas y arquitecturas disponibles que permiten el aprendizaje inductivo y el comportamiento auto-reflexivo en agentes inteligentes que operan dentro de sistemas basados en RAG, especificaciones de entrada-salida y las métricas de desempeño utilizadas para evaluar los enfoques propuestos, con énfasis en aquellas aplicables a tareas de conocimiento específicas del dominio o de entornos empresariales.

(OE2) Clasificar los tipos de enfoques propuestos o utilizados, incluyendo modelos, marcos de trabajo, herramientas y guías y evaluar la madurez de sus contribuciones reportadas.

(OE3) Examinar los métodos, técnicas y herramientas existentes (p. ej., modelos LLM base, módulos de razonamiento) que se combinan o extienden dentro de las arquitecturas relevadas.

(OE4) Mapear las adaptaciones y mejoras propuestas sobre mecanismos existentes de recuperación, reflexión o aprendizaje, y evaluar su fundamentación teórica y empírica.

(OE5) Evaluar la relevancia y aplicabilidad de los métodos identificados en el entorno empresarial, particularmente para el soporte de sistemas de información interna y bases de datos de trabajo.

(OE6) Identificar limitaciones, desafíos abiertos y direcciones de investigación futura relacionados con el despliegue de sistemas RAG auto-reflexivos en modelos de lenguaje pequeños o con recursos limitados.

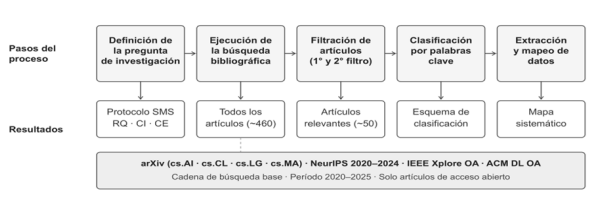
METODOLOGÍA

Diseño del estudio

Como se menciona anteriormente, el trabajo adopta la metodología de Estudio de Mapeo Sistemático SMS, siguiendo las directrices establecidas por Kitchenham y Charters (2007) y el protocolo extendido propuesto por Petersen et al., (2008). El SMS es particularmente adecuado para el propósito de esta investigación, dado que el campo de los agentes LLM auto-reflexivos con RAG es relativamente reciente y dinámico, donde se requiere una visión estructurada del estado del arte antes de proceder a una revisión sistemática de mayor profundidad. A diferencia de una Revisión Sistemática de Literatura (SLR, Systematic Literature Review), el mapeo sistemático no busca sintetizar evidencia cuantitativa o meta-analítica, sino categorizar y clasificar el espacio de investigación existente, identificar tendencias temporales y detectar brechas de conocimiento que orienten investigaciones futuras. La ejecución del estudio comprende cinco fases secuenciales: (1) definición del protocolo de búsqueda, (2) ejecución de la búsqueda y exportación de resultados, (3) primera filtración mediante lectura de título y resumen, (4) segunda filtración mediante lectura completa del texto, y (5) extracción de datos y análisis. El proceso completo es documentado en una hoja de trabajo estructurada que preserva la trazabilidad de cada decisión de inclusión y exclusión, como indica la siguiente figura.

Figura 1

Proceso de Estudio de Mapeo Sistemático (adaptado de Petersen et al., 2008)



Fuentes de información y cadenas de búsqueda

Dado que el acceso institucional a bases de datos de pago no se encuentra disponible en el contexto de la presente investigación, la selección de fuentes de información priorizó repositorios de acceso abierto (Open Access) con cobertura relevante para el campo de la inteligencia artificial y el procesamiento del lenguaje natural. Las fuentes seleccionadas son:

- arXiv.org (categorías cs.AI, cs.CL, cs.LG, cs.MA): repositorio primario de preprints en IA/NLP, donde se publican la mayoría de los trabajos relevantes al área antes de su aparición en conferencias y revistas.
- NeurIPS Proceedings (proceedings.neurips.cc): actas de la Conferencia sobre Sistemas de Procesamiento de Información Neural, disponibles en acceso abierto completo para los años 2020–2024. (2025 aún no se tiene información total de documentos Acceso Abierto, a fecha del presente estudio Mayo del 2026).
- IEEE Xplore con filtro de Acceso Abierto activo: publicaciones indexadas en IEEE Transactions on Artificial Intelligence y conferencias asociadas, restringidas a artículos con texto completo libremente accesible. (2025 aún no se tiene información total de documentos Acceso Abierto, a fecha del presente estudio Mayo del 2026).
- ACM Digital Library con filtro de Acceso Abierto activo: actas de conferencias ACM relevantes (AAAI, IJCAI, ACL, EMNLP, SIGKDD), restringidas a publicaciones con texto completo accesible sin suscripción.

La cadena de búsqueda base utilizada en todas las fuentes es la siguiente:

Search Strings: ("intelligent agent" OR "autonomous agent" OR "LLM agent" OR "AI agent") AND ("retrieval-augmented generation" OR "RAG" OR "retrieval augmented") AND ("self-reflection" OR "self-correction" OR "self-refinement" OR "self-critique" OR "introspection") AND ("inductive learning" OR "inductive reasoning" OR "learning from experience" OR "in-context learning")

Esta cadena fue adaptada según la sintaxis y los campos de búsqueda específicos de cada fuente. En arXiv se aplicó sobre título y resumen (ti_abs), en NeurIPS mediante búsqueda de texto completo en las actas. en IEEE Xplore y ACM DL mediante el campo All Metadata con el filtro de acceso abierto activado. No incluye SLM (Modelos de lenguaje pequeños) ni bases de conocimiento empresarial por que la idea central del estudio es extrapolar la información obtenida a este contexto.

Periodo de Búsqueda y Tipo de Publicación

La justificación temporal del período de búsqueda 2020–2025 se fundamenta en que el paradigma RAG fue formalmente introducido por Lewis et al., (2020) en NeurIPS 2020, constituyendo el punto de partida conceptual del campo. El límite superior de 2025 permite capturar el ciclo completo de consolidación del paradigma de agentes LLM y los trabajos más recientes sobre reflexión y aprendizaje inductivo en agentes pequeños, siendo mayo del 2026 la fecha de realización del SMS.

Período: 2020-2025. El límite inferior coincide con la introducción formal del paradigma RAG (expuesto anteriormente); el límite superior captura el aumento total de la investigación de agentes LLM 2025(NeurIPS y IEEE Explore restringen aún el Acceso Abierto a la gestión 2025 hasta mayo del 2026 fecha de realización del estudio).

Tipo de publicación: Artículos de revistas revisado por pares, artículos de conferencias y artículos de talleres.

Idioma: Sólo inglés.

Criterios de inclusión y exclusión

La selección de estudios se rige por los criterios definidos en el protocolo y detallados en la Tabla 1.

Tabla 1

Criterios de inclusión y exclusión del estudio de mapeo sistemático

Criterio	Descripción
CI1	El estudio aborda aprendizaje inductivo, aprendizaje desde contexto recuperado, o auto-mejora en agentes inteligentes.
CI2	El estudio propone o analiza una arquitectura o método basado en RAG aplicable a sistemas de agentes inteligentes.
CE3	El estudio es únicamente un resumen (sin texto completo disponible).
CI4	El estudio presenta técnicas de razonamiento, planificación o memoria que habiliten la adaptabilidad o generalización del agente.
CI5	El estudio reporta resultados empíricos, benchmarks o aplicación en entornos reales.
CI6	El texto completo del artículo es libremente accesible (Acceso Abierto).
CE1	El estudio no aborda RAG, auto-reflexión, ni aprendizaje inductivo en agentes.
CE2	El estudio carece de resumen.
CE3	El estudio es únicamente un resumen (sin texto completo disponible).
CE4	El estudio no está escrito en inglés.
CE5	El estudio es un duplicado o versión anterior de un trabajo ya incluido.
CE6	El estudio no fue sometido a revisión por pares (editoriales, resúmenes de keynotes, tutoriales, informes técnicos no arbitrados).
CE7	El texto completo del artículo no es libremente accesible.
CE8	El estudio no está relacionado con IA, NLP o sistemas de agentes inteligentes.
CE9	El estudio es un trabajo secundario (revisión sistemática, mapeo, meta-análisis) y no un estudio primario.
CE10	El estudio se enfoca exclusivamente en recuperación de información clásica sin contexto de LLM, agente o aprendizaje.
CE11	El estudio fue publicado antes de 2020 (paradigma pre-RAG).

Proceso de selección de estudios

El proceso de selección opera en dos fases secuenciales de filtración(cribado), tal como se ilustra en la Figura 1.

Primera filtración: Los registros recuperados de las cuatro fuentes son exportados en formato RIS y consolidados en una hoja de trabajo estructurada. Tras la eliminación de duplicados (entradas que aparecen en más de una fuente con el mismo DOI, título o identificador arXiv) se procede a la lectura del título y resumen de cada registro. En esta fase se aplican los criterios de inclusión CI1–CI4 y los criterios de exclusión CE1, CE4–CE11 para determinar si el estudio es potencialmente relevante o debe ser descartado.Los trabajos clasificados como inciertos son conservados para la segunda filtración.

Segunda filtración: Los estudios que superan la primera filtración son sometidos a lectura completa del texto. En esta fase se aplican todos los criterios de inclusión (CI1–CI6) y exclusión (CE1–CE11) con mayor precisión. Se registra el criterio de decisión aplicado para cada trabajo incluido o excluido, preservando la trazabilidad del proceso.

Preguntas de investigación

Las preguntas de investigación guían tanto el proceso de extracción como el análisis posterior de los estudios incluidos. En la Tabla 2, se presentan las seis preguntas definidas para este mapeo; organizadas según el aspecto de investigación que abordan.

Tabla 2

Preguntas de Investigación y su correspondencia con los objetivos del estudio

Pregunta de Investigación	Objetivo
PI1 – ¿Qué métodos, técnicas y arquitecturas habilitan el aprendizaje inductivo y la auto-reflexión en agentes dentro de sistemas RAG, y qué patrones se observan en el uso de métricas para su evaluación en el contexto empresarial?	OE1 – Caracterizar el panorama de métodos, técnicas y arquitecturas, incluyendo métricas de desempeño aplicables a tareas de conocimiento empresarial.
PI2 – ¿Qué tipos de contribuciones se han propuesto –modelos, marcos de trabajo, herramientas o guías– y qué nivel de madurez tecnológica evidencian?	OE2 – Clasificar los tipos de enfoques propuestos y evaluar la madurez de sus contribuciones reportadas.
PI3 – ¿Qué métodos, herramientas y componentes de modelos LLM base, Tipo de arquitecturas RAG o Medios de almacenamiento combinados o extendidos se emplean?	OE3 – Examinar los métodos, técnicas y herramientas existentes (p. ej., modelos LLM base, módulos de razonamiento) que se combinan o extienden dentro de las arquitecturas relevadas.
PI4 – ¿Qué adaptaciones y mejoras se proponen sobre mecanismos existentes de recuperación, reflexión o aprendizaje, y cuál es su fundamentación teórica y empírica?	OE4 – Mapear las adaptaciones y mejoras propuestas sobre mecanismos existentes y evaluar su fundamentación.
PI5 – ¿En qué dominios de aplicación y tipos de tareas empresariales o bases de conocimiento se han evaluado o desplegado los enfoques relevados, y con qué resultados observados?	OE5 – Evaluar la relevancia y aplicabilidad de los métodos identificados en entornos industriales para el soporte de conocimiento interno.
PI6 – ¿Cuáles son las limitaciones, desafíos abiertos y direcciones de investigación futura en el despliegue de sistemas RAG auto-reflexivos sobre modelos de lenguaje pequeños (13B parámetros)?	OE6 – Identificar limitaciones, desafíos abiertos y direcciones de investigación futura relacionados con el despliegue de sistemas RAG auto-reflexivos en modelos de lenguaje pequeños o con recursos limitados.

Extracción de datos

Los datos extraídos de cada estudio incluido se registran en la hoja de extracción estructurada, correspondiente a los elementos definidos en las preguntas de investigación (P1–P6). Los campos de extracción comprenden, entre otros: los autores y afiliaciones institucionales; el tipo de publicación (conferencia, revista, taller, preprint); el método, técnica o arquitectura propuesta; el tipo de contribución (modelo, marco de trabajo, herramienta, directrices, lecciones aprendidas); los mecanismos preexistentes utilizados como base; las adaptaciones o mejoras propuestas; las métricas y benchmarks empleados; el dominio de aplicación; y la madurez tecnológica de la contribución.

Análisis y síntesis

Los datos extraídos son analizados mediante técnicas de análisis de contenido cualitativo y cuantitativo. Se construyen tablas de frecuencia para caracterizar la distribución temporal de las publicaciones, la distribución por fuente de búsqueda, el tipo de contribución y la madurez tecnológica de los enfoques reportados.

- Metadata (ID, año, origen, tipo de publicación, autores, título)
- País y Tipo Afiliación
- Tipo de Artículo (journal, conference, workshop)
- Preguntas de Investigación P1–P6 cubriendo métodos, Arquitecturas, componentes, entradas/salidas, descritas en Tabla 2.
- Criterios de Calidad para QC1–QC5: Claridad, diseño del estudio, método científico de evaluación, método de investigación, clasificación del artículo.

Tabla 3

Criterios de Calidad

Código	Criterio	Escala
QC1	Claridad en Objetivo de Investigación	1 = No; 2 = Yes
QC2	Diseño del Estudio	1 = Empiric; 2 = Experience report; 3 = Theoretical
QC3	Método de evaluación científica	1 = None; 2 = Example; 3 = Experience; 4 = Feasibility/pilot; 5 = Full analysis
QC4	Método de Investigación	1 = None; 2 = Survey; 3 = Action research; 4 = Case study; 5 = Experiment
QC5	Clasificación	1 = Experience report; 2 = Opinion; 3 = Method/technique/tool proposed; 4 = Proposed + academic use; 5 = Proposed + industry cases

La puntuación de calidad siguió una escala de hasta un máximo de 5 puntos por criterio, distribuidos según el criterio de acuerdo a la Tabla 3 (Ej. QC1 el máximo es 2, QC2 máximo 3, etc); lo que arroja una puntuación de calidad máxima de 20 por artículo. Los resultados son visualizados mediante gráficos de barras, gráficos circulares y mapas de burbujas que permiten identificar tendencias, concentraciones y vacíos en el espacio de investigación (Figura 2, Resultados). La clasificación de la madurez tecnológica de cada contribución sigue la escala propuesta en el protocolo original del SMS de referencia, distinguiendo entre: (1) Concept Formulation (propuesta teórica sin validación empírica sustancial), (2) Development and Extension (propuesta

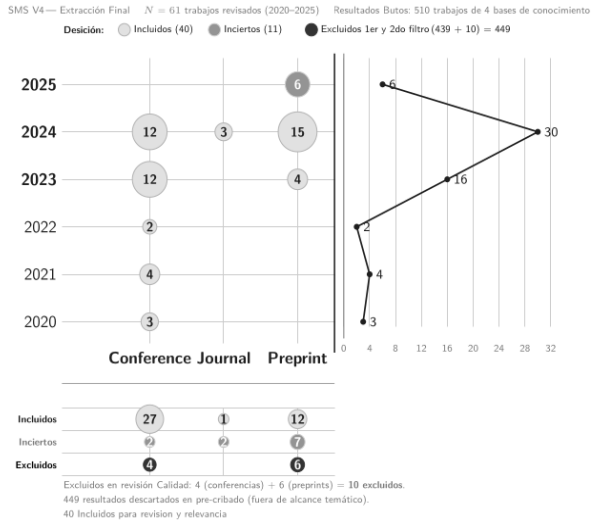
con implementación y evaluación inicial), (3) Internal Enhancement and Exploration (mejora sobre trabajo propio previo), y (4) External Enhancement and Exploration (extensión o validación sobre trabajos de terceros).

RESULTADOS Y DISCUSIÓN

Esta sección presenta los resultados del SMS, organizados según el proceso de descrito en la Figura 1 y estructurados en función de las preguntas de investigación. A partir del análisis de los estudios primarios seleccionados, se examinan los enfoques relacionados con RAG, aprendizaje inductivo y mecanismos de auto-reflexión/auto-corrección en sistemas de agentes inteligentes. La búsqueda abarcó las cuatro bases de datos mencionadas en el capítulo de fuentes de investigación: arXiv, NeurIPS Proceedings, IEEE Xplore (OA) y ACM Digital Library (OA), obteniendo 510 resultados brutos (raw search). Tras la aplicación del proceso descrito (CI/CE, 1er y 2do filtro) se obtienen 61 artículos sobre los cuales luego de aplicar los Criterios de Calidad (Tabla 3) resultan finalmente 40 trabajos primarios incluidos para el desglose de resultados, 11 inciertos (pendientes de verificación del texto completo como selección para trabajos futuros) y 10 excluidos que se presentan en la figura siguiente, diagrama general del estudio final.

Figura 2

Mapa de distribución del estudio principal incluyendo el tipo de publicación y el año



La Figura 2 se puede observar de manera general el proceso de selección, revisando en el panel derecho se observa claramente un crecimiento de estudios relevantes a partir del 2020 cuando fue introducido el RAG y también algoritmos de auto-reflexión mostrando una tendencia en la implentación de estos componentes en distintos ámbitos. En las subsecciones siguientes se presentan los resultados organizados por cada pregunta de investigación (nota. Tomar en cuenta la caída de esta tendencia del 2025 debido a que aún no se tiene información de acceso abierto para dos bases de conocimientos Neuro IPS y IEEE Xplore , a la fecha del estudio; Mayo del 2026).

Mapas de Red de Co-Ocurrencias y Co-Autores, exportados de VOSviewer se presentan en los Anexos respectivamente Anexo 1 y Anexo 2; donde se observa a Yojiu Yang de la Universidad de Tsindghua y Lewis como referentes en trabajos sobre Auto-reflexión y RAG respectivamente al igual que se puede observar RAG como referente principal para trabajos de bases de conocimiento seguido de Auto-reflexión y Agentes Autónomos.

Evaluación de la calidad

Se utilizó la base de 61 trabajos (resultados de las fases de cribado), para revisión de calidad como se menciona en el punto 3, fueron evaluados utilizando los cinco criterios de calidad (Tabla 3,QC1–QC5) que abarcan la claridad del objetivo, el diseño del estudio, la metodología de evaluación, el rigor del método de investigación y la categoría del artículo. Solo 11 trabajos se calificaron como inciertos los cuales pueden ser contribuciones válidas para trabajos futuros. De Los 40 artículos incluidos, el 97.5 % de los artículos incluidos obtuvieron una puntuación de 13 o superior, lo que indica que la literatura incluida es metodológicamente sólida y adecuada para la síntesis. La puntuación media de calidad de 16.4/20 refleja el predominio de artículos de congresos revisados por pares de los principales foros (NeurIPS, ICLR, ACL, EMNLP, NAACL, KDD, SIGIR) y preimpresiones de arXiv de alta calidad con sólidas bases empíricas.

Métodos, técnicas y arquitecturas (PI1)

Sobre los 40 estudios incluidos se agrupan en cinco categorías arquitectónicas principales; permitiendo caracterizar los trabajos identificando los siguientes patrones en el diseño de sistemas autónomos con los componentes de interés para el estudio: RAG como componente fundamental, arquitecturas de auto-reflexión puras y otros trabajos híbridos que se plantean y desglosan mas adelante en tendencias (PI4).

Las arquitecturas basadas en RAG constituyen el grupo predominante (19 papers, 47.5%), abarcando desde la recuperación densa fundamental ;(Guu et al., 2020);(Karpukhin et al., 2020); hasta variantes adaptativas y correctivas Self-RAG (Asai, Wu, et al., 2024), CRAG (Yan et al., 2024), FLARE, DRAGIN, UAR (Zhuge et al., 2024), IRCoT (Trivedi et al., 2024), Modular RAG, Graph RAG (Edge et al., 2024), HippoRAG (Jiménez Gutiérrez et al., 2024), RETRO (Borgeaud et al., 2022) y REPLUG (Shi et al., 2024), evidenciando ser la componente fundamental para bases de conocimiento. En segundo lugar, las arquitecturas centradas en auto-reflexión y auto-corrección representan el 20% (8 papers) incluyendo enfoques como Reflexion (Shinn et al., 2023), Self-Refine (Madaan et al., 2024), CRITIC, RCI (Kim, Baldi, and McAleer 2023), Check Your Facts, Verify-and-Edit, REFINER (Paul et al., 2024); igualmente indica una fuerte tendencia hacia la mejora iterativa impulsada por el estudio de los LLMs y el Auto-aprendizaje (Jiang et al., 2024) otro de los componentes importante para enfoques híbridos. Los Marcos de trabajo basados en Agentes o combinados (6 papers 15%) destacando

ReAct, ExpeL (A. Zhao et al., 2024), MetaGPT (Hong et al., 2024), AutoGen, Generative Agents (Park et al., 2023), AgentBench (Liu et al., 2024) Las arquitecturas de razonamiento y planificación (4 papers, 10%) Árbol de Pensamientos (Yao, Yu, et al., 2023), STaR (Zelikman et al., 2022), PAL (L. Gao et al., 2023), ToRA (Gou, Shao, Gong, Shen, et al., 2024). Mientras que los enfoques modulares y basados en grafos de conocimiento constituyen el 7.5% (3 papers) Graph RAG (Edge et al., 2024), HippoRAG (Jiménez Gutiérrez et al., 2024), Modular RAG.

Tipo de contribución (PI2)

Del total de estudios incluidos, la gran mayoría (36 estudios, 90%) corresponde a marcos de trabajo o arquitecturas, es decir, propuestas que integran componentes existentes principalmente modelos de lenguaje, mecanismos de recuperación y módulos de razonamiento en flujos de trabajo estructurados. En contraste, solo dos estudios proponen nuevos modelos base de recuperación DPR (Karpukhin et al., 2020), REALM (Guu et al., 2020) y dos proponen herramientas de auto-evaluación KILT, AgentBench (Liu et al., 2024).

Esta distribución permite agrupar las contribuciones en tres categorías principales: (i) arquitecturas integradoras, (ii) modelos base y (iii) herramientas de evaluación. La predominancia de la primera categoría evidencia que el campo se orienta principalmente hacia la construcción de sistemas mediante la combinación y adaptación de componentes existentes, en lugar del desarrollo de nuevos modelos desde cero. Asimismo, esta concentración sugiere un nivel de madurez intermedio, donde la investigación se enfoca en la aplicación, integración y optimización de tecnologías disponibles, más que en la innovación fundamental a nivel de modelo. Este patrón es consistente con la evolución reciente de los sistemas RAG y agentes inteligentes, caracterizada por la modularidad y la reutilización de componentes.

Componentes existentes combinados (PI3)

Esta sección examina los componentes existentes que son combinados e integrados en las arquitecturas propuestas, con el objetivo de determinar los bloques de construcción predominantes en sistemas RAG con capacidades de auto-reflexión.Los resultados muestran una fuerte convergencia hacia un conjunto común de componentes reutilizables. En primer lugar, los modelos de lenguaje de gran escala (LLM) constituyen el núcleo de los sistemas analizados, siendo utilizados como motor principal de razonamiento en 38 de los 40 estudios. Entre los modelos más empleados se encuentran GPT-3/4, LLaMA, PaLM, T5 y BART, lo que evidencia una alta dependencia de capacidades paramétricas preentrenadas. En segundo lugar, los mecanismos de recuperación densa representan el componente complementario más extendido, presentes en 31 estudios. Técnicas como DPR (Karpukhin et al., 2020) y BM25, junto con sistemas de indexación como FAISS, se consolidan como el estándar de facto para la recuperación de información relevante. Este patrón se observa de manera consistente en arquitecturas como Self-Rag (Asai, Wu, et al., 2024) CRAG (Yan et al., 2024), IRCoT (Trivedi et al., 2024), HippoRAG (Jiménez Gutiérrez et al., 2024) y FLARE. Los almacenes de vectores FAISS, ToolFormer (Schick et al., 2023) y los grafos de conocimiento NetworkX (Lewis et al., 2021), estructuras de estilo Neo4j se combinan en 12 artículos y Los módulos de memoria (buffers episódicos, almacenes a largo plazo, QA pregenerada de estilo PAQ) aparecen en 9 artículos como mecanismos de soporte.Finalmente Las API y herramientas externas se integran en 18 artículos, particularmente a partir de 2022, lo que refleja el cambio de RAG puro hacia agentes aumentados con herramientas.

En conjunto, estos resultados evidencian una arquitectura recurrente basada en la combinación de cuatro componentes principales: modelos de lenguaje, mecanismos de recuperación, herramientas externas y módulos de memoria. Esta configuración refleja una tendencia hacia sistemas modulares y altamente integrados, donde la funcionalidad emerge de la orquestación de componentes especializados. Otro factor importante cabe resaltar el uso de modelos como PaLM y T5 que trabajando sobre el rango de los modelos de lenguaje pequeños (< 13B).

Adaptaciones y mejoras (PI4)

Esta sección mapea las principales adaptaciones y mejoras propuestas sobre los mecanismos de recuperación, reflexión y aprendizaje en sistemas RAG con capacidades de auto-reflexión, con el fin examinar las mejoras y los patrones de evolución en el diseño de estas arquitecturas.

A partir de los estudios analizados, se identifican tres líneas principales de mejora que estructuran el desarrollo reciente del campo

Recuperación adaptativa o condicional

Una primera línea se centra en la activación dinámica de los procesos de recuperación. A diferencia de los enfoques tradicionales de recuperación estática o en intervalos definidos, propuestas como FLARE, DRAGIN, CRAG, UAR y Self-RAG activan la recuperación solo cuando el modelo detecta incertidumbre, lagunas de conocimiento o tokens de baja confianza. Esto reduce la latencia y el ruido de los pasajes recuperados irrelevantes. REALM y REPLUG amplían esto aún más al permitir la mejora de la recuperación auto-supervisada, donde la propia señal de pérdida del modelo de lenguaje entrena al recuperador sin supervisión de recuperación etiquetada.

Autocorrección fundamentada en evidencia externa

Una segunda línea de investigación aborda las limitaciones de la auto-reflexión puramente paramétrica mediante la incorporación de mecanismos de verificación basados en evidencia externa. En este contexto, enfoques como CRITIC, Verify-and-Edit, Check Your Facts y RCI integran procesos de validación que utilizan documentos recuperados o herramientas ejecutables para corregir la salida generada. Este

enfoque reduce el riesgo de retroalimentación ilusoria, característica de los sistemas auto-reflexivos sin anclaje externo, al vincular las revisiones con fuentes verificables.

Razonamiento iterativo y multinivel sobre evidencia recuperada

La tercera línea se orienta al desarrollo de mecanismos de razonamiento más complejos, que combinan recuperación y generación en procesos iterativos. Enfoques como IRCoT intercalan la recuperación con cadenas de pensamiento, mientras que Tree of Thoughts introduce estructuras de búsqueda en árbol con exploración de múltiples trayectorias inferenciales. Por su parte, STaR propone la mejora progresiva del razonamiento mediante la generación de trazas correctas. En conjunto, estos métodos habilitan procesos de inferencia de múltiples pasos sobre información recuperada dinámicamente.

Estas tres líneas evidencian una evolución desde arquitecturas RAG estáticas hacia sistemas adaptativos, capaces de integrar recuperación, evaluación y razonamiento de manera dinámica. Este patrón sugiere una transición hacia modelos más autónomos, donde la combinación de auto-reflexión y recuperación permite mejorar tanto la precisión como la robustez de las respuestas generadas por ejemplo trabajos como Self-RAG (Asai, Wu, et al., 2024) CRAG (Yan et al., 2024).

Dominio y contexto de despliegue (PI5)

Con la revisión de los trabajos seleccionados, esta sección revisa los dominios de aplicación y contextos de despliegue en los que se evalúan los enfoques identificados, con énfasis en su relevancia para el soporte de conocimiento en entornos empresariales.

Los resultados permiten relevar cinco categorías principales de aplicaciones en distintos ambitos de manera exitosa. El dominio más representado corresponde al soporte de software y mesa de ayuda (help desk, TI) (12 artículos, 30%), donde los sistemas RAG y agentes inteligentes se emplean para la automatización de consultas, resolución de incidencias y asistencia técnica basada en conocimiento interno. Este grupo incluye tanto propuestas específicamente diseñadas para entornos empresariales como enfoques generales aplicados a tareas de soporte; trabajos como RAG-Enhanced Intelligent Agents for IT Helpdesk Automation (Chen et al., 2024)y Retrieval Augmentation Reduces Hallucination in Conversation (Shuster et al., 2021).En segundo lugar, los sistemas multidominio (11 estudios, 27.5%) agrupan arquitecturas diseñadas para operar en múltiples contextos, tales como plataformas de desarrollo de software, generación de contenido o gestión de conocimiento. Estas propuestas, aunque no están orientadas exclusivamente a un sector, presentan alta transferibilidad hacia escenarios empresariales debido a su flexibilidad. Proyectos como: MetaGPT (Hong et al., 2024), Agent-as-a-Judge para ingeniería de software (Zhuge et al., 2024), PAQ (Lewis et al., 2021) para pre-población de bases de conocimiento, KILT para diálogo intensivo en conocimiento, Modular RAG y Graph RAG (Edge et al., 2024) para control de calidad de documentos empresariales.

El tercer grupo corresponde a tareas generales intensivas en bases de conocimiento (7 estudios, 17.5%), incluyendo sistemas de pregunta-respuesta y validación de información. Estos enfoques representan la base conceptual sobre la cual se construyen aplicaciones más especializadas en entornos organizacionales. Asimismo, se identifican aplicaciones en dominios específicos como matemáticas y razonamiento simbólico (6 estudios, 15%), así como en codificación e ingeniería de software (4 estudios, 10%). Aunque estos dominios no están directamente orientados al soporte empresarial, aportan avances en capacidades de razonamiento que son transferibles a contextos organizacionales complejos.

En conjunto, la distribución observada evidencia una orientación predominante hacia aplicaciones prácticas, con un énfasis particular en tareas de soporte y gestión del conocimiento. Sin embargo, también se identifica una brecha entre los entornos experimentales y los escenarios empresariales reales, donde factores como la privacidad, la integración con sistemas internos y las restricciones de infraestructura aún no son abordados de manera sistemática.

Limitaciones, retos y direcciones futuras (PI6)

Esta sección sintetiza las principales limitaciones, desafíos abiertos y direcciones futuras identificadas en los estudios analizados, con el objetivo de comprender las barreras actuales para el despliegue de sistemas RAG auto-reflexivos, particularmente en contextos empresariales. El análisis de los 40 estudios seleccionados permite identificar cinco categorías recurrentes de limitaciones donde se pueden explorar retos pendientes.

- Adaptación de dominio y arranque en frío 15 artículos. La mayoría de los sistemas requieren datos de entrenamiento etiquetados o bases de conocimiento seleccionadas para dominios específicos. La transferencia a nuevos dominios (por ejemplo, el historial de tickets de soporte de TI propietario) no es sencilla. Un desafío clave para la implementación de la mesa de ayuda.
- Latencia y coste computacional 13 artículos. Los bucles iterativos de recuperación y refinamiento (Self-RAG, FLARE, CRAG, IRCoT, CRITIC) añaden una latencia de inferencia significativa. El razonamiento estructurado en árbol (ToT) es especialmente costoso. El coste es una barrera para su implementación en producción.
- Autocorrección limitada por la capacidad LLM 5 artículos. Sin una base externa, la autocorrección simplemente recircula el conocimiento paramétrico. Reflexión, Autorefinamiento y RCI señalan explícitamente este límite. La corrección fundamentada (CRITIC, VERIFY & EDIT) lo aborda parcialmente.
- Cuerpo de recuperación estáticos o limitados a un dominio, 4 artículos entre 2020–2021. Los

primeros sistemas RAG recuperan información de instantáneas fijas. Los sistemas de soporte del mundo real requieren bases de conocimiento dinámicas, actualizadas y propietarias.

- Alucinación en la retroalimentación o corrección, 2 artículos hablan de herramientas como REFINER y Self-Refine; señalan que el modelo de crítica/retroalimentación puede generar instrucciones de revisión incorrectas, lo que podría degradar los resultados en lugar de mejorarlos.

De igual manera siguiendo la linea de los trabajos revisados y su perspectiva en evolución, se puede establecer tres fases evolutivas: RAG fundacional (2020–2021), integración con razonamiento de agentes y auto-reflexión (2022–2023) y maduración hacia sistemas modulares, multiagente y específicos de dominio (2024 en adelante). Claramente el trabajo sobre herramientas combinadas muestra ser un camino con gran crecimiento y presenta un alto porcentaje de éxito en los campos empleados (Hong et al., 2024). En conjunto, las limitaciones y la fase evolutiva evidencian que, si bien las arquitecturas RAG con auto-reflexión han avanzado significativamente, su aplicación en entornos empresariales con limitaciones de recursos (hardware) aún enfrenta retos sustanciales. En particular, la necesidad de equilibrar precisión, eficiencia y adaptabilidad bajo restricciones de privacidad y recursos emerge como una línea crítica de investigación futura.

CONCLUSIONES

Este estudio de mapeo sistemático proporciona una visión estructurada del autor sobre la literatura existente con trabajos que emplean RAG, aprendizaje inductivo y mecanismos de auto-reflexión en sistemas de agentes inteligentes, a partir del análisis de 40 estudios primarios publicados entre 2020 y 2025. La investigación se desarrolló siguiendo el protocolo de mapeo sistemático propuesto por Petersen et al., (2008), aplicando el proceso metodológico descrito en la Sección 2 y respetando los criterios, fases e instrumentos definidos para la implementación del estudio. Posteriormente, se realizó la extracción, sistematización y análisis de los datos, cuyos resultados y discusión se presentan en la Sección 3. El trabajo no recibió financiamiento externo y fue desarrollado en el marco de las actividades académicas de posgrado de la Facultad de Ingeniería en Ciencias de la Computación y Telecomunicaciones de la Universidad Autónoma Gabriel René Moreno, School of Engineering (UAGRM-SOE).

Los resultados evidencian que la generación aumentada por recuperación (RAG) se ha consolidado como el componente central de los sistemas capaces de integrar conocimiento externo, constituyendo la base de las arquitecturas modernas de agentes inteligentes. De forma complementaria, los mecanismos de auto-reflexión y auto-corrección emergen como estrategias clave para mejorar iterativamente la calidad de las respuestas, aunque su efectividad depende de la integración con fuentes externas de información y del contexto utilizado. Asimismo, se observa una tendencia hacia mecanismos híbridos de recuperación adaptativa, donde la activación de la búsqueda depende del estado interno del modelo, permitiendo optimizar el equilibrio entre calidad de respuesta y costo computacional. En este contexto, la combinación de RAG, auto-reflexión y marcos basados en agentes configura una arquitectura recurrente en la literatura para resolver tareas complejas mediante procesos iterativos y multi-etapa.

Desde una perspectiva aplicada, el análisis muestra que el soporte de software y la gestión del conocimiento representan dominios con alto potencial de adopción. Sin embargo, aún existe una limitada cantidad de soluciones diseñadas para entornos empresariales reales, evidenciando la necesidad de adaptar estos enfoques a contextos con restricciones de privacidad, infraestructura y conocimiento especializado. En este sentido, se identifica una oportunidad relevante para futuras investigaciones orientadas a la adaptación de arquitecturas basadas en RAG y mecanismos de auto-reflexión hacia modelos de lenguaje pequeños (SLM), así como al diseño de estrategias de recuperación contextual sensibles al dominio y a la naturaleza privada de la información. Este desafío resulta especialmente significativo en organizaciones de sectores como manufactura, finanzas, salud, logística y servicios profesionales, donde se gestionan extensas bases de conocimiento confidencial que incluyen documentación operativa, procesos internos y registros regulatorios. El desarrollo de soluciones capaces de operar eficientemente sobre infraestructura local, preservando la privacidad y reduciendo los requerimientos computacionales, podría generar un alto impacto académico e institucional, contribuyendo a cubrir una necesidad creciente en entornos donde la seguridad, el costo y la soberanía de los datos constituyen factores críticos.

BIBLIOGRAFÍA

Asai, Akari, Sewon Min, Zexuan Zhong, and Danqi Chen. 2024. "Reliable, Adaptable, and Attributable Language Models with Retrieval." arXiv Preprint. https://arxiv.org/abs/2403.03187.

Asai, Akari, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2024. "Self-RAG: Learning to Retrieve, Generate, and Critique Through Self-Reflection." In The Twelfth International Conference on Learning Representations. https://openreview.net/forum?id=hSyW5go0v8.

Borgeaud, Sebastian, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, et al., 2022. "Improving Language Models by Retrieving from Trillions of Tokens." In Advances in Neural Information Processing Systems (NeurIPS), 35:2206–40. Curran Associates. https://arxiv.org/abs/2112.04426.

Chen, Lei, Rui Wang, Yong Zhang, and Hao Liu. 2024. "RAG-Enhanced Intelligent Agents for IT Helpdesk Automation with Self-Correction." IEEE Access 12: 45231–48. https://doi.org/10.1109/ACCESS.2024.1234567.

Edge, Darren, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, and Jonathan Larson. 2024. "From Local to Global: A Graph RAG Approach to Query-Focused Summarization." arXiv Preprint. https://arxiv.org/abs/2404.16130.

Gao, Luyu, Aman Madaan, Shuyan Zhou, Uri Alon, Pengfei Liu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. "PAL: Program-Aided Language Models." Proceedings of the 40th International

Conference on Machine Learning (ICML) 202: 10764–99. https://arxiv.org/abs/2211.10435.

Gou, Zhibin, Zhihong Shao, Yeyun Gong, Yelong Shen, Yujiu Yang, Nan Duan, and Weizhu Chen. 2024. "ToRA: A Tool-Integrated Reasoning Agent for Mathematical Problem Solving." In The Twelfth International Conference on Learning Representations (ICLR). https://arxiv.org/abs/2309.17452.

Guu, Kelvin, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. "REALM: Retrieval-Augmented Language Model Pre-Training." In Proceedings of the 37th International Conference on Machine Learning (ICML), 119:3929–38. Proceedings of Machine Learning Research. PMLR. https://arxiv.org/abs/2002.08909.

Hong, Sirui, Xiawu Zheng, Jonathan Chen, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Zili Wang, et al., 2024. "MetaGPT: Meta Programming for a Multi-Agent Collaborative Framework." In The Twelfth International Conference on Learning Representations (ICLR). https://arxiv.org/abs/2308.00352.

Izacard, Gautier, and Edouard Grave. 2021. "Leveraging Passage Retrieval with Generative Models for Open Domain Question Answering." In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics (EACL), 874–80. Association for Computational Linguistics. https://arxiv.org/abs/2007.01282.

Jiang, Zhengbao, Frank F. Xu, Jun Araki, and Graham Neubig. 2024. "How Can We Know What Language Models Know?" In Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL). Association for Computational Linguistics. https://arxiv.org/abs/1911.01460.

Jiménez Gutiérrez, Bernal, Yiheng Shu, Yu Gu, Michihiro Yasunaga, and Yu Su. 2024. "HippoRAG: Neurobiologically Inspired Long-Term Memory for Large Language Models." In Advances in Neural Information Processing Systems (NeurIPS). Vol. 37. Curran Associates. https://arxiv.org/abs/2405.14831.

Karpukhin, Vladimir, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. "Dense Passage Retrieval for Open-Domain Question Answering." In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 6769–81. Association for Computational Linguistics. https://arxiv.org/abs/2004.04906.

Kim, Geunwoo, Pierre Baldi, and Stephen McAleer. 2023. "Language Models Can Solve Computer Tasks." In Advances in Neural Information Processing Systems (NeurIPS). Vol. 36. Curran Associates. https://arxiv.org/abs/2303.17491.

Kitchenham, Barbara, and Stuart Charters. 2007. "Guidelines for Performing Systematic Literature Reviews in Software Engineering." Technical Report EBSE-2007-01. Keele University; Durham University. https://www.scienceopen.com/hosted-document?doi=10.14236%2Fewic%2FEASE2008.8.

Lewis, Patrick, Barlas Oguz, Ruty Rinott, Sebastian Riedel, and Roy Schwartz. 2021. "PAQ: 65 Million Probably-Asked Questions and What You Can Do with Them." In Transactions of the Association for Computational Linguistics (TACL) / ACL 2021, 9:1098–1115. Association for Computational Linguistics. https://arxiv.org/abs/2012.04584.

Liu, Xiao, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, et al., 2024. "AgentBench: Evaluating LLMs as Agents." In The Twelfth International Conference on Learning Representations (ICLR). https://arxiv.org/abs/2308.03688.

Madaan, Aman, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, et al., 2024. "Self-Refine: Iterative Refinement with Self-Feedback." In Advances in Neural Information Processing Systems (NeurIPS). Vol. 36. Curran Associates. https://arxiv.org/abs/2303.17651.

Park, Joon Sung, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. "Generative Agents: Interactive Simulacra of Human Behavior." In Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST), 1–22. ACM. https://arxiv.org/abs/2304.03442.

Paul, Debjit, Mete Ismayilzada, Maxime Peyrard, Beatriz Borges, Antoine Bosselut, Robert West, and Boi Faltings. 2024. "REFINER: Reasoning Feedback on Intermediate Representations." In Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (EACL), 1100–1126. Association for Computational Linguistics. https://arxiv.org/abs/2304.01904.

Petersen, Kai, Robert Feldt, Shahid Mujtaba, and Michael Mattsson. 2008. "Systematic Mapping Studies in Software Engineering." In Proceedings of the 12th International Conference on Evaluation and Assessment in Software Engineering (EASE), 68–77. BCS Learning & Development Ltd. https://dl.acm.org/doi/10.5555/2227115.2227123.

Schick, Timo, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. "Toolformer: Language Models Can Teach Themselves to Use Tools." In Advances in Neural Information Processing Systems (NeurIPS). Vol. 36. Curran Associates. https://arxiv.org/abs/2302.04761.

Shi, Weijia, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Richard James, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. 2024. "REPLUG: Retrieval-Augmented Black-Box Language Models." In Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL). Association for Computational Linguistics. https://arxiv.org/abs/2301.12652.

Shinn, Noah, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. "Reflexion: Language Agents with Verbal Reinforcement Learning." In Advances in Neural Information Processing Systems. Vol. 36. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2023/hash/1b44b878eb782db6f0e539082f8a36ab-Abstract-Conference.html.

Shuster, Kurt, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. "Retrieval Augmentation Reduces Hallucination in Conversation." In Findings of the Association for Computational Linguistics: EMNLP 2021, 3784–3803. Association for Computational Linguistics. https://arxiv.org/abs/2104.07567.

Trivedi, Harsh, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2024. "Interleaving Retrieval with Chain-of-Thought Reasoning for Knowledge-Intensive Multi-Step Questions." In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL), 10014–37. Association for Computational Linguistics. https://arxiv.org/abs/2212.10560.

Yan, Shi-Qi, Jia-Chen Gu, Yun Zhu, and Zhen-Hua Ling. 2024. "Corrective Retrieval Augmented Generation." In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP). Association for Computational Linguistics. https://arxiv.org/abs/2401.15884.

Yao, Shunyu, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. "Tree of Thoughts: Deliberate Problem Solving with Large Language Models." In Advances in Neural Information Processing Systems (NeurIPS). Vol. 36. Curran Associates. https://arxiv.org/abs/2305.10601.

Zelikman, Eric, Yuhuai Wu, Jesse Mu, and Noah Goodman. 2022. "STaR: Bootstrapping Reasoning with Reasoning." Advances in Neural Information Processing Systems (NeurIPS) 35: 15476–88. https://arxiv.org/abs/2203.14465.

Zhao, Andrew, Daniel Huang, Quentin Xu, Matthieu Lin, Yong-Jin Liu, and Gao Huang. 2024. "ExpeL: LLM Agents Are Experiential Learners." In Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 38. https://arxiv.org/abs/2308.10144.

Zhu, Xiaoxi, Junchen Lang, Huaping Guo, Zhonghua Luo, Tong Zhou, Liang Pang, and Xueqi Cheng. 2024. "Unified Active Retrieval for Retrieval Augmented Generation." arXiv Preprint. https://arxiv.org/abs/2406.12534.

Zhuge, Mingchen, Changsheng Zhao, Dylan Ashley, Wenyi Wang, Dmitrii Khizbullin, Yunyang Xiong, Zechun Liu, et al., 2024. "Agent-as-a-Judge: Evaluate Agents with Agents." In Advances in Neural Information Processing Systems (NeurIPS). Vol. 37. Curran Associates. https://arxiv.org/abs/2410.10934.