Pontificia Universidad Católica del Ecuador

Facultad de Hábitat, Infraestructura y Creatividad

Maestría en Sistemas de Información mención Data Science

Trabajo previo para la obtención del título de Magister en Sistemas de Información mención Data Science

Tema:

Diseño e Implementación de un Sistema Agéntico para la Gestión de Conocimiento con Inteligencia Artificial Generativa.

Autor:

Bryan Alexander Saravia Avila

Tutor:

Eduardo José Montero Bermudez Mg.

Quito - marzo 2026

Dedicatoria

Dedicado a todos quienes busquen aplicar y ampliar sus conocimientos de Inteligencia Artificial Agéntica, que este trabajo sirva como guía en este apasionante mundo de máquinas que aprenden a “razonar”.

Agradecimientos

A todos quienes me han brindado soporte en este camino de aprendizaje, familia, profesores, compañeros y amigos son fundamentales para que este trabajo llegue a su conclusión.

Tabla de contenido

Lista de tablas

Lista de figuras

Introducción

La gestión del conocimiento (Knowledge Management, KM por sus siglas en inglés) constituye un componente fundamental para la explotación efectiva del conocimiento organizacional. Existen estudios recientes que demuestran que las prácticas de KM explican más del 80 % de la variabilidad del desempeño empresarial mediante mecanismos de innovación, aprendizaje organizacional y eficiencia operativa (Ştefan et al., 2024) , es así, de manera complementaria, (Isa & Rahmah, 2023) muestran que la relación entre KM y desempeño es mediada por capacidades dinámicas con un efecto indirecto significativo, reafirmando que la ausencia de procesos formales y tecnología adecuada limita la creación, uso y transferencia del conocimiento crítico.

Las organizaciones enfrentan diversos retos estructurales, entre ellos la fragmentación documental, la existencia de repositorios heterogéneos, la duplicidad de información, la pérdida de conocimiento experto y las dificultades para acceder a datos no estructurados. Estas brechas impactan la capacidad para responder a necesidades de negocio basadas en conocimiento, situación identificada como un obstáculo en la literatura reciente de KM e innovación organizacional. En este sentido, marcos internacionales como ISO 30401, TOGAF® Standard y DAMA-DMBOK2 establecen que la gestión del conocimiento debe abordarse como un sistema formal integrado a la arquitectura empresarial, con procesos medibles y mecanismos de mejora continua.

En paralelo, el auge de la inteligencia artificial generativa (GenAI) está redefiniendo la forma en que las organizaciones gestionan su conocimiento. Según McKinsey & Company (Chui, Yee, et al., 2023), menos del 33 % de las empresas ha adoptado GenAI de manera regular, pese a su potencial para incrementar la productividad laboral entre 0.1 % y 0.6 % anual, pudiendo alcanzar hasta 3.4 % en combinación con otras tecnologías. La baja adopción se debe, en parte, a la falta de integración de GenAI con marcos robustos de KM y a la dificultad para operacionalizar el conocimiento no estructurado.

En este marco adquieren especial relevancia los sistemas agénticos, junto con las técnicas de Retrieval-Augmented Generation (RAG) y GraphRAG, ya que posibilitan la integración de la capacidad generativa de los modelos de inteligencia artificial con mecanismos de recuperación de información verificada, trazabilidad y razonamiento estructurado. Estudios recientes demuestran que el enfoque RAG contribuye a la reducción de alucinaciones y, entre sus principales beneficios, destaca la mejora en la recuperación contextual de la información, así como el apoyo a la generación y actualización de activos de conocimiento organizacional (Gao et al., 2024; James et al., 2025). De forma complementaria, los enfoques basados en grafos, como GraphRAG, facilitan la exploración de relaciones semánticas complejas y aportan mayor explicabilidad, aspectos esenciales en entornos corporativos donde el conocimiento debe ser confiable, auditable y coherente con los procesos de negocio.

Considerando este escenario, se identifica una brecha crítica, pues, las organizaciones aún carecen de sistemas unificados que integren IA generativa, recuperación contextual (RAG) y principios de gestión del conocimiento alineados a estándares internacionales.

Esta brecha motiva el desarrollo del presente estudio, orientado al diseño e implementación de una arquitectura tecnológica de un sistema agéntico apoyado en inteligencia artificial generativa y técnicas de recuperación de información, con el fin de mejorar los procesos de gestión del conocimiento organizacional.

Objetivos

Objetivo General:

Diseñar, implementar y evaluar un sistema agéntico basado en Retrieval-Augmented Generation (RAG) para la gestión documental del conocimiento organizacional, que permita recuperar y generar respuestas fundamentadas en evidencia documental, y validar su desempeño mediante un experimento con un conjunto de preguntas y métricas de extracción, recuperación y generación.

Objetivos Específicos:

- Desarrollar un pipeline de procesamiento y extracción de texto capaz de manejar diferentes estructuras documentales, incluyendo documentos: i) PDF textuales, ii) documentos escaneados, iii) archivos PDF con tablas estructuradas y iv) presentaciones en formato PPTX.
- Implementar un sistema de almacenamiento basado en una base de datos vectorial para arquitectura RAG, que permita la indexación y recuperación eficiente de fragmentos documentales, garantizando un desempeño mínimo del 80 % en precisión de recuperación de información.
- Diseñar, implementar y evaluar un patrón de inteligencia artificial agéntica basada en Large Language Models (LLM) que permita realizar consultas semánticas y léxicas con enrutamiento de conocimiento, sobre la información almacenada y generar respuestas automáticas con un nivel mínimo de precisión del 70 %.

Justificación

El crecimiento de la inteligencia artificial generativa y de los sistemas agénticos apunta a consolidarlos como componentes cada vez más presentes dentro de las arquitecturas tecnológicas de las organizaciones. Su adopción representa una oportunidad significativa para impulsar procesos de transformación digital, particularmente en la gestión de datos no estructurados, tales como documentos, contratos, manuales, procesos y flujos de trabajo, entre otros.

El conocimiento almacenado en la documentación organizacional se presenta en múltiples formatos, que incluyen soportes digitales, documentos escaneados, manuscritos e incluso archivos multimedia. Esta heterogeneidad plantea desafíos relevantes en términos de almacenamiento, consulta y acceso, dado que puede constituir una barrera para la atención eficiente de requerimientos que demandan la integración de diversas fuentes documentales.

En este contexto, el presente proyecto propone diseñar, implementar y evaluar un sistema agéntico basado en inteligencia artificial generativa y técnicas de recuperación de información que permita unificar diferentes fuentes documentales considerando tanto su estructura como su contenido. Esta propuesta busca contribuir al fortalecimiento de los procesos de gestión del conocimiento organizacional, facilitando el procesamiento, almacenamiento y consulta de información proveniente de múltiples formatos y dominios documentales.

Asimismo, el sistema permitirá superar las limitaciones asociadas a los procesos manuales de estandarización documental, posibilitando que las organizaciones accedan y exploten el conocimiento contenido en sus documentos de manera ágil mediante consultas en lenguaje natural.

Preguntas de Investigación

Pregunta Principal:

¿De qué manera el diseño e implementación de un sistema agéntico basado en Retrieval-Augmented Generation (RAG) contribuye a mejorar la recuperación y consulta de información en documentos no estructurados, como soporte a la gestión del conocimiento organizacional?

Preguntas Secundarias:

1. ¿Qué técnicas de procesamiento y extracción de texto permiten la ingestión eficiente de documentos no estructurados en múltiples formatos dentro de un sistema RAG agéntico?
2. ¿Cómo influye el uso de bases de datos vectoriales en la indexación y recuperación eficiente de fragmentos documentales en un sistema RAG agéntico?
3. ¿En qué medida un patrón agéntico basado en LLM mejora la recuperación contextual y la generación de respuestas fundamentadas en evidencia documental?

Desarrollo

Marco teórico

El presente capítulo establece los fundamentos conceptuales que sustentan el diseño e implementación de una arquitectura agéntica basada en Retrieval-Augmented Generation (RAG) para la recuperación y consulta de información en documentos no estructurados, como soporte a la gestión del conocimiento organizacional.

La segunda sección analiza cómo el conocimiento se constituye como un activo estratégico, determinando su impacto en la innovación, la formulación de estrategias y la toma de decisiones.

La tercera sección describe la gestión documental y el conocimiento no estructurado, abordando la problemática de fragmentación, la heterogeneidad y limitaciones de los sistemas tradicionales, las tipologías documentales y la importancia del análisis del diseño de documentos (Document Layout Analysis) para la extracción de conocimiento. La sección cuatro profundiza en la Inteligencia Artificial Generativa en contextos organizacionales, en este punto se explora la evolución de la IA, capacidades de los modelos de lenguaje de gran escala (LLM) y barreras de adopción en las empresas.

La sección cinco se enfoca en RAG y GraphRAG, mostrando su pipeline, beneficios frente a modelos generativos puros, razonamiento basado en grafos y habilitación de activos de conocimiento organizacional. La sección seis desarrolla los sistemas agénticos basados en IA, incluyendo agentes autónomos, arquitecturas agénticas con LLM y RAG agéntico con patrón de routing, y su aplicación en la gestión del conocimiento.

La sección siete aborda la evaluación de sistemas RAG y agénticos, ya que, pretende resaltar la importancia de métricas objetivas, frameworks automáticos de evaluación y el uso de LLM como juez para asegurar la confiabilidad y trazabilidad de la información generada.

1. Gestión del Conocimiento Organizacional

1.1 Importancia y evolución de la Gestión del Conocimiento

La gestión del conocimiento como activo estratégico en la era de la inteligencia artificial

La Gestión del Conocimiento (*Knowledge Management*, KM) puede entenderse, en términos generales, como el conjunto de procesos sistemáticos a través de los cuales las organizaciones crean, capturan, organizan, almacenan, comparten y aplican conocimiento, con la finalidad de fortalecer su desempeño, su capacidad de toma de decisiones y su potencial de innovación (Al Ahbabi & others, 2024; Alavi & Leidner, 2001) . Es decir, no se trata únicamente de administrar información, sino de gestionarla de manera estratégica para que genere valor dentro de la organización. Y en esta línea la KM no debe concebirse solo como un repositorio estático de

contenidos, sino más bien como un habilitador de capacidades dinámicas que permiten a las organizaciones adaptarse a contextos complejos y en constante cambio. De este modo, contribuye a impulsar procesos de innovación y a sostener la competitividad organizacional en el tiempo (Al Ahbabi & others, 2024; Scuotto & others, 2024).

La literatura sobre KM reconoce que las organizaciones enfrentan grandes e importantes desafíos en la gestión efectiva de su conocimiento. Estudios recientes han evidenciado que los enfoques tradicionales, basados en taxonomías estáticas, motores de búsqueda por palabras clave y repositorios documentales aislados, son insuficientes para atender las necesidades actuales de las organizaciones, especialmente en contextos donde el conocimiento es voluminoso, heterogéneo y distribuido en múltiples formatos, eso lo fundamenta Ahbadi (Al Ahbabi & others, 2024; Alavi & Leidner, 2001). Estos enfoques totalmente pueden soportar operaciones básicas de almacenamiento y consulta, pero no logran capturar la riqueza contextual ni facilitar un acceso semántico profundo al conocimiento organizacional, limitando la capacidad para tomar decisiones complejas basadas en información corporativa dispersa.

La investigación actual propone la integración de tecnologías avanzadas en los sistemas de KM, de modo que estos tengan posibilidad de proporcionar los factores indispensables: recuperación contextualizada del conocimiento, asociación semántica y generación de respuestas coherentes a consultas humanas complejas (Al Ahbabi & others, 2024; Scuotto & others, 2024). Además, el diseño e implementación de un sistema agéntico basado en Retrieval-Augmented Generation (RAG) se presenta como una respuesta tecnológica frente a las limitaciones identificadas por la literatura en la gestión del conocimiento organizacional, se conoce que estos sistemas combinan técnicas de recuperación de información (basadas en vectores) con modelos generativos que entienden y responden en lenguaje natural, ofreciendo una solución más robusta para transformar documentos y otros activos informacionales en conocimiento accesible, relevante y accionable.

- **El conocimiento como activo estratégico**

En las organizaciones contemporáneas, el conocimiento se entiende, cada vez con mayor claridad, como un activo estratégico fundamental para la generación de valor, así como para la formulación de estrategias y la sostenibilidad del desempeño organizacional. Desde la teoría, se plantea que las organizaciones tienen como meta, integrar, coordinar y aplicar conocimiento especializado que se encuentra distribuido entre individuos, procesos y sistema. Esta capacidad es justamente, lo que las diferencia de manera sustantiva de sus competidores (Grant, 1996).

Diversos estudios empíricos confirman que una gestión efectiva del conocimiento se asocia de forma positiva con el aprendizaje organizacional y el desempeño sostenido, tanto en sectores privados como públicos (Kassa & Ning, 2023a; Nonaka et al., 2000). De igual forma, se tiene que el conocimiento actúa como un habilitador clave de capacidades dinámicas que permite a las organizaciones identificar oportunidades y respondan a entornos complejos y cambiantes (Teece,

2007). Bajo esta lógica, la gestión del conocimiento trasciende el almacenamiento de información y se orienta hacia la creación, integración y reutilización continua del conocimiento organizacional.

Aun así, se observa que la literatura reciente señala que muchas organizaciones presentan limitaciones estructurales y tecnológicas que dificultan la explotación efectiva del conocimiento como activo estratégico. Además limitaciones en la fragmentación de repositorios documentales, baja interoperabilidad entre sistemas, dependencia de motores de búsqueda de léxicos y escasa capacidad para gestionar información no estructurada en múltiples formatos (Mahadevkar et al., 2024a; Sternad Zabukovšek et al., 2023a). Entonces, como resultado, el conocimiento disponible pierde impacto en la formulación de estrategias y en la toma de decisiones, reduciendo su contribución al desempeño organizacional (Gelashvili-Luik et al., 2025).

Es entonces que la incorporación de tecnologías basadas en inteligencia artificial se posiciona como un factor importante para superar estas limitaciones, particularmente mediante el uso de modelos de lenguaje de gran escala capaces de procesar, razonar y generar conocimiento a partir de fuentes documentales heterogéneas (Minaee et al., 2024a).

Sin embargo, se debe observar que la adopción de modelos generativos sin mecanismos de control adecuados introduce riesgos relevantes como la generación de información incorrecta, además, la pérdida de trazabilidad y la vulnerabilidad a ataques de seguridad y privacidad, aspectos ampliamente documentados en la literatura reciente (Bai et al., 2024a; X. Liu et al., 2024; Yao et al., 2024a).

En este marco, el diseño de sistemas inteligentes capaces de facilitar el acceso contextualizado y oportuno al conocimiento se alinea en directo con los objetivos de esta investigación. En particular, las arquitecturas agénticas basadas en Retrieval-Augmented Generation permiten integrar modelos generativos con procesos de recuperación estructurada y trazable, reduciendo riesgos de alucinación y fortaleciendo la confiabilidad del conocimiento utilizado en los procesos organizacionales (Peng et al., 2024; Yu et al., 2024a). Dichos sistemas contribuyen a potenciar el valor estratégico del conocimiento organizacional, y se consolida como un activo central para la toma de decisiones y la gestión empresarial basada en evidencia.

- **Evolución de enfoques documentales a sistemas inteligentes**

Con el incremento del volumen de información y la diversificación de formatos documentales, estos enfoques comenzaron a mostrar limitaciones significativas. La proliferación de documentos no estructurados, como reportes extensos, presentaciones, contratos escaneados y contenido multimodal, dificultó la localización eficiente del conocimiento relevante. La gestión documental tradicional pasó de ser un habilitador del conocimiento a convertirse, en muchos casos, en una barrera para su aprovechamiento estratégico (Mahadevkar et al., 2024a).

Las organizaciones iniciaron una transición progresiva hacia sistemas más avanzados, incorporando técnicas de recuperación semántica, ontologías y tecnologías basadas en inteligencia artificial. La gestión documental evolucionó hacia enfoques orientados al conocimiento donde el énfasis ya no se encuentra solo en el documento como unidad de información, sino en el significado y el contexto del contenido almacenado (Ristoski & Paulheim, 2016).

Los sistemas inteligentes basados en Retrieval-Augmented Generation (RAG) pueden entenderse como una evolución natural de los enfoques documentales tradicionales, en la medida en que estos integran procesos de recuperación semántica desde repositorios documentales con modelos generativos, lo que permite que las respuestas producidas se fundamenten en conocimiento previamente almacenado y contextualizado (Gao & others, 2024; Yu et al., 2024a) y esto implica superar la lógica de búsqueda estática para avanzar hacia un modelo más dinámico y contextual de acceso al conocimiento.

La incorporación de arquitecturas de agentes en estos sistemas inteligentes refuerza dicha evolución, ya que permite la orquestación de tareas como selección de fuentes, el razonamiento contextual y control de calidad de las respuestas. La gestión documental deja de ser un proceso pasivo y se transforma en un sistema activo de apoyo a la toma de decisiones, alineado con los objetivos estratégicos de la organización y con los principios modernos de la gestión del conocimiento (Cook et al., 2025a).

1.2 Tipologías del conocimiento organizacional

La diferencia entre conocimiento explícito y conocimiento tácito constituye uno de los fundamentos conceptuales de la gestión del conocimiento organizacional, permite comprender cómo se origina el conocimiento, se comparte y se aplica dentro de las organizaciones, así como las limitaciones de los enfoques tradicionales para su gestión.

El conocimiento explícito se refiere a aquel que puede ser formalizado, codificado y almacenado en soportes tangibles, tales como documentos, manuales, bases de datos o procedimientos estandarizados. Este tipo de conocimiento es fácilmente transferible, debido a su naturaleza estructurable, y ha sido históricamente el principal foco de los sistemas de información organizacionales (Nonaka, 1994). De donde muchas estrategias de gestión del conocimiento se han centrado en su captura y organización.

Por otro lado, el conocimiento tácito es inherentemente personal, contextual y difícil de formalizar. Se encuentra ligado a la experiencia, intuición y habilidades de los individuos, por lo que su transferencia no puede realizarse de manera directa mediante documentos o sistemas formales (Polanyi, 1966). Así, la literatura reconoce que una gestión del conocimiento efectiva debe considerar mecanismos que permitan articular ambos tipos de conocimiento y facilitar su interacción.

- **Conocimiento estructurado y no estructurado**

Otra clasificación relevante para el análisis de la gestión del conocimiento es la que distingue entre conocimiento estructurado y no estructurado. Esta diferencia trascendental resulta especialmente pertinente en contextos de la transformación digital y el crecimiento exponencial de información disponible en las organizaciones.

El conocimiento estructurado corresponde a información que está organizada bajo esquemas predefinidos, como bases de datos relacionales, hojas de cálculo o sistemas transaccionales, este tipo de conocimiento facilita su almacenamiento, consulta y análisis mediante herramientas tradicionales de gestión de la información (Sternad Zabukovšek et al., 2023a).

Por otro lado, el conocimiento no estructurado se presenta en formatos que no siguen una estructura rígida, como documentos de texto, correos electrónicos, informes, presentaciones o contenidos multimedia. Estudios señalan que una parte significativa del conocimiento organizacional se encuentra en este tipo de fuentes, lo que dificulta su explotación mediante sistemas convencionales (Mahadevkar et al., 2024a).

Es decir, la literatura reciente destaca el rol de la inteligencia artificial y, en particular, de los modelos de lenguaje de gran escala, como habilitadores claves para transformar conocimiento no estructurado en información accionable. Estas tecnologías permiten interpretar el contenido semántico de los documentos y facilitar su integración en procesos de toma de decisiones estratégicas (Yu et al., 2024a).

- **Documentos como repositorios primarios de conocimiento explícito**

Históricamente, los documentos han constituido el principal medio para almacenar y preservar el conocimiento explícito dentro de las organizaciones. Manuales operativos, políticas internas, reportes técnicos y documentos estratégicos han sido utilizados como mecanismos formales para capturar y difundir conocimiento organizacional. Por eso, los sistemas de gestión documental surgieron con el objetivo de organizar estos repositorios y facilitar el acceso a la información. Sin embargo, la literatura evidencia que dichos sistemas suelen limitarse a funciones de almacenamiento y búsqueda por palabras clave, sin considerar el contexto (Sternad Zabukovšek et al., 2023a). En consecuencia, el valor estratégico del conocimiento contenido en los documentos no siempre se traduce en apoyo efectivo a la toma de decisiones.

El crecimiento sostenido del volumen documental ha incrementado la complejidad para localizar información relevante de manera oportuna. Este escenario ha puesto en evidencia la necesidad de evolucionar desde repositorios documentales pasivos hacia sistemas inteligentes capaces de interpretar, contextualizar y relacionar el conocimiento explícito almacenado (Kassa & Ning, 2023a).

Con esta idea en mente, las arquitecturas basadas en Retrieval-Augmented Generation (RAG) permiten redefinir el rol de los documentos, y no solo como fuentes de información, sino también como activos dinámicos que alimentan sistemas agénticos orientados a la generación de conocimiento contextualizado, alineándose con los objetivos de esta investigación.

1.3 Gestión del conocimiento y desempeño organizacional

- **Relación KM–desempeño**

La literatura académica ha establecido de manera consistente una relación positiva entre la gestión del conocimiento y el desempeño organizacional. La gestión sistemática del conocimiento permite a las organizaciones mejorar la calidad de sus decisiones, optimizar procesos internos y fortalecer su capacidad de respuesta ante entornos competitivos y cambiantes (Kassa & Ning, 2023a). Entonces, el conocimiento deja de ser un recurso pasivo y se dispone a convertirse en un factor determinante del rendimiento de la organización.

Diversos estudios empíricos señalan que aquellas organizaciones que conservan prácticas maduras de gestión del conocimiento presentan mejores resultados en términos de eficiencia operativa, innovación y ventaja competitiva sostenible (Nonaka, 1994). Sin embargo, en la literatura también advierte que estos beneficios no se materializan únicamente a través del almacenamiento de información, sino mediante la capacidad de acceder al conocimiento adecuado en el momento oportuno y en el contexto correcto.

Desde esta perspectiva, existen limitaciones de los sistemas documentales tradicionales que afectan directamente la relación entre gestión del conocimiento y el desempeño. La dificultad para localizar información relevante y contextualizada reduce el impacto del conocimiento en la toma de decisiones estratégicas, y, es por ello, que la incorporación de sistemas inteligentes capaces de facilitar el acceso dinámico al conocimiento se alinea con la necesidad de fortalecer el vínculo entre gestión del conocimiento y desempeño organizacional, objetivo central de esta investigación.

- **Capacidades dinámicas como mecanismo mediador**

Las capacidades dinámicas se definen como la habilidad de una organización para integrar, construir y reconfigurar sus recursos y competencias en respuesta a cambios del entorno (Teece, 2007). En este sentido, el conocimiento actúa como un insumo clave para el desarrollo de dichas capacidades.

La literatura sugiere que la gestión del conocimiento no impacta directamente en el desempeño, sino que lo hace a través de capacidades dinámicas que permiten a la organización identificar oportunidades, tomar decisiones informadas y adaptar sus estrategias de manera oportuna. Estas capacidades incluyen la detección de cambios, la asimilación de nueva información y la reconfiguración de procesos organizacionales (Teece, 2007).

2. Estándares y marcos de referencia para la Gestión del Conocimiento

La Gestión del Conocimiento ha evolucionado desde enfoques conceptuales y prácticas aisladas hacia la adopción de estándares y marcos de referencia que buscan formalizarla como una capacidad organizacional transversal, además estos proporcionan estructuras sistemáticas para definir procesos, roles, métricas y mecanismos de mejora continua, permitiendo que el conocimiento sea gestionado de manera consistente y alineada con la estrategia organizacional. El ámbito actual que es caracterizado por la digitalización y el crecimiento exponencial de información no estructurada, y dichos estándares son relevantes pues sirven como base para la incorporación de soluciones tecnológicas avanzadas, incluyendo arquitecturas de inteligencia artificial y sistemas inteligentes orientados a la gestión del conocimiento.

2.1 Norma ISO 30401: Sistemas de Gestión del Conocimiento

- **Enfoque sistémico de la gestión del conocimiento**

Las normas internacionales, particularmente la ISO 30401, señalan que la gestión del conocimiento debe implementarse como un sistema de gestión formal e integrado, alineado con la estrategia organizacional. El conocimiento se entiende como un recurso que fluye a través de procesos interrelacionados y gobernados, los cuales deben diseñarse, operarse y evaluarse de forma sistemática. Pues, no se trata simplemente de almacenar información, sino de estructurar un modelo que garantice su circulación y aprovechamiento dentro de la organización. La gestión del conocimiento no se puede asumir como una iniciativa aislada ni únicamente tecnológica, sino más bien como un sistema organizacional que articula liderazgo, cultura, procesos y tecnologías de soporte de manera coherente. (International Organization for Standardization, 2018a).

La ISO 30401 enfatiza que un sistema de gestión del conocimiento debe integrarse con otros sistemas de gestión existentes y con la arquitectura organizacional, lo que permite asegurar coherencia, trazabilidad y sostenibilidad. Entonces, el enfoque sistémico se convierte en un requisito para reducir la fragmentación documental y así facilitar el acceso consistente al conocimiento crítico. La literatura académica adicionalmente respalda esta visión al señalar que los sistemas de gestión del conocimiento más maduros son aquellos que se encuentran formalizados y alineados con marcos normativos y de gobernanza organizacional (Alavi & Leidner, 2001; Kassa & Ning, 2023a).

Así, la incorporación de tecnologías basadas en inteligencia artificial debe responder a los principios establecidos por las normas ISO, actuando como habilitadores del sistema y no solo como sustitutos de los procesos de gestión del conocimiento. Investigaciones recientes indican que los sistemas inteligentes orientados al análisis documental deben diseñarse para operar dentro de un marco de gestión formal, asegurando control, coherencia y orientado a la alineación estratégica (Mahadevkar et al., 2024a). Por lo tanto, el sistema agéntico propuesto en esta investigación se

concibe como un componente tecnológico integrado a un sistema de gestión del conocimiento conforme a los lineamientos de la ISO 30401.

- **Principios, procesos y mejora continua**

La norma ISO 30401 define un amplio conjunto de principios y procesos que orientan la implementación efectiva de la gestión del conocimiento en las organizaciones. Estos procesos abarcan la identificación, creación, captura, organización, protección, compartición y aplicación del conocimiento, los cuales deben ser formalizados y documentados para garantizar su consistencia y repetibilidad (International Organization for Standardization, 2018a). Ya que la gestión del conocimiento se estructura como un sistema operativo, esta permite transformar información dispersa en conocimiento útil para la toma de decisiones.

Un elemento central de la ISO 30401 es la mejora continua, entendida como la capacidad del sistema para adaptarse a cambios organizacionales, tecnológicos y contextuales. La norma establece que las organizaciones deben evaluar el desempeño, de forma constante, al sistema de gestión del conocimiento, identificar brechas y aplicar acciones correctivas y preventivas. La mejora continua se convierte en una alternativa sumamente importante para asegurar la vigencia y efectividad del sistema a lo largo del tiempo, especialmente en entornos caracterizados por el crecimiento acelerado de información no estructurada.

Este enfoque resulta particularmente relevante al incorporar soluciones basadas en inteligencia artificial generativa. Estudios recientes sobre evaluación de sistemas de recuperación aumentada dices que el desempeño de estos sistemas debe ser medido y ajustado continuamente para así garantizar coherencia, precisión y confiabilidad (Saad-Falcon et al., 2024a; Yu et al., 2024a). La mejora continua propuesta por la ISO 30401 se alinea directamente con la necesidad de optimizar de manera iterativa los sistemas agénticos orientados a la gestión del conocimiento documental.

- **Medición y trazabilidad del conocimiento**

La ISO 30401 establece que la medición y la trazabilidad son requisitos esenciales para demostrar la eficacia del sistema de gestión del conocimiento y su contribución a los objetivos organizacionales. La norma claramente enfatiza la necesidad de definir indicadores que permitan evaluar el desempeño de los procesos de gestión del conocimiento, así como son la calidad y accesibilidad del conocimiento gestionado (International Organization for Standardization, 2018a). Es entonces que esta medición se convierte en un instrumento fundamental para sustentar la toma de decisiones basada en evidencia.

Diversos estudios señalan que cuando no existe una adecuada trazabilidad, las organizaciones ven reducida su capacidad para validar el conocimiento que sustenta decisiones estratégicas y operativas (Sternad Zabukovšek et al., 2023a). En la práctica esto impacta directamente en la confianza y en la transparencia del sistema. Por eso la trazabilidad del conocimiento resulta clave, ya que permite identificar de dónde proviene, cómo ha evolucionado y de qué manera se utiliza dentro de los procesos organizacionales. En otras palabras, no se trata solo de un requisito técnico, sino de un elemento fundamental para la gobernanza y el cumplimiento normativo.

En los sistemas inteligentes basados en recuperación aumentada, la trazabilidad cobra todavía mayor importancia. Se dice sobre la evaluación de RAG que vincular de manera explícita las respuestas generadas con sus fuentes documentales originales constituye un factor decisivo para disminuir errores y aumentar la confiabilidad del sistema (Saad-Falcon et al., 2024a; Yu et al., 2024a). Dado esto, los criterios de medición y trazabilidad establecidos por la ISO 30401 ofrecen un marco normativo sólido para valorar el sistema agéntico propuesto en esta investigación y asegurar su coherencia con las prácticas internacionales de gestión del conocimiento.

2.2 TOGAF y la gestión del conocimiento en la arquitectura empresarial

TOGAF es, en la práctica, uno de los marcos de referencia más usados para diseñar y gestionar la arquitectura empresarial porque propone principios, métodos y buenas prácticas que buscan alinear la estrategia del negocio con los sistemas de información y con las capacidades organizacionales. En esta tesis su inclusión cobra sentido ya que permite entender la gestión del conocimiento no como un proyecto aislado o meramente tecnológico, sino como una capacidad estructural, integrada y gobernada dentro de la organización. Así facilita relacionarla directamente con el desempeño, la toma de decisiones y la sostenibilidad organizacional en el tiempo.

- **Arquitectura empresarial y capacidades organizacionales**

La arquitectura empresarial puede entenderse como un marco estructurado que ayuda a comprender y diseñar las capacidades organizacionales necesarias para ejecutar la estrategia y afrontar entornos cada vez más complejos. TOGAF se define como la representación organizada de los componentes del negocio, la información, las aplicaciones y la tecnología, junto con sus interrelaciones, con el propósito de facilitar el logro de los objetivos organizacionales (The Open Group, 2022a). Visto así, la arquitectura empresarial no se reduce a un ejercicio solamente técnico; más bien, funciona como un mecanismo que permite identificar, desarrollar y gobernar capacidades organizacionales clave de manera coherente con la estrategia.

Desde esta perspectiva, las capacidades organizacionales representan lo que la organización es capaz de hacer de manera consistente y repetible. La literatura destaca que dichas capacidades emergen de la integración coordinada de procesos, personas, información y tecnologías, siendo el conocimiento un componente transversal que influye directamente en su desempeño (Alavi & Leidner, 2001). De este modo, la arquitectura empresarial se convierte en un habilitador

fundamental para estructurar la gestión del conocimiento como una capacidad organizacional alineada con la estrategia y los procesos del negocio.

En el contexto de la transformación digital, TOGAF adquiere relevancia al proporcionar un lenguaje común y un método sistemático para integrar iniciativas de gestión del conocimiento con soluciones tecnológicas avanzadas. Investigaciones recientes señalan que la ausencia de una arquitectura empresarial clara limita la efectividad de los sistemas de gestión del conocimiento, especialmente cuando estos incorporan tecnologías basadas en inteligencia artificial y análisis documental a gran escala (Sternad Zabukovšek et al., 2023a). El enfoque, en síntesis, de una arquitectura empresarial propuesto por TOGAF constituye un pilar conceptual para el diseño del sistema agéntico desarrollado en esta investigación.

- **Dominios funcionales según TOGAF**

Desde la óptica de la gestión del conocimiento, los dominios definidos por TOGAF resultan especialmente relevantes para abordar la diversidad y heterogeneidad de los activos documentales:

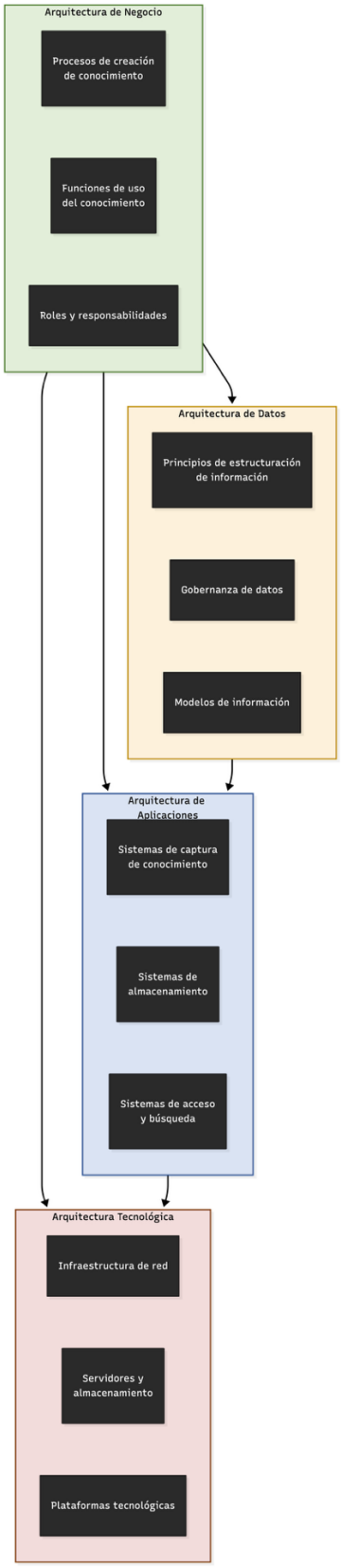


Figura 1. Estructura de dominios según TOGAF

Desde la perspectiva de TOGAF, los dominios arquitectónicos cumplen roles específicos que en conjunto permiten evaluar y diseñar sistemas de gestión del conocimiento integrados y coherentes con la organización (The Open Group, 2022a). Pues, la arquitectura de negocio delimita los procesos y funciones donde el conocimiento se crea y utiliza en la práctica. Además, que la arquitectura tecnológica por su parte provee la infraestructura necesaria para que todo el sistema pueda operar. En cuanto a la arquitectura de datos esta define los principios que orientan cómo se estructura y gobierna la información dentro de la organización. Finalmente, la arquitectura de aplicaciones determina qué sistemas respaldan la captura el almacenamiento y el acceso al conocimiento.

La alineación entre estos dominios es un factor crítico para el éxito de iniciativas de gestión del conocimiento, especialmente en entornos donde predominan documentos no estructurados y múltiples fuentes de información (Mahadevkar et al., 2024b). Entonces, la estructuración de la muestra documental por dominios funcionales, como se plantea en esta investigación, encuentra sustento teórico en la descomposición propuesta por TOGAF, garantizando una evaluación más representativa del sistema agéntico.

- **Integración de KM como capacidad organizacional**

TOGAF plantea que las capacidades organizacionales son el puente entre lo que una organización se propone y lo que efectivamente ejecuta, razón por la cual deben modelarse y gobernarse dentro de la arquitectura empresarial. Desde ahí la gestión del conocimiento puede entenderse como una capability transversal que atraviesa múltiples dominios funcionales y procesos de negocio. Integrándola en la arquitectura empresarial se logra coherencia, alineación estratégica y sostenibilidad en el tiempo (The Open Group, 2022a).

La norma ISO 30401 apunta en la misma dirección al exigir que la gestión del conocimiento funcione como un sistema formal con roles, procesos y métricas definidos. La convergencia entre ambos marcos permite entonces conceptualizar la gestión del conocimiento como una capacidad que puede gobernarse y medirse. Varios autores señalan que esta integración facilita además la incorporación de tecnologías emergentes al ofrecer un marco claro para ubicarlas dentro del ecosistema organizacional (Kassa & Ning, 2023a).

En el contexto de esta investigación esa integración habilita el diseño de un sistema agéntico basado en RAG que opera como un componente arquitectónico alineado con los dominios funcionales y los procesos de negocio. El enfoque propuesto no solo atiende los retos técnicos de la gestión documental, sino que involucra marcos normativos y arquitectónicos reconocidos internacionalmente lo que fortalece la validez y aplicabilidad del sistema.

3. Gestión documental y conocimiento no estructurado.

La gestión documental ocupa un lugar central dentro de la gestión del conocimiento organizacional sobre todo cuando gran parte del conocimiento explícito vive en documentos no estructurados, con el crecimiento en volumen y diversidad de la información los documentos han dejado de ser simples soportes administrativos para convertirse en activos estratégicos. Gestionarlos bien condiciona directamente el acceso al conocimiento existente y la posibilidad de reutilizarlo para generar valor.

El problema es que los documentos organizacionales rara vez son homogéneos. Distintos formatos, estructuras internas dispares, calidad variable y diferentes grados de digitalización hacen difícil tratarlos de manera uniforme. Eso limita lo que los sistemas tradicionales pueden hacer y el resultado habitual es conocimiento fragmentado que no llega a los procesos de decisión ni al aprendizaje organizacional.

Esta sección aborda precisamente la gestión documental vista desde el conocimiento no estructurado, revisando tanto los problemas organizacionales que genera como los desafíos técnicos del diseño y análisis de documentos. Con ello se sienta la base para entender por qué hacen falta enfoques más avanzados como el análisis de diseño documental y los sistemas de parsing inteligente capaces de convertir documentos complejos en fuentes de conocimiento accesibles y aprovechables.

3.1 Problemática de la gestión documental en las organizaciones

Planteémonos la siguiente pregunta ¿Cómo puede una organización gestionar su conocimiento cuando sus documentos ni siquiera hablan entre sí? Esa es la realidad de muchas organizaciones contemporáneas. Correos electrónicos, informes técnicos, presentaciones y archivos escaneados coexisten sin estructura común ni criterios homogéneos de clasificación. La fragmentación es el resultado de años de proliferación de fuentes, formatos diversos y repositorios distribuidos que nunca fueron diseñados para integrarse. Consolidar una memoria organizacional coherente se vuelve entonces una tarea compleja, y con ella se resiente tanto la toma de decisiones como la eficiencia de los procesos internos (Alavi & Leidner, 2001; Davenport & Prusak, 1998).

A esto se suma un problema quizás menos visible pero igualmente costoso. Duplicidad, obsolescencia y pérdida progresiva de conocimiento experto son consecuencias directas de la ausencia de controles formales sobre versiones, vigencia y autoría. El conocimiento explícito pierde confiabilidad. Los documentos antiguos donde se tiene el conocimiento de expertos que ya no están se vuelven difíciles de localizar y prácticamente imposibles de reutilizar. (Nonaka, 1994; Walsh & Ungson, 1991).

Los sistemas tradicionales de búsqueda no resuelven esto, lo agravan. Basados en coincidencias léxicas y metadatos definidos manualmente, son incapaces de capturar el significado semántico ni

el contexto del contenido. Los resultados son extensos. Poco relevantes. El tiempo de búsqueda crece y los usuarios simplemente dejan de usarlos. Interpretar estructuras internas, relaciones conceptuales o jerarquías de información está fuera de su alcance, lo que deja sin aprovechar una proporción significativa de los activos informacionales de cualquier organización (Becerra-Fernandez & Sabherwal, 2015).

Sobre lo anterior hay una debilidad estructural que pocas organizaciones reconocen explícitamente: la gestión documental y la estrategia de conocimiento operan desconectadas. La primera se concibe como función operativa, centrada en almacenar y custodiar archivos. La segunda aspira a crear, transferir y aplicar conocimiento. Entre ambas hay un abismo. Los documentos se tratan como artefactos aislados y no como elementos vivos dentro de los flujos de conocimiento organizacional, lo que frena el aprendizaje y la innovación (Jennex & Olfman, 2005).

La conclusión. Gestionar documentos como se ha hecho hasta ahora ya no es suficiente. El volumen, la complejidad y la diversidad de la información organizacional exigen enfoques más avanzados, orientados a la comprensión semántica y al acceso contextualizado al conocimiento.

Ese es precisamente el punto de partida de esta investigación: entender la gestión documental no como un proceso administrativo sino como un componente crítico para construir arquitecturas de conocimiento robustas y con valor estratégico real.

3.2 Tipologías documentales y complejidad estructural

No todos los documentos son iguales, y esa diferencia importa más de lo que parece. Las organizaciones gestionan una amplia variedad de materiales que difieren en forma, estructura interna y complejidad semántica. Esas diferencias tienen efecto directo en la capacidad de cualquier sistema para indexar, recuperar y explotar el conocimiento que contienen. La tipología documental se convierte así en un factor crítico al momento de diseñar soluciones tecnológicas para gestionar conocimiento no estructurado (Becerra-Fernandez & Sabherwal, 2015; Sternad Zabukovšek et al., 2023a).

Los documentos en PDF y DOCX son los más comunes en entornos organizacionales. Parecen simples. Sin embargo, encabezados, tablas, listas y elementos gráficos introducen una estructura implícita que los sistemas tradicionales de búsqueda no reconocen. (Alavi & Leidner, 2001; Mahadevkar et al., 2024a).

Las presentaciones PPTX plantean una complejidad distinta. El conocimiento no fluye de forma lineal, sino que se distribuye en diapositivas con jerarquías visuales, fragmentos de texto, gráficos y esquemas conceptuales. Aquí la relación entre lo visual y lo textual es lo que da sentido al contenido. Analizarlas correctamente exige enfoques capaces de leer tanto la estructura lógica como la disposición visual para preservar el conocimiento representado (Sternad Zabukovšek et al., 2023a).

Distinto es el caso de los documentos escaneados. Procesados mediante reconocimiento óptico de caracteres, el reto es la calidad de la digitalización. Errores de reconocimiento, pérdida de formato y ausencia de metadatos estructurales se acumulan y el resultado es conocimiento difícil de explotar. La confiabilidad de los sistemas de recuperación de información se resiente directamente (Mahadevkar et al., 2024a).

¿Y qué ocurre con los PDF que contienen tablas complejas? Representan quizás el reto estructural más exigente. Filas, columnas y relaciones implícitas entre datos no siempre sobreviven los procesos de extracción automática. Se pierden. Y con ellas se pierde conocimiento crítico, especialmente en contextos financieros, operativos o analíticos donde precisamente ese tipo de información concentra el mayor valor. Esto evidencia la necesidad de técnicas avanzadas de análisis documental alineadas con los objetivos de esta investigación (Becerra-Fernandez & Sabherwal, 2015; Mahadevkar et al., 2024a). En Figura 2, se puede observar una representación sobre las Tipologías Documentales.

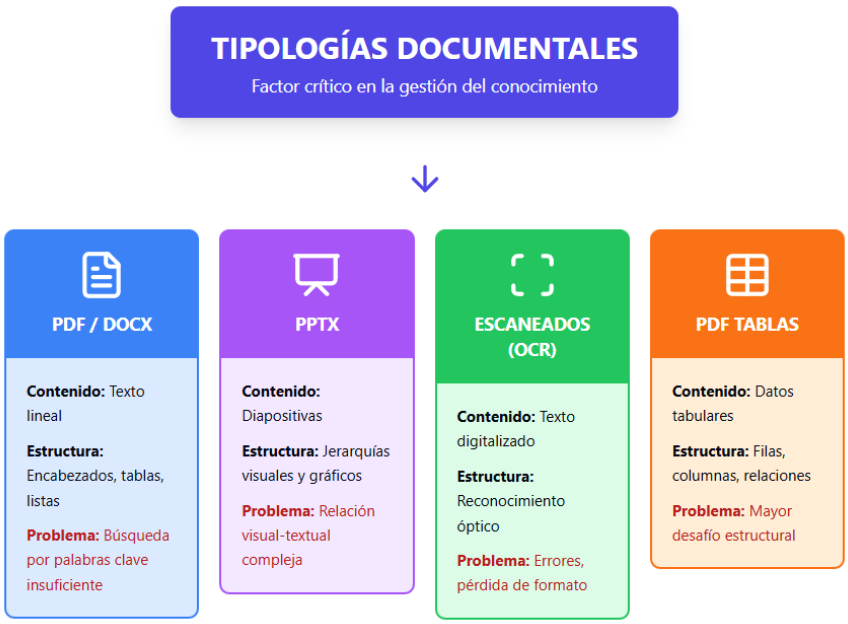


Figura 2. Tipos de documentos empresariales

3.3 Análisis de Diseño de Documentos (Document Layout Analysis)

Extraer texto de un documento no es lo mismo que entenderlo. Esa distinción, aparentemente simple, es el punto de partida del análisis de diseño documental. La creciente complejidad de los materiales que circulan en las organizaciones ha dejado en evidencia que la extracción directa de texto es insuficiente para comprender y explotar el conocimiento que contienen. Encabezados, secciones, párrafos, tablas, listas y la disposición espacial del contenido no son detalles estéticos. Son estructura. Y la estructura es la que preserva el significado, el contexto y las relaciones

semánticas del documento, especialmente en archivos corporativos de alta complejidad como informes técnicos, normativas y reportes organizacionales (Wu & others, 2023; Xu et al., 2020).

Nos planteamos la siguiente pregunta ¿Qué cambia cuando se incorpora ese análisis estructural? Principalmente, la segmentación mejora. La identificación de unidades informativas relevantes se vuelve más precisa. Y la recuperación posterior del conocimiento gana en coherencia y confiabilidad. El análisis del diseño documental se convierte así en un habilitador clave, no solo para extraer conocimiento sino para evaluar qué tan bien los sistemas lo interpretan de forma automática. Estudios recientes muestran que combinar técnicas de análisis estructural con modelos avanzados de lenguaje eleva significativamente la precisión de los sistemas inteligentes de gestión del conocimiento (Mahadevkar et al., 2024a; Minaee et al., 2024a).

La sintetiza los tres componentes centrales del análisis de diseño de documentos: la descomposición visual y estructural, la relevancia para la extracción de conocimiento y la evaluación de parsing documental. Como se puede observar cada componente articula dimensiones específicas que van desde la identificación de títulos, tablas y jerarquías de contenido hasta la segmentación inteligente y la evaluación de fidelidad del conocimiento extraído. En conjunto estos elementos ilustran por qué el análisis estructural no es un paso accesorio sino la base sobre la que se construyen sistemas inteligentes de recuperación aumentada capaces de operar con precisión en entornos organizacionales complejos.

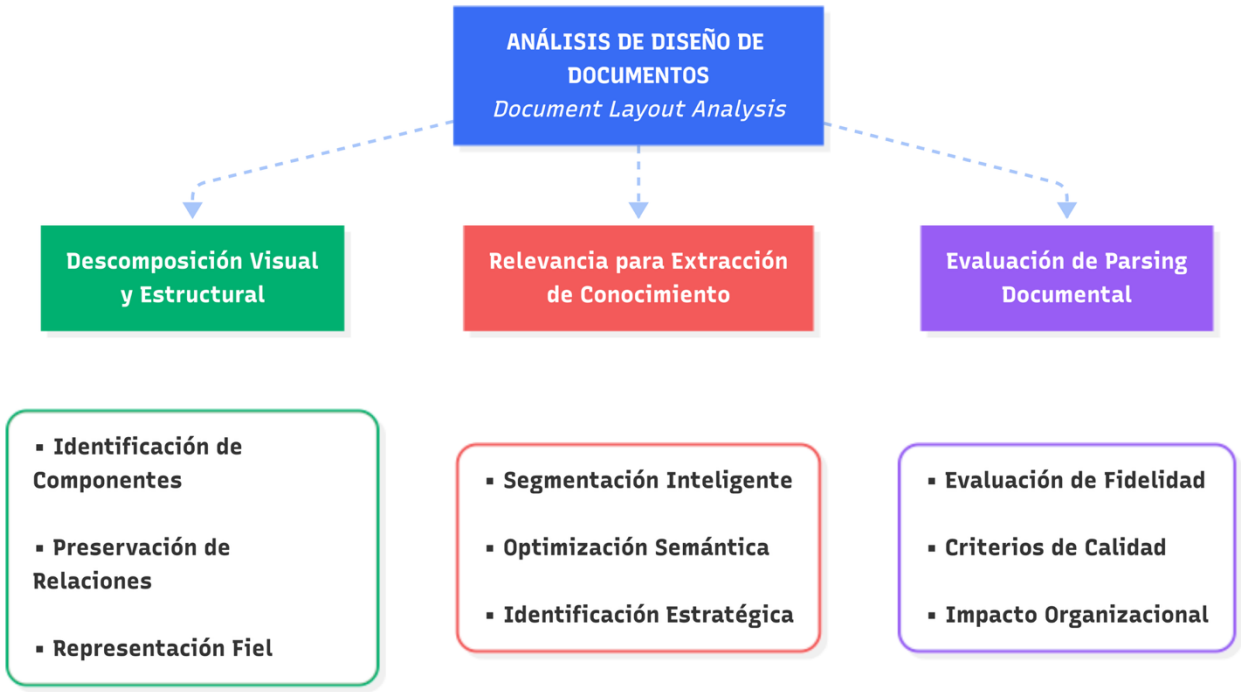


Figura 3. Análisis de diseño de documentos (DLA)

- **Descomposición visual y estructural**

Un documento no es solo texto. Detrás de cada página hay una arquitectura de títulos, encabezados, párrafos, tablas, listas e imágenes que organiza el contenido y le da sentido. La descomposición visual y estructural es precisamente el proceso que permite identificar y separar esos componentes preservando las relaciones jerárquicas y espaciales entre ellos. Sin ese paso el conocimiento explícito contenido en los documentos llega distorsionado a los sistemas de gestión, especialmente cuando se trabaja con información no estructurada (Xu et al., 2020).

¿Cómo funciona esto desde el punto de vista computacional? El análisis se apoya en características como alineación, la proximidad, el tamaño y formato para leer la disposición espacial de los elementos dentro del documento. Lo que se busca es diferenciar secciones lógicas y unidades semánticas que el orden lineal del texto extraído simplemente no refleja. Perder esas relaciones estructurales durante el parsing tiene un costo directo, y es que la calidad de la información recuperada cae y los sistemas inteligentes interpretan mal el conocimiento (Wu & others, 2023).

En esta investigación ese problema es especialmente concreto. La diversidad de tipologías documentales analizadas, textuales, presentaciones, escaneados y archivos con tablas complejas, hace que la segmentación estructural no sea opcional sino un requisito previo. Sin ella no hay fragmentos documentales coherentes. Sin fragmentos coherentes no hay almacenamiento vectorial confiable. Estudios recientes confirman que una segmentación estructural adecuada mejora la precisión de la recuperación semántica y reduce errores en sistemas basados en modelos de lenguaje, lo que resulta clave para arquitecturas RAG y sistemas agénticos orientados a la gestión del conocimiento organizacional (Minaee et al., 2024a).

- **Relevancia para extracción de conocimiento**

La estructura interna de un documento condiciona directamente la calidad del conocimiento que se puede extraer de él. Los sistemas que interpretan únicamente texto plano tienden a omitir conceptos clave, relaciones semánticas, datos críticos y reglas implícitas que sostienen los procesos de negocio. El análisis del diseño documental cobra relevancia precisamente ahí, al contribuir de forma directa a la identificación de esas unidades informativas significativas que una lectura lineal del contenido no logra capturar. La literatura especializada es clara al respecto, pues, la extracción basada solo en texto plano limita la calidad del conocimiento recuperado y su utilidad para la toma de decisiones organizacionales (Mahadevkar et al., 2024a; Xu et al., 2020).

La coherencia semántica del contenido se preserva cuando la información visual y estructural se integra durante el proceso de extracción. Reportes técnicos, normativas y archivos con tablas densas son casos donde el significado depende de la disposición espacial y jerárquica de los elementos, no solo de las palabras. Los enfoques que combinan análisis estructural con modelos avanzados de lenguaje incrementan la precisión en la identificación de conocimiento relevante y

reducen ambigüedades durante la recuperación contextual, aspecto fundamental para sistemas basados en recuperación aumentada (Minaee et al., 2024a; Yu et al., 2024a).

El sistema agéntico propuesto en esta investigación depende directamente de esa capacidad. Una extracción de conocimiento que respeta la estructura original del documento mejora la calidad de los embeddings, fortalece la recuperación semántica y favorece la generación de respuestas fundamentadas en evidencia documental real. El Document Layout Analysis se consolida así, como un componente clave para habilitar sistemas de gestión del conocimiento confiable, explicable y alineado con los objetivos de precisión y coherencia definidos en el diseño experimental del estudio.

- **Relación con evaluación de parsing documental**

La evaluación del parsing documental constituye un componente crítico para determinar la efectividad de los sistemas orientados a la extracción y gestión del conocimiento. No se limita a la conversión de documentos en texto legible por máquina, sino que implica la correcta identificación de la estructura lógica y visual del contenido, así como la preservación de las relaciones semánticas entre los distintos elementos que conforman el documento. Entonces, la calidad del parsing influye directamente en la fidelidad del conocimiento extraído y en su confiabilidad, los sistemas que dependen de estos procesos para la recuperación y generación de información (Mahadevkar et al., 2024a).

La evaluación del parsing documental exige ir más allá de las métricas tradicionales de exactitud en la extracción de texto. La correcta segmentación estructural la identificación de secciones relevantes y la coherencia de los fragmentos generados son criterios igualmente determinantes. Los errores en el parsing inicial no se quedan ahí: se propagan a lo largo del pipeline afectando la calidad de los embeddings la recuperación contextual y la generación final de respuestas. Mecanismos de evaluación sistemáticos y trazables no son opcionales sino una necesidad justificada por la evidencia (Saad-Falcon et al., 2024a; Yu et al., 2024a).

La relación entre el análisis del diseño de documentos y la evaluación del parsing documental resulta fundamental en esta investigación para validar el desempeño del sistema agéntico propuesto. La estructura original del documento es el punto de referencia desde el cual se evalúa la calidad del parsing: desde ahí se identifican limitaciones técnicas y oportunidades de optimización concretas. Los fragmentos almacenados y recuperados representan fielmente el conocimiento explícito solo cuando esa evaluación es rigurosa. La robustez del sistema se construye desde ese nivel y con ella la capacidad de cumplir los umbrales de precisión y coherencia establecidos en el diseño experimental del estudio.

4. Inteligencia Artificial Generativa en contextos organizacionales

La Inteligencia Artificial Generativa no es una versión más sofisticada de la automatización tradicional. Su integración en el ámbito corporativo representa una ruptura. Las máquinas pasan a realizar tareas cognitivas complejas: sintetizar información, generar código, razonar sobre datos no estructurados. La productividad organizacional se redefine desde sus bases (Chui, Hazan, et al., 2023).

4.1 Evolución de la IA aplicada a la gestión del conocimiento

La gestión del conocimiento ha recorrido un largo camino. Los repositorios pasivos de documentos han dado paso a ecosistemas interactivos donde la información no solo se almacena, sino que se transforma. La eficacia de la KM en las organizaciones modernas depende hoy de la capacidad del sistema para convertir conocimiento tácito en explícito, mismo que habilita una recuperación que va mucho más allá de la simple coincidencia de palabras (Kassa & Ning, 2023b).

- **Motores de búsqueda vs. IA generativa**

Los motores de búsqueda convencionales operan bajo una lógica de recuperación estadística. Identifican dónde está la información. La IA generativa hace algo distinto, ya que, la interpreta. La comprensión semántica reemplaza al keyword matching y el resultado es información presentada de forma contextualizada, no solo localizada.

- **Limitaciones de la búsqueda por palabras clave**

El volumen y la heterogeneidad de los datos organizacionales exponen con claridad los límites de los sistemas tradicionales de gestión documental. La búsqueda por palabras clave falla al enfrentarse a documentos no estructurados porque no captura la intención del usuario ni las relaciones latentes entre distintas piezas de información. Así, el conocimiento organizacional se fragmenta. Se vuelve inaccesible. Y los procesos de decisión pierden el soporte que necesita (Mahadevkar et al., 2024).

Tabla 1. Comparativa de los motores de búsqueda y los sistemas de IA generativa.

Dimensión	***Motores de Búsqueda Tradicionales***	***Sistemas basados en IA Generativa***
Mecanismo de consulta	Palabras clave exactas (Keywords).	Lenguaje natural y contexto semántico.
Procesamiento	Indexación de texto plano.	Embeddings vectoriales y razonamiento.
Resultado final	Lista de documentos jerarquizados.	Respuesta directa, síntesis o acción.
Contexto	Limitado al documento individual (Sternad Zabukovšek et al., 2023).	Relacional (conecta múltiples fuentes) (Abou Ali & coautores, 2025).

4.2 Inteligencia Artificial Generativa y Large Language Models

Los Modelos de Lenguaje de Gran Escala constituyen la base tecnológica sobre la que opera la IAG. Redes neuronales profundas entrenadas en corpus masivos de texto, los LLM abordan una amplia gama de tareas de procesamiento de lenguaje natural mediante la predicción de secuencias probabilísticas (Minaee et al., 2024). Su escala no es un detalle menor. Es precisamente lo que les permite generalizar, razonar y producir contenido con un nivel de coherencia que los modelos anteriores no lograban.

- **Capacidades de los LLM**

Superar la mera generación de texto es lo que distingue a los LLM en entornos organizacionales. Los modelos multimodales actuales integran la comprensión de elementos visuales como gráficos, tablas y firmas con el contenido textual habilitando una auditoría documental verdaderamente integral. Según Wu et al. (2023), un motor de inferencia capaz de extraer datos estructurados desde fuentes caóticas y heterogéneas es lo que el modelo se convierte cuando opera en ese nivel.

- **Riesgos: Alucinación y falta de trazabilidad**

El poder de los LLM viene acompañado de riesgos que no pueden ignorarse. Información factualmente incorrecta pero lingüísticamente coherente es lo que produce el fenómeno conocido como alucinación, uno de los riesgos técnicos más críticos identificados en la literatura (Bai et al.,

2024b) En gestión documental las consecuencias son concretas y graves. Una fecha equivocada o un monto incorrecto bastan para invalidar un proceso legal o financiero. La trazabilidad se convierte entonces en el mecanismo de defensa central: cada salida del modelo debe estar anclada a una referencia fuente verificable, especialmente en arquitecturas que no incorporan métodos de alineación humana como RLHF.

4.3 Barreras de adopción de IA generativa en organizaciones

La madurez digital de una organización determina en gran medida el éxito o el fracaso de estas implementaciones. Sin una infraestructura que soporte el ciclo de vida de los documentos el despliegue de IA generativa tiende al fracaso antes de mostrar resultados (Sternad Zabukovšek et al., 2023b) La tecnología no llega a un vacío, sino, llega a una organización con sistemas, procesos y niveles de preparación muy concretos.

- **Desafíos de Integración y Seguridad**

Operar en silos es uno de los problemas más frecuentes de los LLM en entornos organizacionales. Sin acceso al contexto histórico de la empresa la integración con los sistemas de gestión del conocimiento existentes se vuelve superficial. A esto se suman riesgos de privacidad y seguridad que los departamentos de cumplimiento no pueden ignorar, estos son la exposición de datos sensibles a través de prompts representa una preocupación de primer orden en cualquier despliegue corporativo (Yao et al., 2024b).

- **Necesidad de arquitecturas controladas**

Transitar de modelos de propósito general hacia arquitecturas controladas es la respuesta más coherente a esas barreras. La autonomía creativa no deseada cede paso a la precisión técnica cuando la generación queda bastante limitada por la base de conocimientos organizacional. Marcos de gobernanza como el estándar NIST AI 600-1 apuntan exactamente en esa dirección al exigir que la IA generativa se despliegue bajo protocolos de seguridad robustos que garanticen control, trazabilidad y alineación con los objetivos institucionales (Y. Liu et al., 2022; Qin et al., 2024; Yao et al., 2024b).

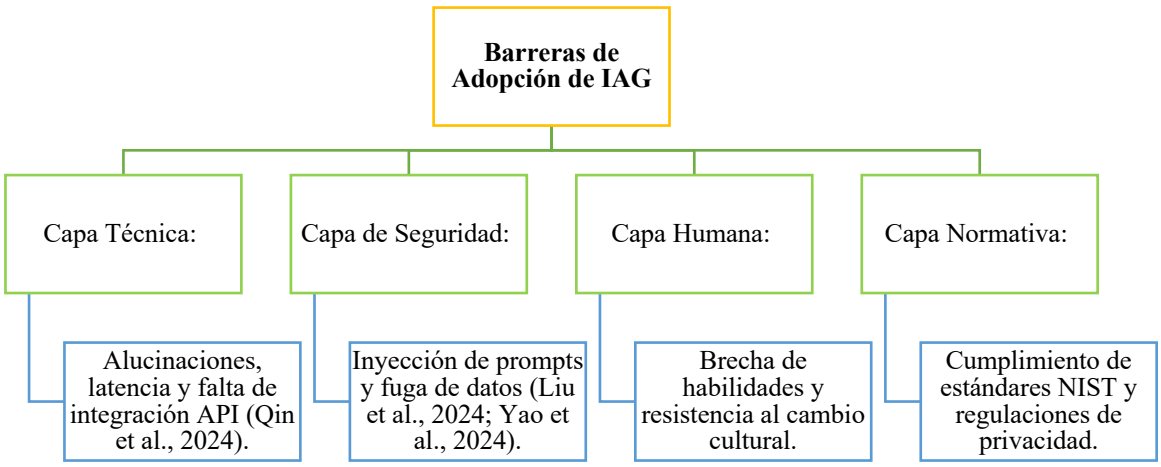


Figura 4. Barreras de adopción de IAG

A pesar del notable avance de los modelos LLM, su uso en entornos organizacionales plantea preguntas relevantes respecto a la fiabilidad del conocimiento generado. Desde una perspectiva epistemológica, los LLM no operan sobre una base de conocimiento estructurada ni verifican la veracidad de la información que producen, sino que generan respuestas a partir de patrones estadísticos aprendidos en su entrenamiento. Como consecuencia, los modelos pueden producir textos altamente coherentes y plausibles que, sin embargo, no necesariamente corresponden con información factual. Este fenómeno, conocido como *alucinación* (*hallucination*), constituye una de las principales limitaciones de los sistemas basados en inteligencia artificial generativa, especialmente en contextos organizacionales donde las decisiones requieren información confiable y verificable.

En este contexto, la incorporación de inteligencia artificial generativa en procesos de gestión del conocimiento requiere también abordar aspectos relacionados con la gobernanza de los dominios de conocimiento organizacional. La gobernanza implica establecer mecanismos, roles y responsabilidades que permitan garantizar la calidad, trazabilidad y confiabilidad de la información utilizada por los sistemas inteligentes.

Diversos estudios citados anteriormente, destacan la necesidad de integrar prácticas como la referencia explícita a las fuentes documentales utilizadas por el sistema, la validación del contenido generado y la supervisión humana en procesos críticos de toma de decisiones. De esta manera, la adopción de modelos de lenguaje en entornos organizacionales debe entenderse como un proceso socio-técnico, en el que las capacidades tecnológicas se complementan con mecanismos de control

y validación que aseguren la transparencia y confiabilidad del conocimiento recuperado para la generación de respuestas.

5. Generación Aumentada por Recuperación (RAG)

Cerrar la brecha entre el conocimiento generalista de los modelos de lenguaje y las necesidades de precisión factual de las organizaciones es precisamente lo que la arquitectura RAG ha logrado consolidar como solución técnica predominante. Además, el paradigma optimiza la respuesta del modelo recuperando fragmentos relevantes de un corpus externo antes de proceder con la generación del texto (Gao et al., 2024). El razonamiento lo procesa el LLM. La memoria vive en bases de datos externas. Esa separación permite operar con información actualizada sin necesidad de costosos procesos de reentrenamiento.

El flujo de trabajo técnico se articula en tres etapas críticas, estas son indexación recuperación y aumento (Yu et al., 2024b). Los documentos organizacionales se transforman primero en representaciones vectoriales densas. Ante una consulta el sistema identifica los segmentos con mayor similitud semántica. Y, finalmente el modelo integra esos fragmentos en su ventana de contexto para generar una respuesta fundamentada. En sistemas que prescinden de alineación por refuerzo humano la veracidad no depende del comportamiento del modelo sino de la evidencia documental recuperada.

5.1 Beneficios de RAG frente a modelos generativos puros

La superioridad de RAG sobre los modelos puramente generativos reside en la mitigación sistemática de la incertidumbre. Las alucinaciones en modelos de gran escala suelen derivar de lagunas en el conocimiento paramétrico o de confusión entre datos similares (Bai et al., 2024b). RAG resuelve eso anclando cada respuesta a un activo de conocimiento específico. La caja negra se convierte en un sistema de consulta auditable y verificable en tiempo real.

La flexibilidad operativa es otra ventaja que los modelos estáticos simplemente no pueden igualar. Un modelo puro requiere procesos de fine-tuning para incorporar nuevos datos. Un pipeline RAG se actualiza instantáneamente al modificar su base de conocimientos externa. Esa capacidad de actualización garantiza que el agente mantenga relevancia factual en entornos normativos o técnicos que cambian con frecuencia proporcionando además citas explícitas que refuerzan la confianza del usuario final (Yu et al., 2024b).

Tabla 2. Comparativa de confiabilidad, Modelos puros vs Arquitectura RAG

Criterio de evaluación	Modelo generativo puro	Arquitectura RAG
Origen de la información	Pesos internos (entrenamiento).	Documentos externos (indexación).
Riesgo de alucinación	Alto (inherentemente probabilístico).	Bajo (anclado a fuentes reales).
Trazabilidad	Nula (no cita fuentes).	Alta (proporciona referencias directas).
Actualización de datos	Requerimiento de reentrenamiento.	Actualización inmediata del índice.
Referencia	(Minaee et al., 2024b)	(Gao et al., 2024; Yu et al., 2024b)

5.2 RAG como habilitador de activos de conocimiento organizacional

La implementación de estas arquitecturas transforma la documentación pasiva en activos de conocimiento dinámicos que cumplen con los objetivos de la gestión del conocimiento moderna. (James et al., 2025) sostienen que RAG es el componente tecnológico que permite a las organizaciones materializar los principios de la norma ISO 30401, al asegurar que el conocimiento no solo sea almacenado, sino aplicado y reutilizado de manera sistemática para la toma de decisiones.

Finalmente, la integración de RAG en los sistemas de gestión documental (DMS) cierra la brecha entre el dato y la acción. (Kassa & Ning, 2023b) destacan que la síntesis de información es el valor agregado más esperado en la gestión pública y privada; en este sentido, un agente basado en RAG actúa como un mediador capaz de consolidar el saber experto de la organización. Al alinear estas capacidades con la madurez digital descrita por (Sternad Zabukovšek et al., 2023b), las organizaciones pueden garantizar un ciclo de vida documental donde la información se actualiza, audita y consume de forma autónoma y segura.

Los enfoques basados en Retrieval-Augmented Generation han surgido como una estrategia para mitigar algunas de las limitaciones de los modelos generativos puros, particularmente el problema de las alucinaciones y la falta de trazabilidad de las respuestas. Al incorporar mecanismos de recuperación de información desde fuentes documentales verificables, los sistemas RAG permiten contextualizar las respuestas generadas por los modelos de lenguaje y vincularlas con evidencias documentales. Sin embargo, su efectividad depende en gran medida de la calidad de los procesos de ingestión, segmentación e indexación de los documentos. Problemas en la estructuración del contenido o en las estrategias de recuperación pueden afectar la relevancia del contexto recuperado y, por tanto, la precisión de las respuestas generadas. Por esta razón, la literatura citada enfatiza la importancia de complementar las arquitecturas RAG con técnicas avanzadas de procesamiento documental y mecanismos de evaluación que permitan garantizar la confiabilidad de los sistemas de consulta basados en inteligencia artificial.

6. Sistemas agénticos basados en IA

La evolución de la Inteligencia Artificial ha trascendido la generación estática de contenido hacia la configuración de agentes autónomos. Según (Xi et al., 2024), un sistema agéntico se define por su capacidad de operar en un ciclo continuo de percepción, razonamiento y acción. A diferencia de los sistemas tradicionales basados en reglas deterministas, los agentes utilizan modelos de lenguaje de gran escala (LLM) como un "cerebro" que les permite interpretar instrucciones ambiguas, descomponer tareas complejas en subtareas y ejecutar acciones en entornos dinámicos para alcanzar objetivos específicos.

La distinción esencial con el software convencional radica en la autonomía del flujo de control existente. Mientras que los sistemas tradicionales siguen rutas de ejecución predefinidas (hard-coded), los sistemas agénticos determinan su propio camino de resolución de problemas basándose en el contexto y las herramientas disponibles. Esta capacidad de auto-planificación es lo que permite que el agente actúe como un mediador inteligente entre el usuario y los activos de conocimiento de la organización porque de esa manera adapta su estrategia de búsqueda y síntesis según la dificultad de la consulta que se haya planteado.

6.1 Arquitecturas agénticas con LLM y RAG agéntico con patrón de routing

La arquitectura de un sistema RAG agéntico ubica al LLM como un orquestador cognitivo encargado de la toma de decisiones, de este modo (Chen et al., 2025) en su propuesta de arquitecturas multi-agente (MAO-ARAG) destacan que la eficiencia no depende de un único proceso de búsqueda, sino de un patrón de enrutamiento dinámico, mismo que en este esquema, el agente evalúa la naturaleza de la consulta para determinar qué herramienta de recuperación es la más adecuada para realizar una búsqueda semántica para conceptos abstractos, una búsqueda léxica para términos técnicos precisos, o una consulta a un GraphRAG para entender relaciones jerárquicas entre documentos.

Este enfoque de enrutamiento dinámico es esencial para garantizar la trazabilidad y la explicabilidad en entornos críticos, de esta manera según (Cook et al., 2025b), el uso de patrones de routing permite al sistema acceder de forma directa a documentos fuente cuando se detecta una consulta altamente sensible, reduciendo así la latencia y maximizando la precisión real, de este modo al segmentar las fuentes de información y los métodos de acceso, la arquitectura agéntica se convierte en un sistema modular y auditable que justifica cada paso de su razonamiento lógico superando las limitaciones de los sistemas RAG que son estáticos.

6.2 Sistemas agénticos para la gestión del conocimiento organizacional

La integración de agentes en la arquitectura empresarial permite la automatización de todo el ciclo completo del conocimiento, desde la captura hasta la aplicación. (James et al., 2025) argumentan que un sistema agéntico sirve no solo como una interfaz de consulta sino también como un habilitador de la arquitectura de información que puede identificar brechas de conocimiento y sugerir actualizaciones de forma proactiva. Al permitir consultas en lenguaje natural sobre conjunto de datos que antes eran inaccesibles, por esto el agente democratiza el acceso a la pericia técnica dentro de la organización, alineándose con los objetivos de eficiencia operativa.

La eficiencia de estos sistemas destaca en la capacidad de integrarse con la infraestructura existente de la organización, es así que un agente bien diseñado no reemplaza el sistema de gestión del conocimiento, sino que actúa como una capa de inteligencia que orquesta la interacción entre los usuarios y los activos digitales, por ello esta integración asegura que la toma de decisiones se apoye en una verdad única extraída de documentos oficiales que proporcionan un marco de trabajo donde la IA generativa no es un generador de contenido creativo, sino es un motor de búsqueda y análisis altamente controlado y confiable.

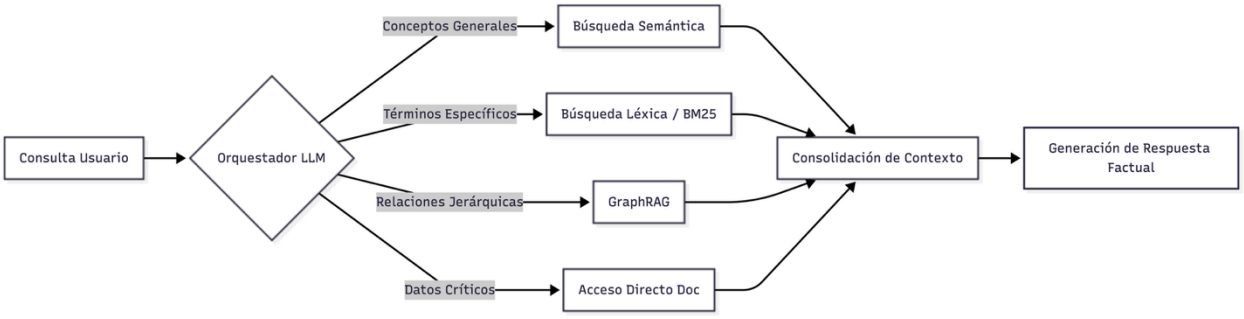


Figura 5. Flujo Routing del RAG agéntico

Los sistemas agénticos basados en modelos de lenguaje representan una evolución significativa en la forma en que los sistemas de inteligencia artificial interactúan con información y ejecutan tareas complejas. A diferencia de los modelos generativos tradicionales, las arquitecturas agénticas incorporan capacidades de memoria, uso de herramientas externas y razonamiento iterativo, lo que

permite abordar procesos más complejos de consulta y análisis de información. No obstante, su implementación introduce nuevos desafíos relacionados con la coordinación de procesos, el control del flujo de información y la trazabilidad de las decisiones generadas por el sistema.

En contextos organizacionales, donde la gestión del conocimiento requiere altos niveles de confiabilidad y transparencia, resulta fundamental complementar estas arquitecturas con mecanismos de recuperación documental, evaluación de respuestas y control del comportamiento del agente. En este sentido, la integración de sistemas agénticos con arquitecturas basadas en Retrieval-Augmented Generation permite aprovechar las capacidades de razonamiento de los modelos LLM, al mismo tiempo que se mantiene una vinculación explícita con fuentes documentales verificables, contribuyendo así a mejorar la confiabilidad y utilidad de los sistemas de consulta de conocimiento organizacional.

7. Evaluación de sistemas RAG y agénticos

La evaluación de los sistemas de inteligencia artificial se considera un requisito crítico para mitigar los riesgos operativos y éticos que se asocian a la generación de lenguaje. Asimismo, según Yu et al. (2024), el despliegue de modelos sin una validación exhaustiva puede derivar en la propagación de sesgos, errores factuales y la pérdida de confianza por parte de los usuarios, por este motivo en el contexto de la gestión documental donde la precisión es imperativa, la evaluación deja de ser una opción de desarrollo para convertirse en un componente fundamental de la gobernanza tecnológica.

La necesidad de métricas que sean objetivas responde a la naturaleza probabilística de los LLM, mientras que en el software tradicional se mide el éxito por resultados que son binarios, en cambio en los sistemas agénticos la calidad se evalúa en un espectro de relevancia y fidelidad de información, es por esto que como señala el marco del NIST AI 600-1 (Autio & others, 2024) que "una métrica robusta ayuda a cuantificar el desempeño del sistema frente a riesgos específicos, lo que proporciona una base empírica para la toma de decisiones sobre la escalabilidad del agente dentro de la arquitectura empresarial".

7.1 Métricas de recuperación y generación

La evaluación de un sistema RAG se divide en dos dimensiones interdependientes, la primera que es la calidad de la recuperación y la segunda que se relaciona con la calidad de la generación, por tanto, para la recuperación se utilizan métricas convencionales de recuperación de información como Precision y Recall los que miden la capacidad del sistema para identificar los fragmentos de conocimiento correctos. Sin embargo, (Yu et al., 2024b) destacan que la similitud semántica que se mide a través de la distancia de coseno entre vectores es hoy una métrica superior para entender si el contexto recuperado es realmente útil para el agente, más allá de la coincidencia exacta de palabras que pueda existir.

En la dimensión de la generación, el enfoque se traslada hacia la coherencia y la veracidad, dado que la Tríada de RAG que se basa en la fidelidad, relevancia de la respuesta y relevancia del contexto permite descomponer el proceso para identificar dónde falla el sistema, de esta manera al fidelidad, en particular, es crítica en este estudio al no contar con RLHF pues esta métrica asegura que la respuesta del modelo sea una derivada lógica y exclusiva de los documentos que han sido recuperados, eliminando la dependencia de la creatividad o el conocimiento paramétrico del modelo.

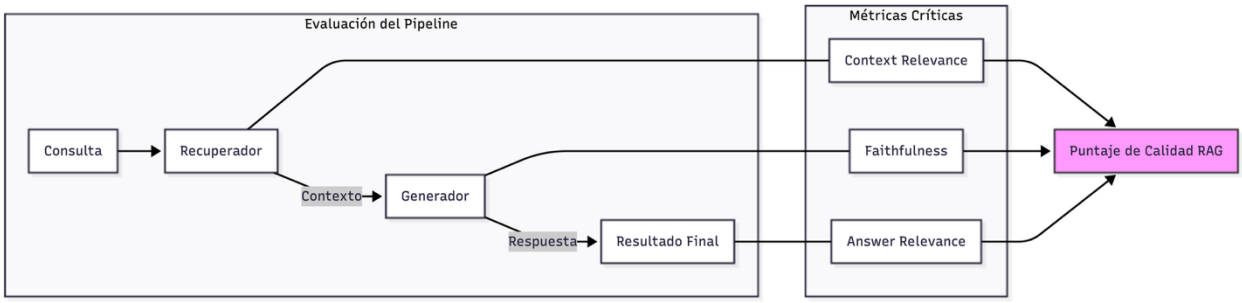


Figura 6. La Tríada de Evaluación RAG: Métricas de interdependencia entre recuperación y generación. Basado en (Yu et al., 2024b).

7.2 Frameworks de evaluación automática: RAGAS, ARES y DeepEval

Dada la limitación de realizar evaluaciones humanas a escala, han surgido entornos de evaluación automatizada, por ello (Saad-Falcon et al., 2024b) presentan marcos como RAGAS (RAG Assessment) y ARES los que utilizan modelos de lenguaje de mayor capacidad para puntuar de manera automática las respuestas generadas de sistemas más pequeños, de esta forma estos frameworks permiten realizar pruebas unitarias sobre el conocimiento organizacional para garantizar que el agente mantenga su desempeño a medida que se añaden nuevos activos de conocimiento a la base de datos.

El uso de herramientas como DeepEval facilita la integración de la evaluación en el ciclo de desarrollo continuo. Según (Saad-Falcon et al., 2024b), el valor de estos frameworks radica en su capacidad para proporcionar métricas interpretables que imitan el juicio humano, de esa forma al automatizar la detección de alucinaciones y la relevancia contextual las organizaciones pueden iterar rápidamente sobre la arquitectura del agente el cual ajusta el tamaño de los fragmentos o los algoritmos de búsqueda con la seguridad de que cada cambio mejora la precisión real del sistema.

7.3 Evaluación de LLM como juez

El paradigma del LLM como juez utiliza modelos de frontera como por ejemplo GPT-4o o Claude 3.5 para evaluar la calidad de las respuestas de modelos de menor escala, es así que esta técnica ofrece ventajas considerables en términos de velocidad y consistencia comparada con la evaluación humana que se usa tradicionalmente. Sin embargo, no se encuentra libre de limitaciones pues los

modelos como juez pueden presentar sesgos hacia sus propias respuestas o hacia textos de mayor longitud, por ello es fundamental calibrar el modelo juez mediante conjuntos de datos de referencia previamente validados por expertos.

Un hallazgo crítico en la literatura actual es la relación que puede existir entre el juicio automático y el humano, por esto diversos estudios, incluyendo el análisis de (Saad-Falcon et al., 2024b), proponen que un sistema alcanza un estado de madurez de despliegue cuando existe una correlación superior al 80% entre el juicio del LLM y el de evaluadores expertos, por este motivo alcanzar este umbral del 80% indica que el sistema de evaluación automática es lo suficientemente robusto para actuar como un sustituto confiable de la supervisión humana, permitiendo así que el ciclo de mejora del agente sea autónomo y escalable bajo un marco de gobernanza técnico estricto.

Tabla 3. Comparativa de frameworks de Evaluación Automática

FRAMEWORK	***ENFOQUE***	***REFERENCIA DE EVALUACIÓN***	***VENTAJA***
RAGAS	Pipeline RAG completo.	Referencia-less (no necesita verdad única).	Especializado en métricas de la tríada RAG.
ARES	Clasificadores aprendidos.	Datos sintéticos y humanos.	Alta correlación con juicio humano.
DEEPEVAL	Unit Testing para LLMs.	Basado en tests de aserción.	Integración fácil en flujos de DevOps.

De esta forma, el marco teórico que se desarrolló a lo largo de este estudio demuestra que la gestión del conocimiento constituye un activo estratégico para las organizaciones porque es capaz de potenciar la innovación, el aprendizaje y la eficiencia operativa que pueda llegar a existir. Por otro lado, los enfoques tradicionales de gestión documental y de búsqueda basada en palabras clave presentan restricciones frente a la heterogeneidad y complejidad de los documentos actuales, por este motivo es fundamental incorporar principios de gestión del conocimiento, estándares internacionales como ISO 30401 y marcos de arquitectura empresarial como TOGAF con soluciones tecnológicas avanzadas, incluyendo sistemas agénticos y agentes RAG, a su vez la revisión de la literatura también destacó la importancia de distinguir entre conocimiento explícito y tácito, estructurado y no estructurado, así como la necesidad de contar con mecanismos con mayor robustez de análisis del diseño de documentos, segmentación visual y estructural, y

evaluación del parsing documental garantizando de este modo la fidelidad y trazabilidad de la información procesada.

Los hallazgos detallados proporcionan un sustento conceptual sólido que fundamenta la fase de diseño e implementación del estudio, de este modo la sección siguiente, metodología de investigación, describe el enfoque experimental y exploratorio que se llevará a cabo, además las técnicas de recolección de datos, la selección de la muestra documental y los criterios de evaluación de desempeño del sistema agéntico y los agentes RAG propuestos, así mismo se detallará cómo se operacionalizan los conceptos de inteligencia artificial generativa, agentes autónomos y recuperación aumentada para así asegurar que la arquitectura diseñada permita la extracción, almacenamiento y consulta semántica de conocimiento con precisión y coherencia, cumpliendo los objetivos planteados para la administración eficiente de información no estructurada en entornos organizacionales.

En este contexto, el presente estudio propone el diseño e implementación de un sistema agéntico basado en RAG que permita centralizar la consulta de conocimiento organizacional garantizando trazabilidad y confiabilidad de la información para mejorar la gestión documental de dominios de conocimiento heterogéneos en formato y contenido.

Metodología

Metodología de investigación

Enfoque metodológico: Design Science Research (DSR)

La presente investigación se desarrolla bajo el enfoque de Design Science Research (DSR), dado que su enfoque es diseñar, implementar y evaluar un artefacto tecnológico orientado a resolver un problema (Hevner & Salvatore, 2004).

En este estudio, el artefacto corresponde a una arquitectura agéntica basada en Retrieval-Augmented Generation (RAG), concebida como un sistema de apoyo a la gestión documental en el contexto de la gestión del conocimiento organizacional. En consecuencia, el aporte principal de la investigación se centra en la construcción de una solución tecnológica evaluable, cuya utilidad se examina a través de métricas asociadas a las etapas de extracción, recuperación y generación de información.

Fases de la investigación

Conforme al enfoque de Design Science Research adoptado, el desarrollo metodológico del estudio se estructura en cuatro fases principales: identificación del problema, diseño del artefacto, implementación y evaluación.

En la primera fase, se realiza la identificación del problema de investigación a partir de la revisión de literatura sobre gestión documental, conocimiento no estructurado y sistemas de recuperación de información, lo que permite establecer las limitaciones.

En la segunda fase, se lleva a cabo el diseño del artefacto tecnológico propuesto, consistente en una arquitectura agéntica basada en Retrieval-Augmented Generation (RAG). Esta etapa incluye la definición de dominios de conocimiento emulando una estructura empresarial real, selección de la muestra mediante recopilación de documentos públicos heterogéneos, definición funcional de la arquitectura del sistema, tales como el pipeline de ingesta y procesamiento documental, el almacenamiento vectorial y el patrón agéntico de consulta con enrutamiento de conocimiento.

En la tercera fase, se procede a la implementación del artefacto diseñado mediante la integración de herramientas de procesamiento documental, bases de datos vectoriales y modelos fundacionales LLM y Embeddings, configurando un entorno experimental que permite operar el sistema sobre el conjunto de no estructurados seleccionados.

Finalmente, en la cuarta fase, se realiza la evaluación técnica del artefacto mediante un conjunto de preguntas de prueba y métricas estándar para sistemas agénticos con RAG asociadas a las etapas de extracción, recuperación y generación de respuestas. Esta evaluación tiene como propósito

determinar el nivel de desempeño alcanzado por la arquitectura propuesta en términos de precisión de recuperación, similitud semántica, factualidad y coherencia de las respuestas generadas.

Fase 1: Identificación del problema

La primera fase del proceso de investigación se orienta a delimitar el problema relacionado con la gestión y consulta de información contenida en documentos organizacionales. El conocimiento institucional de las organizaciones reside mayoritariamente en repositorios documentales heterogéneos que presentan estructuras internas diversas, calidad variable y diferentes grados de digitalización. Esta heterogeneidad genera fragmentación del conocimiento, duplicidad, obsolescencia y pérdida progresiva de conocimiento experto, con consecuencias directas en la toma de decisiones y en la eficiencia de los procesos internos. Los sistemas tradicionales de búsqueda no resuelven este problema, sino que lo agravan: incapaces de capturar el significado semántico ni el contexto del contenido, producen resultados poco relevantes que los usuarios terminan por abandonar e inclusive hay información que no se llega a acceder. Este escenario configura el punto de partida de la investigación: la necesidad de una arquitectura tecnológica que transforme los documentos organizacionales en activos de conocimiento accesibles, trazables y aprovechables para la gestión institucional.

La complejidad del problema se amplifica al considerar la diversidad estructural de los documentos organizacionales, analizada en la sección 3. Gestión documental y conocimiento no estructurado.. Documentos PDF textuales, archivos escaneados, presentaciones PPTX y PDF con tablas complejas coexisten en los repositorios corporativos, cada uno con características estructurales propias que los sistemas de extracción genéricos no logran preservar adecuadamente. Como desarrolla la sección 3.3, la extracción de texto sin análisis de diseño documental (Document Layout Analysis) resulta insuficiente: encabezados, tablas, jerarquías y la disposición espacial del contenido son una estructura que preserva el significado semántico.

La sección 4. Inteligencia Artificial Generativa en contextos organizacionales analiza y discute la comparativa entre búsqueda con motores tradicionales basados en palabras clave y sistemas basados en IA generativa, donde se evidencian con claridad las ventajas del procesamiento semántico y las limitaciones de la búsqueda por coincidencia exacta de palabras. Los motores convencionales operan bajo una lógica de recuperación estadística incapaz de capturar la intención del usuario ni las relaciones semánticas latentes entre distintos documentos, fragmentando el conocimiento organizacional y limitando el soporte a los procesos de decisión. No obstante, la solución no reside en adoptar modelos generativos puros. La sección 4.2 documenta que los LLM producen el fenómeno de alucinación: respuestas lingüísticamente coherentes pero factualmente incorrectas, cuyas consecuencias en gestión documental pueden invalidar procesos legales, financieros u operativos. La sección 4.3 refuerza esta limitación al señalar que los LLM operan en silos sin acceso al contexto histórico de la organización, por lo que su adopción corporativa exige arquitecturas controladas que garanticen precisión técnica, trazabilidad y alineación con la base de conocimiento institucional.

La arquitectura Retrieval-Augmented Generation, analizada en la sección 5. Generación Aumentada por Recuperación (RAG), es la respuesta técnica propuesta a las limitaciones identificadas. RAG contextualiza cada respuesta generada en fragmentos documentales específicos recuperados de un corpus externo, mitigando la caja negra de los LLM y transformando el RAG en un sistema de consulta auditable y verificable. La sección 5.1 documenta que esta arquitectura reduce significativamente el riesgo de alucinación, proporciona trazabilidad hacia las fuentes documentales y permite actualizar la base de conocimiento sin necesidad de reentrenamiento del modelo. La sección 5.2 subraya, sin embargo, que su efectividad depende críticamente de la calidad de los procesos de ingesta, segmentación e indexación: problemas en la estructuración del contenido o en las estrategias de recuperación afectan la relevancia del contexto recuperado y, por tanto, la precisión de las respuestas.

Sin embargo, los sistemas RAG estáticos presentan limitaciones propias al enfrentarse a consultas complejas sobre repositorios documentales diversos. La sección 6. Sistemas agénticos basados en IA muestra que las arquitecturas agénticas superan estas restricciones al incorporar un ciclo continuo de percepción, razonamiento y acción, donde el LLM actúa como orquestador cognitivo. La sección 6.1 describe cómo el patrón de enrutamiento dinámico (routing) permite al agente evaluar la naturaleza de cada consulta para seleccionar la estrategia de recuperación más adecuada en función de las bases de conocimiento disponibles. Finalmente, la sección 7. Evaluación de sistemas RAG y agénticos establece que la naturaleza probabilística de los LLM convierte la evaluación sistemática en un requisito de gobernanza y no en una opción de desarrollo. Métricas como precisión de recuperación, fidelidad (faithfulness), relevancia de la respuesta y relevancia del contexto conocida como la "Tríada de RAG". Este conjunto de limitaciones identificadas y soluciones documentadas en la literatura abre la discusión del diseño e implementación de un sistema agéntico basado en RAG como artefacto de investigación.

Fase 2: Diseño del artefacto

La segunda fase del proceso metodológico corresponde al diseño del artefacto tecnológico propuesto. En esta etapa se define la arquitectura conceptual del sistema agéntico basado en RAG, estableciendo los componentes funcionales necesarios para el procesamiento, almacenamiento y consulta de información documental en entornos organizacionales. La especificación y detalle técnico se encuentran en el capítulo de Desarrollo e implementación del artefacto.

El proceso inicia con la ingesta del corpus documental definido en la investigación. Los documentos son procesados mediante herramientas de procesamiento documental que permiten extraer el contenido textual, identificar la estructura del documento y preservar metadatos asociados a secciones, títulos y elementos estructurales. Este proceso incorpora técnicas de reconocimiento óptico de caracteres (OCR) y análisis de estructura, con el propósito de transformar documentos heterogéneos en representaciones estructuradas que puedan ser interpretadas por el sistema.

Posteriormente, el contenido textual extraído es segmentado en fragmentos semánticamente coherentes mediante técnicas de particionamiento basadas en tokenización. Estos fragmentos constituyen las unidades mínimas de conocimiento que serán utilizadas para el proceso de recuperación de información. Cada fragmento es transformado en una representación vectorial mediante modelos de embeddings, lo que permite capturar relaciones semánticas entre los diferentes contenidos documentales.

Las representaciones vectoriales generadas son almacenadas en una base de datos vectorial que permite realizar búsquedas semánticas eficientes dentro del repositorio documental. Este almacenamiento incorpora metadatos asociados a cada fragmento, facilitando la trazabilidad de la información recuperada y permitiendo realizar filtrados contextuales durante el proceso de consulta.

Finalmente, se implementa el agente de consulta basado en RAG, encargado de gestionar la interacción con el usuario. El agente recibe consultas formuladas en lenguaje natural, ejecuta procesos de recuperación semántica sobre la base vectorial y utiliza los fragmentos recuperados como contexto para la generación de respuestas mediante un modelo de lenguaje de gran escala. Este mecanismo permite que las respuestas generadas estén fundamentadas en información documental previamente procesada por el sistema.

Como resultado de esta fase se obtiene una línea base conceptual, de la arquitectura del sistema a agéntico funcional capaz de procesar documentos organizacionales, estructurar su contenido en representaciones semánticas (base de conocimiento) y responder consultas en lenguaje natural basadas en conocimiento documental.

Fase 3: Implementación del sistema

La tercera fase del proceso metodológico corresponde a la implementación del sistema agéntico diseñado en la etapa anterior. En esta fase se materializa la arquitectura conceptual mediante la integración de herramientas, modelos y componentes tecnológicos que permiten procesar, representar y consultar información documental en lenguaje natural como se muestra en el desarrollo de la arquitectura, detallada en la sección Arquitectura del Sistema.

El diseño del sistema se estructura como una arquitectura modular compuesta por cuatro capas principales: i) ingesta y procesamiento documental, ii) particionamiento del contenido, iii) vectorización y almacenamiento iv) implementación del agente de consulta RAG, como se ve en la Figura 7.

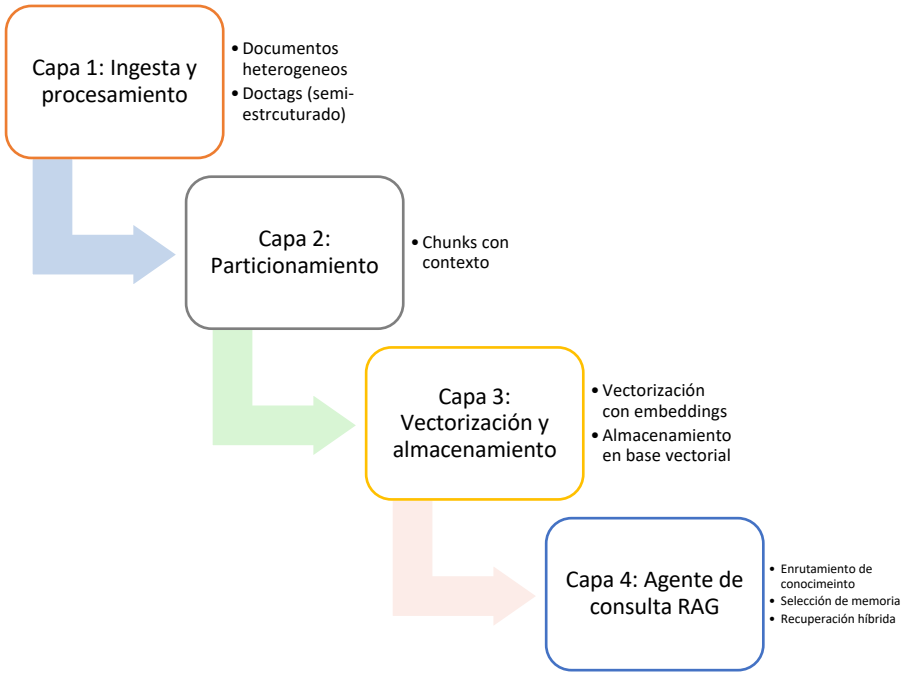


Figura 8. Capas de diseño del artefacto propuesto, Arquitectura en contexto C1.

En la primera capa se define el pipeline de ingesta y procesamiento de documentos, cuyo propósito es transformar documentos heterogéneos en representaciones estructuradas procesables por el sistema. Para ello se emplea la herramienta de procesamiento documental Docling, que permite realizar extracción de texto y análisis del diseño documental mediante modelos especializados. Este pipeline incorpora modelos de Document Layout Analysis, reconocimiento óptico de caracteres (OCR) mediante RapidOCR y modelos multimodales para la interpretación de contenido visual presente en los documentos. Como resultado de este proceso, los documentos son transformados en una representación semi-estructurada denominada Doctags, que conserva tanto el texto como metadatos asociados a la estructura del documento.

Posteriormente, en la segunda capa se realiza el particionamiento del contenido textual, con el objetivo de generar unidades de información adecuadas para su recuperación semántica. En esta etapa los documentos transformados a Doctags son segmentados en fragmentos de texto o chunks. Para ello se utiliza el tokenizador del modelo XLM-RoBERTa, estableciendo un tamaño máximo de aproximadamente 400 tokens por fragmento y un mínimo de 100 tokens. Este proceso permite preservar la coherencia semántica de cada fragmento y facilitar su posterior representación vectorial. Adicionalmente, se incorpora información contextual derivada de la estructura jerárquica del documento, permitiendo asociar cada fragmento con secciones, capítulos o subsecciones del documento original.

En la tercera capa se define el mecanismo de vectorización y almacenamiento del conocimiento, donde cada fragmento textual es transformado en un vector semántico mediante el modelo de embeddings Qwen3-Embedding-4B, que genera representaciones vectoriales de 2560 dimensiones. Estas representaciones son almacenadas en una base de datos vectorial implementada sobre PostgreSQL con la extensión pgvector, lo que permite realizar operaciones eficientes de búsqueda semántica mediante índices de tipo HNSW. La estructura de la base de datos contempla una tabla vectorial por cada dominio de conocimiento, incorporando metadatos que permiten realizar filtrado por fuente documental, etiquetas y contexto estructural del documento.

Finalmente, en la cuarta capa se diseña el agente de consulta basado en RAG, encargado de orquestar el proceso de recuperación y generación de respuestas. Este agente se implementa utilizando los frameworks LangChain y LangGraph, que permiten estructurar el flujo de decisiones del sistema mediante un modelo de agentes con herramientas especializadas. El agente recibe consultas en lenguaje natural, identifica el dominio de conocimiento correspondiente y ejecuta un proceso de recuperación híbrida que combina búsqueda semántica en la base vectorial con filtrado contextual. Los fragmentos recuperados son posteriormente utilizados como contexto para la generación de respuestas mediante un modelo de lenguaje de gran escala, garantizando que las respuestas generadas se fundamenten en evidencia documental.

El diseño arquitectónico resultante permite integrar procesos avanzados de procesamiento documental, recuperación semántica y léxica; y generación de lenguaje natural dentro de una arquitectura modular y escalable orientada a la consulta de conocimiento organizacional.

Fase 4: Evaluación del sistema

La cuarta fase corresponde a la evaluación de la arquitectura implementada como un sistema agéntico, con el propósito de analizar su desempeño en los procesos de recuperación y generación de respuestas basadas en información documental referenciada. Esta evaluación se realiza utilizando el corpus documental definido previamente y un conjunto de consultas en lenguaje natural diseñadas para simular escenarios de búsqueda de conocimiento organizacional.

Durante este proceso, el agente recibe consultas formuladas en lenguaje natural y ejecuta el mecanismo de RAG, el cual combina la recuperación semántica de fragmentos documentales con la generación de respuestas mediante un modelo de lenguaje. Las respuestas generadas por el sistema se comparan posteriormente con respuestas de referencia definidas dentro del conjunto de datos sintético, permitiendo evaluar el desempeño del sistema en términos de recuperación de información y generación de conocimiento.

La evaluación del sistema se realiza mediante un conjunto de métricas que permiten analizar distintas dimensiones del funcionamiento del sistema:

- Precision@k: mide la proporción de fragmentos relevantes dentro de los k documentos recuperados por el sistema durante el proceso de búsqueda semántica. Esta métrica permite evaluar la capacidad del sistema para identificar correctamente los fragmentos documentales más pertinentes para responder una consulta.
- Similaridad semántica (Cosine Similarity): mide el grado de similitud semántica entre la respuesta generada por el sistema y la respuesta de referencia definida en el conjunto de evaluación. Esta métrica se calcula utilizando representaciones vectoriales generadas mediante modelos de embeddings.
- Factualidad o fidelidad de la respuesta (Faithfulness): evalúa si la respuesta generada por el sistema se encuentra respaldada por los fragmentos documentales recuperados durante el proceso de búsqueda, permitiendo identificar posibles casos de generación de información no sustentada por el contexto documental.
- Coherencia de la respuesta: analiza la consistencia y claridad de la respuesta generada por el sistema en relación con la consulta realizada, evaluando si la respuesta mantiene una estructura lógica y una relación adecuada con la información recuperada.

Adicionalmente, el proceso de evaluación puede apoyarse en frameworks especializados para la evaluación de sistemas RAG, como RAGAS y ARES, los cuales permiten otras dimensiones además de las propuestas en este estudio y diferentes dimensiones de calidad en las respuestas generadas mediante técnicas automatizadas de evaluación LLM-as-a-Judge.

Finalmente, el análisis de los resultados obtenidos permite identificar fortalezas y limitaciones del sistema agéntico en distintos escenarios documentales, aportando evidencia sobre la efectividad de la arquitectura basada en RAG agéntico para facilitar la recuperación y utilización de conocimiento contenido en repositorios documentales organizacionales.

Desarrollo e implementación del artefacto

Configuración Experimental

Para la implementación de este sistema se debe realizar una configuración especializada de hardware y software, debido a la naturaleza de los módulos y modelos fundacionales empleados, detallados en la siguiente tabla.

Tabla 4. Descripción a detalle de la configuración experimental donde se desplegó el sistema.

Componente	***Descripción***	***Uso***
Procesador	Intel Core i7 11700K	Ejecución de programas y soporte de cargas distribuidas.
RAM	64 GB DDR4 @ 3200 MHz	Ejecución intensiva de consultas y procesamiento de ventanas de contexto de LLMs.
GPU	NVIDIA GeForce RTX 3060 12GB VRAM, CUDA 12.9	Ejecución de modelos fundacionales locales con llama.cpp
Sistema Operativo	Ubuntu 24.04.3 LTS	Ejecución de contendores, intérpretes de código y motores de base de datos.
Motor de contenedores	Docker 29.2.0	Despliegue de contendores de base de datos e inferencia de modelo fundacionales.
Lenguajes de programación	Python 3.13.11 PostgreSQL 17.8	Construcción del sistema.

Para desplegar el sistema en la nube se deberá considerar que las instancias de cómputo tengan como mínimo características similares para operar de la forma óptima.

Corpus documental sintético

Técnica Principal: Simulación y Análisis Documental

Se recolectará una muestra sintética de documentos que imite la gestión documental real de una organización, para de esta manera evitar problemas de privacidad y confidencialidad vinculado al uso de datos reales de empresas. Con ellos se obtiene el control total sobre las variables de prueba (complejidad y formato).

Instrumentos: Se aprovecharán frameworks de web scrapping para la recolección de datos que además, permitirá la evaluación automática de la respuesta generada por el agente frente a la respuesta del set de datos sintéticos (Saad-Falcon et al., 2024).

Por otra parte, se aplica la técnica del análisis de diseño de documentos, ya que permite descomponer visual y estructuralmente los archivos, para poder evaluar la eficacia de los algoritmos de extracción, implementados en el sistema agéntico.

Población y Muestra

Se ha definido un muestreo estratificado no probabilístico, puesto que, la estratificación se alinea con las capacidades organizacionales definidas por el estándar TOGAF y los principios de gestión del conocimiento de la norma ISO 30401 (International Organization for Standardization, 2018b; The Open Group, 2022b). Esta estructura garantiza que el sistema agéntico sea evaluado en diversos contextos semánticos.

Diseño de la Muestra: El tamaño total de la muestra debe ser de aproximadamente 120 documentos, distribuidos equitativamente en 6 dominios funcionales (estratos), con un total de 20 documentos por dominio.

Estratos (Dominios Funcionales):

1. Clientes y mercado: contiene información referente al protocolo de servicio al cliente, benchmark de negocios, términos y condiciones de los contrato y políticas de tratamiento de clientes.

2. Productos y servicios: contiene información sobre catálogos de productos/servicios, detalles de los productos/servicios, preguntas frecuentes y demás detalles.

3. Operaciones, cadena de suministro y procesos: contiene información sobre normativas industriales, políticas de procedimientos, flujos industriales y empresariales.

4. Finanzas, seguros y riesgos: contiene información sobre resultados financieros, cierres contables, auditorias financieras y de riesgo.

5. Legal, cumplimiento normativo: contiene información de políticas legales y normativas a cumplir en las industrias

6. Talento humano y desarrollo organizacional.: contiene información sobre políticas de empleados, capacitaciones, gestión del talento humano.

Distribución por Tipología (Variable de Control): Dentro de cada dominio, los documentos respetan la siguiente distribución porcentual para asegurar la representatividad técnica:

Tabla 5. Distribución de la muestra documental por dominio de conocimiento

Porcentaje	***Tipo de Documento***	***Características / Complejidad***
60%	PDF / DOCX Textuales	Baja complejidad; contenido principalmente de texto.
15%	Presentaciones PPTX	Estructura jerárquica y elementos visuales.
15%	Escaneados / Imágenes OCR	Presencia de ruido visual; requiere reconocimiento de caracteres.
10%	PDF con Tablas Complejas	Alta densidad de datos; requiere extracción estructurada.

Esta composición respeta la distribución real de reportes promedio de repositorios documentales corporativos reales, y asegura que el sistema sea probado mayoritariamente en texto estándar, pero con diferentes tipologías (tablas e imágenes) que justifican el uso de modelos multimodales.

Arquitectura del Sistema

Este sistema se basa en la arquitectura de Retrieval Augemented Generation (RAG) agéntico, discutida anteriormente en la sección Marco teórico. En este sistema se propone que existan por lo menos tres roles clave para la ejecución y operación.

1. Dueño de dominio de conocimiento: Experto de procesos o negocio que garantiza la factualidad y coherencia del contenido en los documentos del dominio de conocimiento. Su rol es el más relevante, pues aquí nace la información con la que el agente genera las respuestas. Para cumplir de la mejor forma con su rol aplica las recomendaciones de la

norma ISO 30401 y es responsable por el contenido en los documentos, así como que estén actualizados y con los metadatos correspondientes.

2. Administrador de base de conocimiento: Experto técnico en ingeniería de datos-conocimiento, vela por que las bases de conocimiento cumplan con los estándares técnicos definidos en la arquitectura del sistema, opera los flujos de ingesta, vectorización y monitorea la integridad de la base de conocimiento (vectorial, grafos, etc).
3. Usuario: Es el consumidor de las fuentes de conocimiento y su rol va más allá de realizar preguntas y esperar respuestas. Este usuario debe ser embajador del uso responsable de la Inteligencia Artificial (RAI por sus siglas en inglés), esto implica revisión de las fuentes entregadas y análisis de impacto para evitar la conducción al error. Así como de la retroalimentación de las respuestas del agente para conseguir un ciclo de mejora continua.

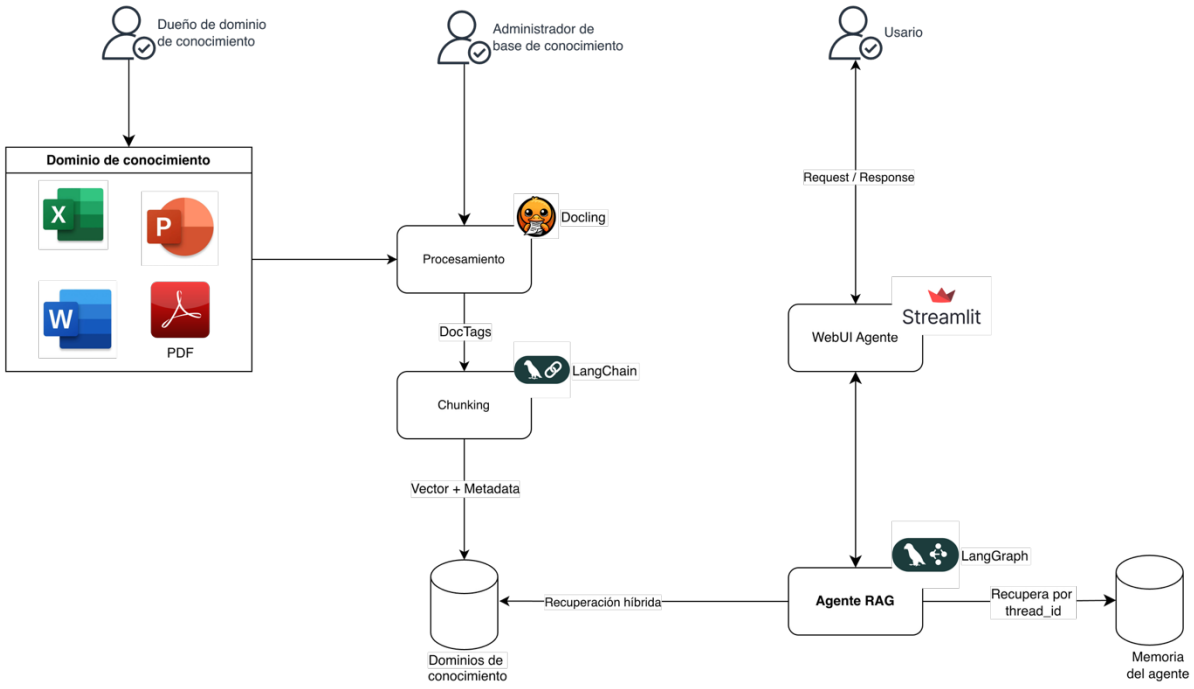


Figura 9. Arquitectura C2, sistema de RAG Agéntico

A continuación, se detalla cada una de las etapas y componentes implementados en la arquitectura.

1. Ingesta y procesamiento

Como se ha detallado anteriormente esta arquitectura cuenta con múltiples formatos de documentos y múltiples bases de conocimiento por lo que se vuelve imprescindible usar un flujo de procesamiento de datos no estructurados robusto, que soporte diferentes fuentes. Tras realizar una revisión de literatura se ha determinado que la herramienta open source software Docling es la mejor opción para un sistema de RAG agéntico (Livathinos et al., 2025).

Esta herramienta cuenta con un conjunto amplio de métodos y modelos especializados en: extracción, interpretación y análisis de documentos en diferentes formatos, haciendo uso de modelos fundacionales de IA Generativa. Docling por defecto emplea modelos desarrollados y liberados por IBM. En este sistema se ha buscado hacer un uso avanzado de la herramienta mediante usar configurar modelos LLM y VLM externos especializados en el análisis de documentos e imágenes (Docling, 2026).

El proceso de extracción de texto de los documentos parte con la configuración del pipeline de Docling, donde se recibe un directorio como entrada, y dentro se encuentran todos los documentos en los distintos formatos.

El proceso consta de las siguientes etapas:

1. Extracción de texto y etiquetas de document layout, mediante el modelo granite-docling-258M
2. Extracción de texto con un motor de Reconocimiento Óptico de Caracteres (OCR) en caso de ser necesario. El modelo es provisto en el framework Rapid OCR.
3. Anotación y transcripción de texto de imágenes con VLM – qwen3-vl-8B.
4. Generación de documento semi-estructurado Doctags (estructura especializada compuesta por el fragmento de texto, pagina, sección, identificadores, otros)

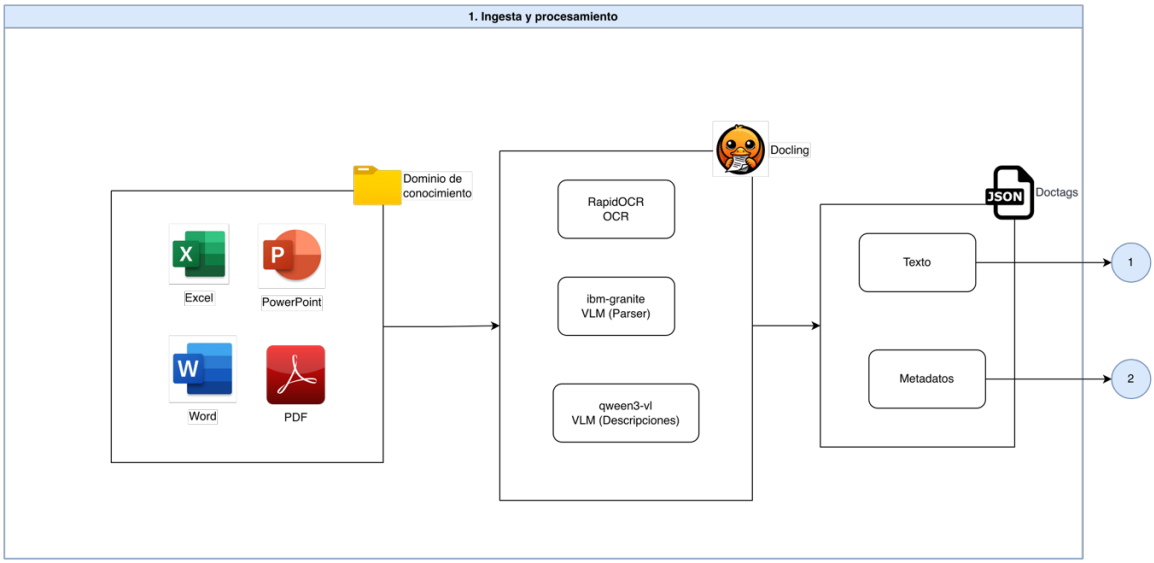


Figura 10. Arquitectura C3, fase ingesta y procesamiento

2. Particionamiento de texto

Cuando el texto ha sido transformado los documentos a Doctags se tiene: tanto texto como metadatos, listos para ser procesados en fragmentos de texto o "chunks". Para este propósito se emplea un modelo de FacebookAI/xlm-roberta-base que permite realizar la tokenización de los textos y establecer el tamaño máximo del texto tomado como 400 tokens máximo y 100 como mínimo.

Adicional, mediante el método de Docling contextualize, se consigue tener un enriquecimiento contextual en que se adiciona la estructura jerárquica del documento.

Por ejemplo: **Capítulo 3 > Sección 3.1 > Subsección 3.1.2 [texto del chunk aquí].**

3. Vectorizado y almacenamiento

Los chunks son generados para cada documento, los cuales se agregan en una lista única pero discriminados por dominio de conocimiento.

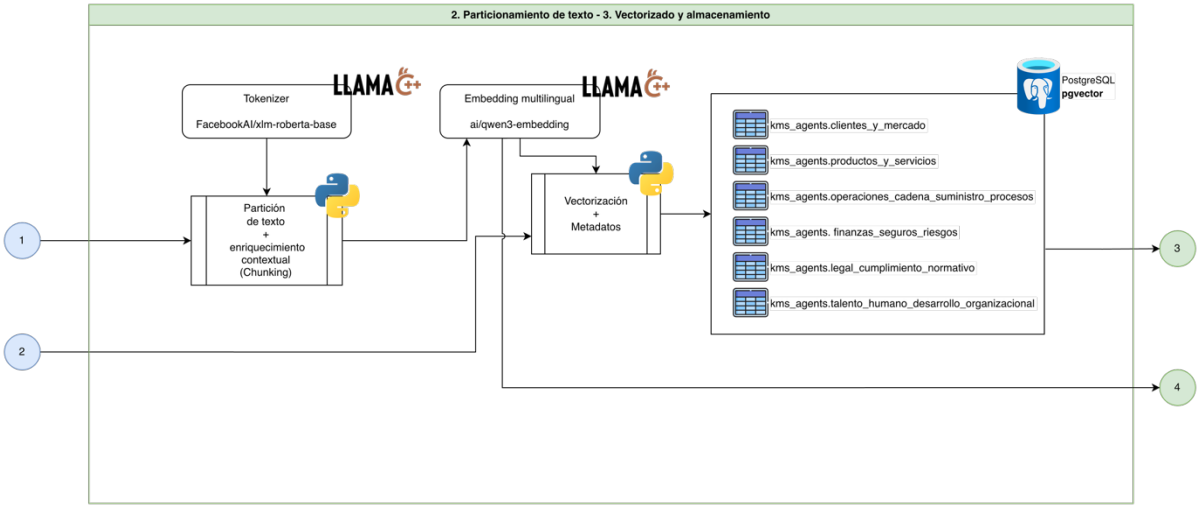


Figura 11. Arquitectura C3, Particionamiento de texto, vectorización y almacenamiento.

La base de datos vectorial es desplegada en una base de datos PostgreSQL versión 17 con la extensión pgvector versión 0.8.0. La imagen base se encuentra en Docker Hub como pgvector/pgvector:pg17. Esta configuración permite realizar las operaciones de recuperación semántica y léxica requeridas para implementar el RAG. El diseño de la base de datos vectorial tiene un esquema donde se tiene una tabla/índice vectorial por cada uno de los seis dominios de conocimiento. Cada tabla de la base de conocimiento está estructurada conforme el esquema detallado en la imagen a continuación.

Figura 12. Estructura de tabla de la base vectorial

Los atributos de cada columna de la base vectorial se detallan a continuación:

Tabla 6. Descripción de los atributos de la tabla vectorial por cada dominio de conocimiento.

Columna	***Tipo PostgreSQL***	***Descripción***
id	varchar (PK)	UUID v4 en texto, generado por Python
embedding	halfvec(2560)	Vector de 2560 dimensiones en media precisión (16-bit float). Propio de la extensión pgvector
document	character varying	Texto completo del chunk (con contexto enriquecido)
cmetadata	jsonb	JSON binario con metadatos: source, mimetype, page, all_pages, labels, annotations, context, chunk_id
custom_id	character varying	ID legible para identificar el archivo del que proviene: {archivo}_{chunk_id}

Para hacer una recuperación semántica y léxica adecuada es indispensable tener índices en la base de datos, estos se detallan a continuación:

Tabla 7. Índices de la base de conocimiento

Índice	***Tipo***	***Sobre***	***Criterio***
HNSW principal	hnsw (halfvec_l2_ops)	embedding	Búsqueda ANN (similitud vectorial). m=32, ef=128
GIN metadata	gin	cmetadata	Filtros por metadatos (ej. source, labels)
GIN FTS	gin (to_tsvector(...))	cmetadata->>'context' || source	Búsqueda full-text en español sobre headings y nombre de archivo

El flujo desarrollado se ejecuta de la siguiente forma:

1. Se genera una lista de chunks como un objeto Langchain document por dominio de conocimiento.
2. Se realiza la vectorización de la lista de chunks con el modelo qwen3-embedding 4B de 2560 dimensiones, la inferencia de este modelo se realiza con hardware local en GPU.
3. Se forma un .json con la estructura de la tabla y se aplica una operación de tipo INSERT en la base de datos.

Esta base vectorial sirve como la fuente del conocimiento para que el agente pueda recuperar mediante sus herramientas de enrutamiento y recuperación de contexto.

4. Implementación del agente RAG

De acuerdo con la sección 6. Sistemas agénticos basados en IA, por definición una agente de IA debe tener un modelo fundacional LLM o VLM, tools, y memoria, para poder cumplir con su tarea.

Para el desarrollo de este agente se ha empleado el framework langchain versión 1.2.7 y langgraph versión 1.0.7 con el uso de la Functional API de langgraph.

Este agente busca la tener autonomía y la escalabilidad mediante modularidad, por lo tanto, se implementaron los siguientes componentes:

1. Agente enrutador de conocimiento: Analiza la pregunta del usuario y determina cual es la base de conocimiento más adecuada para recuperar contexto.
2. Tool base de conocimiento: Utiliza el conector pyscopg para ejecutar operación de consulta semántica y léxica a la base de conocimiento.

Para propósitos de una recuperación de contexto efectiva se implementa el algoritmo de Fusión RAG mediante la técnica Reciprocal Rank Fusión (RRF) propuesta por (Cormack et al., 2009). Descrito por la siguiente formula:

$$RRF_{Score} = \frac{Peso_{semántico}}{k_{rrf}+Rank_{semántico}} + \frac{Peso_{léxico}}{k_{rrf}+Rank_{léxico}}$$

Donde:

- $Peso_{semántico}$ y $Peso_{semántico}$: Corresponden a los coeficientes o pesos asignados a cada tipo de recuperación (vectorial y por palabras clave
- k_{rrf}Es una constante de suavizado que evita que los documentos con rangos muy bajos (posiciones iniciales) tengan un peso desproporcionadamente alto en la suma.
- $Rank_{semántico}$: Es la posición que ocupa el documento en los resultados obtenidos mediante la búsqueda por similitud semántica.
- $Rank_{léxico}$: Es la posición que ocupa el documento en los resultados de la búsqueda tradicional basada en coincidencia de términos.

Para esta implementación se aplicó la configuración de Pesos: Se asigna un 0.75 al componente semántico y 0.25 al léxico para priorizar el significado sobre la coincidencia exacta.

La constante de Suavizado (k_{rrf}) es establecida en 60 para evitar que los primeros resultados dominen excesivamente la puntuación, por recomendación experimental de los autores de la técnica.

En caso de no existir un contexto a recuperar la herramienta tiene un manejador de resultados

3. Agente de respuesta: Genera la respuesta referenciada con base en la pregunta, el contexto y el histórico (si el agente decide que es relevante). El histórico depende de los puntos de control (checkpoints); se guardan y se recuperan mediante la clave el thread_id.
4. Puntos de control (checkpointers) de persistencia: Se está empleando el módulo de persistencia como recomienda el framework de langgraph. A continuación, se puede ver el diagrama entidad relación propuesto. Este modelo de base de datos se implementa en el motor PostgreSQL en un esquema distinto al de la base de conocimiento.

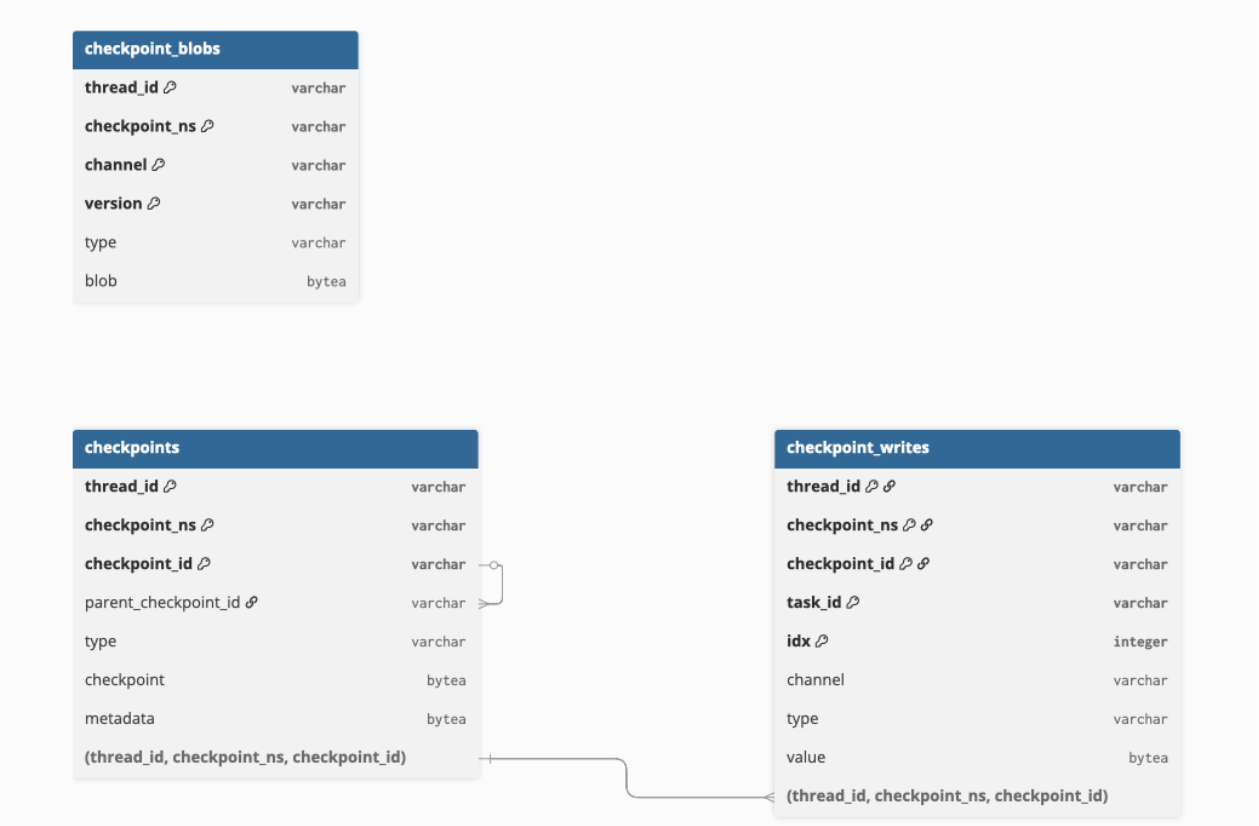


Figura 13. Diagrama entidad relación de checkpoints del agente

A continuación, se muestra el algoritmo implementado para el agente:

Figura 14. Algoritmo de agente RAG implementado

Entrada: Consulta del usuario q, Historial de conversación H

Salida: Respuesta final R, Estado actualizado S

1. **Fase 1: Agente enrutador de conocimiento**
    - Recibir consulta a través del controlador principal.
    - **Identificación de Dominio:** Ejecutar modelo de lenguaje (LLM) para clasificar q.
    - **Si** $q \in Dominio\ Conocido$
        - $Contexto\ (C) \leftarrow$ Recuperar información mediante técinica híbrida con la herramienta $Base\ de\ Conocimiento$.
    - **Si no**:
        - $C \leftarrow$ Ejecutar *Fallback Handler* para manejo de excepciones.
    - Fin de la condición.
2. **Fase 2: Agente de respuesta**
    - **Decisión de Memoria:** Evaluar necesidad de contexto histórico.
    - **Si** requiere memoria:
        - $C_{total} \leftarrow$ Combinar contexto C con historial H.
    - **Si no**:
        - $C_{total} \leftarrow C$
    - Fin de la condición.
    - **Generación:** $R \leftarrow$ Generar respuesta utilizando C_{total}.
3. **Fase de Persistencia**
    - S Calcular nuevo estado del sistema.
    - Almacenar S y R en base de datos PostgreSQL como punto de control (*checkpoint*).
    - **Retornar** R al usuario.

La implementación del agente a nivel de tecnología se detalla en la siguiente figura:

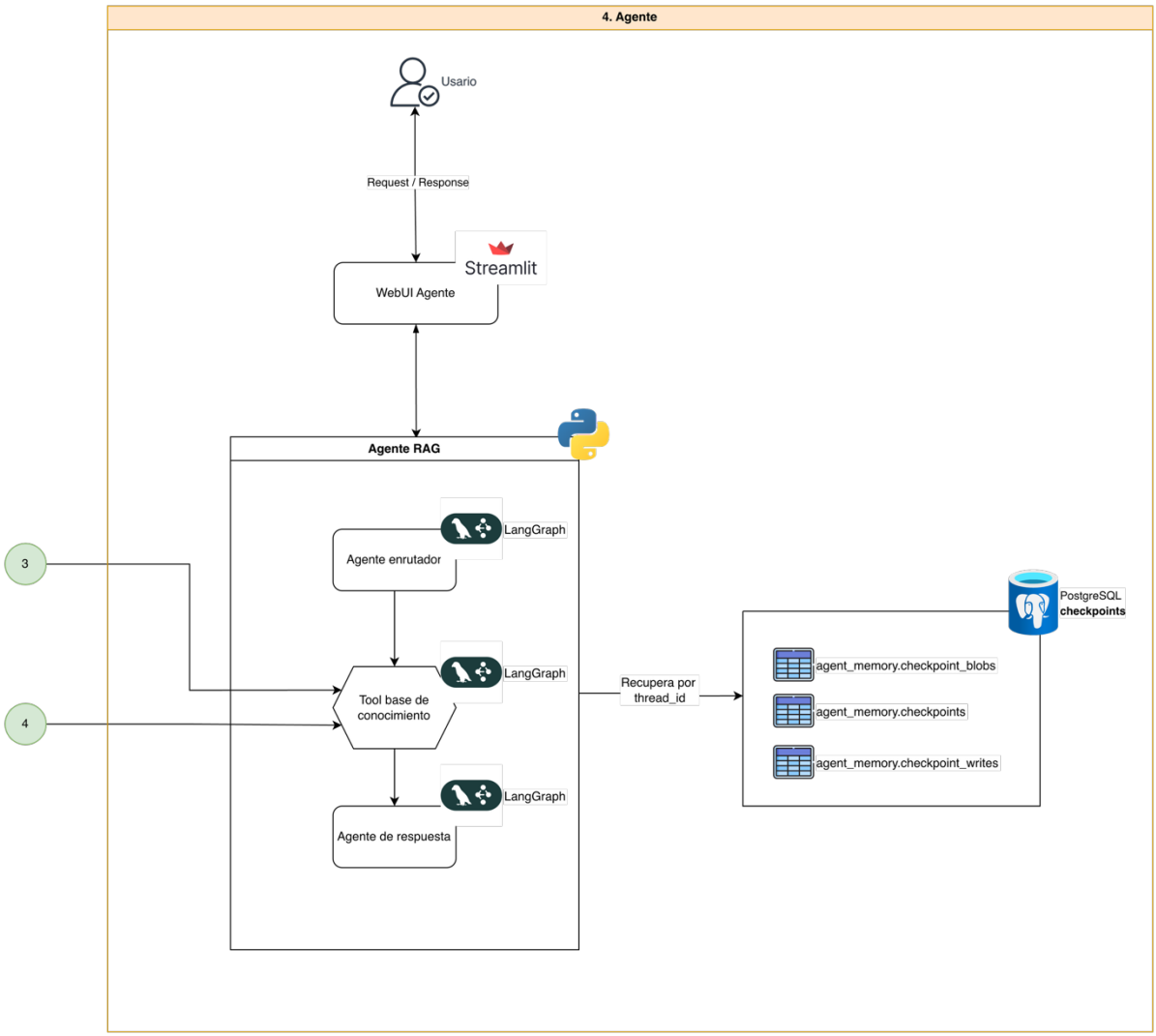


Figura 15. Arquitectura C3, agente RAG

Modelos Fundacionales del sistema

Los modelos fundacionales de inteligencia artificial generativa son el núcleo de este sistema, para el proceso de selección de dichos modelos se ha tomado como referencia los model card y benchamarks globales de finales de 2025 (Hugging Face Community, 2025; *Vision AI Leaderboard - Best Image & Multimodal Models*, 2025). A continuación, se presenta el resumen de los modelos fundacionales y el propósito para el que emplea cada uno de ellos.

Tabla 8. Listado de modelos fundacionales empleados en las etapas de la arquitectura del sistema.

Modelo Funcional	***Parte de la arquitectura***	***Propósito***	***Ranking finales 2025***	***Enlace Model Card***
knowledgator/gliner-bi-large-v1.0	Análisis NLP	NER (reconocimiento de entidades) flexible/zero-shot mediante GLiNER (bi-encoder).	El model card describe el enfoque y sección de benchmarks; no reporta un ranking único 2025.	https://huggingface.co/knowledgator/gliner-bi-large-v1.0
spaCy es_core_news_lg	Análisis NLP	Etiquetado POS (y pipeline lingüístico: parser, lemmatizer, NER, etc.) para español.	Métricas reportadas: POS_ACC 98.51; DEP_LAS 88.19; ENTS_F 89.72 (entre otras).	https://spacy.io/models/es/#es_core_news_lg
josmunpen/mt5-small-spanish-summarization	Análisis NLP	Resumen/generación de titulares en español (mT5-small fine-tuned).	ROUGE-1 44.03; ROUGE-2 28.29; ROUGE-L 40.54; ROUGE-Lsum 40.56 (según model card).	https://huggingface.co/josmunpen/mt5-small-spanish-summarization
FacebookAI/xlm-roberta-base	Particionamiento de texto	Tokenizador/backbone multilingüe (XLM-RoBERTa) para tareas NLP; preentrenamiento MLM.	NA	https://huggingface.co/FacebookAI/xlm-roberta-base

ibm-granite/granite-docling-258M	Ingesta y Procesamiento	Parser/VLM para conversión end-to-end de documentos en Docling (preserva estructura: tablas, ecuaciones, layout).	Release reportado: 17-Sep-2025; modelo ~258M con mejoras (tablas/ecuaciones/estabilidad) según model card.	https://huggingface.co/ibm-granite/granite-docling-258M
TableFormer (Docling)	Ingesta y Procesamiento	Reconocimiento de estructura de tablas (filas/columnas/celdas) dentro del pipeline Docling.	TEDS reportado para TableFormer: 95.4 (simple), 90.1 (complex), 93.6 (all tables) en Docling Models.	https://huggingface.co/docling-project/docling-models
RapidOCR (RapidAI)	Ingesta y Procesamiento	OCR multi-backend/multiplataforma para extraer texto de imágenes y PDFs escaneados.	NA	https://rapidai.github.io/RapidOCRDocs/main/
ai/qwen3-vl:latest (Qwen3-VL-8B Instruct, GGUF cuantizado)	Ingesta y Procesamiento	Visión-lenguaje para comprensión de documentos/imágenes.	Top 10 en LLM arena para visión. MMMU (val): 74.1; MathVista (Main): 62.7; MathVision (Main): 77.7; RealWorldQA: 72.5; OCRBench: 819.0; (Multi-turn score): 8.0	https://hub.docker.com/r/ai/qwen3-vl
Qwen/Qwen3-Embedding-4B	Agente RAG/Vectorizaci	Embeddings para búsqueda semántica/recuperación/cluste	MTEB (Multilingual) Mean(Task)=69.45;	https://huggingface.co/Qwen/Qwe

	ón y almacenamiento	ring/clasificación (texto y código) con dimensión hasta 2560.	Mean(Type)=60.86 (Qwen3Embedding4B)	n3-Embedding-4B
gpt-5.2-chat (Azure OpenAI / Microsoft Foundry)	Agente RAG/Vectorización y almacenamiento	Agente conversacional empresarial (chat) para tareas multi-turn con razonamiento adaptativo.	OpenAI (11Dic2025): SWEbench Verified 80.0%, GPQA Diamond 92.4%, AIME 2025 100.0% (GPT5.2 Thinking).	https://ai.azure.com/catalog/models/gpt-5.2-chat

Evaluación y resultados

Análisis descriptivo de documentos

Conforme se definió en la sección anterior Población y Muestra. El análisis descriptivo de los dominios de conocimiento muestra los 6 dominios de forma resumida en un análisis en cuanto a la muestra considerando el tamaño, tipo y metadatos.

Esta sección presenta el análisis descriptivo del corpus documental procesado, considerando variables de tamaño, tipología y metadatos estructurales distribuidos en los 6 dominios funcionales definidos en la muestra. De los 120 documentos recopilados, con 20 por cada dominio, 113 completaron exitosamente el pipeline de ingesta y procesamiento, lo que corresponde a una Tasa de Éxito de Parsing del 94,2%. La Tabla 9 reporta las estadísticas descriptivas calculadas sobre los 113 documentos efectivamente procesados.

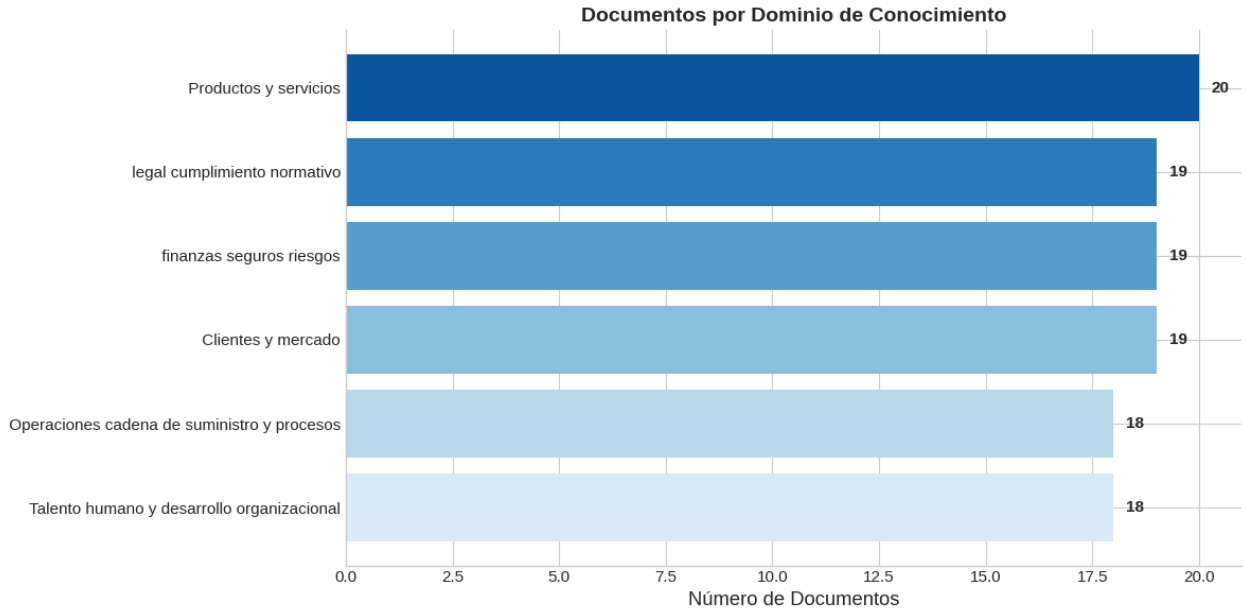


Figura 16. Número de documentos por dominio

El dominio de conocimiento donde se consiguió el mayor número de documentos es Productos y Servicios, esto se debe a la variedad de documentos de acceso público en este dominio. En la siguiente tabla se puede observar la distribución por dominio de conocimiento.

Tabla 9. Estadísticas descriptivas de documentos por dominio

Dominio de Conocimiento	***Número de documentos***	***Tamaño promedio en kB***	***Páginas promedio***
Productos y servicios.	20	3122.80	21.20
Clientes y mercado	19	2802.99	23.37
Finanzas, seguros y riesgos.	19	889.02	25.00
Legal, cumplimiento normativo.	19	1015.64	12.42
Operaciones, cadena de suministro y procesos.	18	1085.05	13.56
Talento humano y desarrollo organizacional.	18	862.07	19.28

En cuanto a tamaño se cuenta con documentos de alrededor de 5Kb es decir 5 Mb con 5 casos excepcionales donde el peso llega a un máximo de 20 Mb. En cuanto longitud de los documentos se mantiene alrededor de las 20 páginas sin embargo existen 3 documentos donde la extensión llegan a 80 y 160 páginas[1].

[1] Para propósito de una mejor visualización se ha omitido los valores extremos en los gráficos.

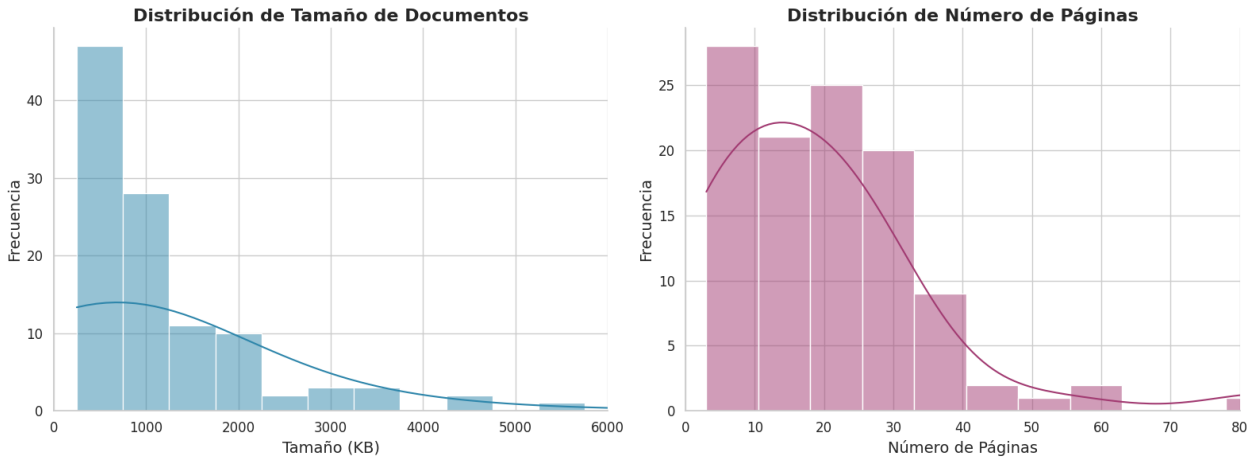


Figura 17. a) En la imagen izquierda se describe la distribución en frecuencia del peso en Kilo Bytes de todos los documentos en la muestra; b) En la imagen derecha longitud de páginas por documento.

A partir de las representaciones textuales obtenidas en el pipeline de ingesta, se realizó un análisis de lenguaje natural (NLP) que comprende el análisis de la distribución en cuanto a longitud de texto a nivel de caracteres, longitud de texto a nivel de tokens. La representación de los tokens ha sido obtenida mediante el modelo knowledgator/gliner-bi-large-v1.0, modelo que también permite la extracción de entidades nombradas (NER). El proceso se describe a en la figura a continuación.

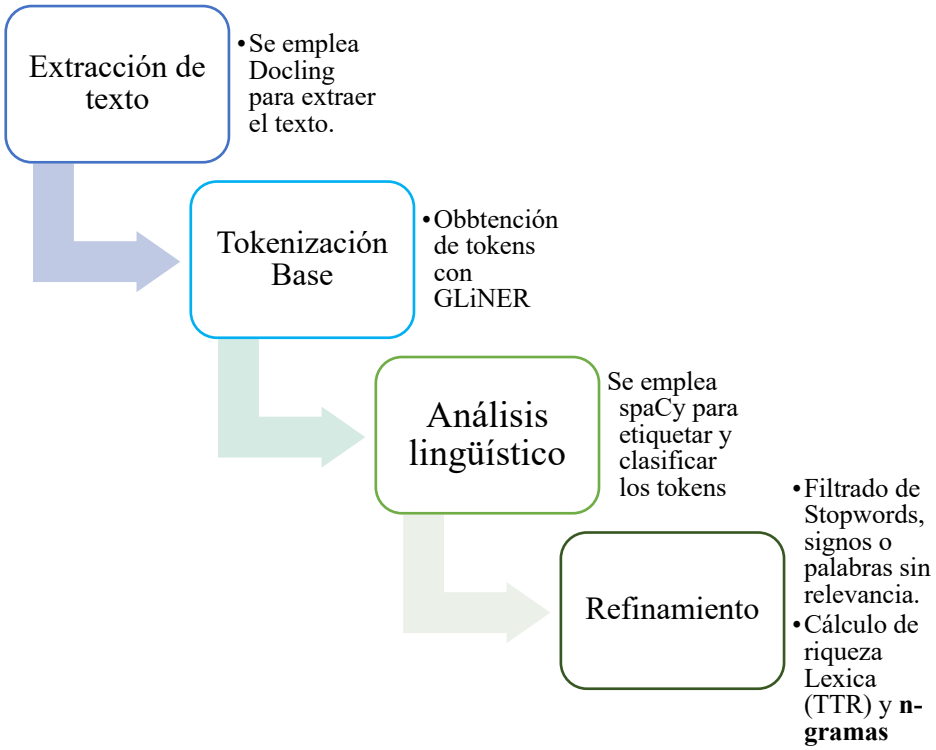


Figura 18. Análisis de textos aplicado a todos los dominios de conocimiento.

Como resultado de este análisis para todos los dominios se muestran las siguientes distribuciones.

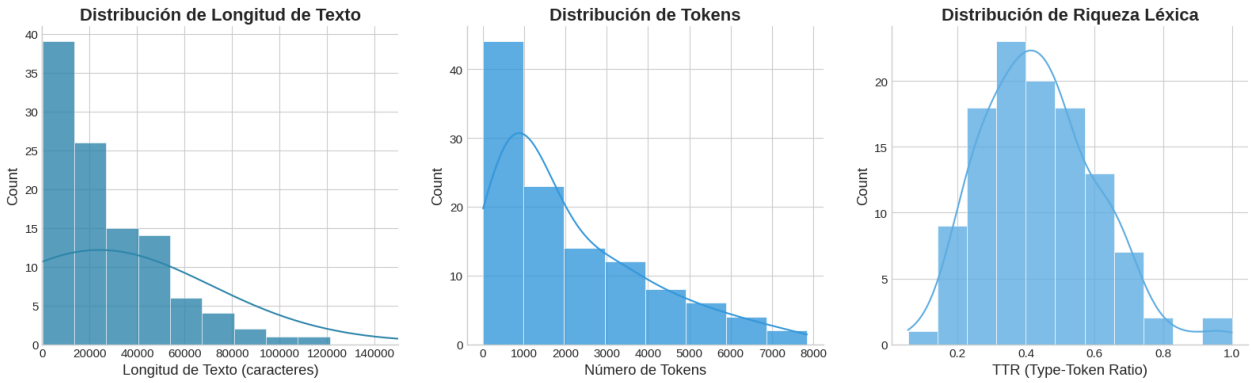


Figura 19. a) Longitud de texto a nivel caracteres; b) Número de tokens en los textos; c) Riqueza léxica

Las distribuciones de longitud de texto nivel de carácter y tokens tiene un comportamiento de ley de tipo potencia con cola pesada a la izquierda, esto implica que la mayoría de texto en los dominios de conocimiento constan de textos cortos cerca de los 100 000 caracteres y 1000 tokens. En cuanto a la riqueza léxica, medida mediante el índice Type-Token Ratio (TTR), sigue una distribución normal, donde este ratio se calculada según la fórmula:

$$TTR = \frac{Types}{Tokens}$$

Donde:

- Types: Número de palabras únicas en el texto (exclusión de palabras repetidas)
- Tokens: Número total de palabras

La interpretación del TTR implica que cuando se obtiene un valor cercano a 1.0 casi cada palabra en el texto es única (textos muy densos o técnicos) por otro lado cuando el valor es cercano a 0 indica un vocabulario sumamente repetitivo. En el caso del presente estudio los dominios analizados se reflejan una distribución estándar alrededor de 0.4.

Análisis de lenguaje natural - NLP

Cada dominio es independiente y tiene su propio vocabulario típico del dominio de conocimiento, ordenando por frecuencia se tiene las siguientes nubes de palabras.

Figura 20. Nubes de palabras de frecuencia de aparición de palabras por dominio de conocimiento.

En términos de todos los dominios de conocimiento se tiene las siguientes palabras como las más frecuentes:

*Tabla 10. Top 10 palabras más frecuentes en todos los dominios de conocimiento (frecuencia global), **

Posición	***Palabra***	***Frecuencia***	***Porcentaje***
1º	*Producto*	*989*	*1.82%*
2º	*Servicio*	*955*	*1.75%*
3º	*Cliente*	*812*	*1.49%*
4º	*Riesgo*	*803*	*1.47%*
5º	*Información*	*698*	*1.28%*
6º	*Personal*	*571*	*1.05%*
7º	*Control*	*560*	*1.03%*
8º	*Empresa*	*539*	*0.99%*
9º	*Crédito*	*495*	*0.91%*
10º	*Dato*	*489*	*0.90%*

***** Este análisis es considerando un total de 54473 menciones (distribuidas en 990 palabras únicas).

En cuanto a la interpretación de los valores de la tabla, la frecuencia indica el número total de veces que aparece una palabra, permitiendo identificar los conceptos más repetidos de forma directa. Por

otro lado, el porcentaje revela el peso real de cada término frente al volumen total de palabras en el corpus, demostrando qué tan dominante es un tema en los documentos.

Figura 21. Representación de nube de palabras de frecuencias de todos los dominios

Estadísticas de estructura

La estructura de los documentos analizados comprende una distribución que se concentra mayormente en el formato PDF donde se presentan las diferentes casuísticas (Documentos textuales, Escaneados, Textos con tablas complejas) mostradas en la Tabla 5,

Tabla 11. Tipologías de documentos en los dominios de conocimiento.

Tipo de documento	***Frecuencia***
PDF	99
Power Point	10
Excel	3
HTML	1

Evaluación de Implementación de la arquitectura

Métricas de evaluación del sistema

Para poder evaluar el éxito del agente se han planteado cinco métricas técnicas en tres dimensiones clave, que se detallan a continuación:

Tabla 12: Métricas técnicos según la fase del agente

Fase del Sistema	***Métrica Clave***
Extracción	Tasa de éxito de procesamiento documental
Recuperación	Precisión (Precision@k)
Recuperación	Similitud Semántica (Cosine Similarity)
Generación	Factualidad (Faithfulness)
Generación	Coherencia de Respuesta

Según los dos principales marcos de trabajo ARES y RAGAS, dentro de la fase de extracción se ha planteado evaluar la tasa de éxito de procesamiento documental. Para la fase de recuperación se han planteado dos métricas, i) tasa de precisión o precision@k, ii) tasa de similitud semántica o cosine similarity. Finalmente, en la fase de generación también se ha planteado dos Métricas, i) tasa de factualidad o faithfulness y ii) tasa de coherencia de respuesta. (Es et al., 2025; Saad-Falcon et al., 2024a). Para poder calcular las métricas se ha creado un conjunto de 50 preguntas, con las cuales se obtendrán los insumos necesarios para dichos indicadores. El conjunto de 50 preguntas se encuentra disponible en el Anexo 1.

Métricas de la Fase de extracción:

1) **Tasa de éxito de procesamiento:**

Para esta validación se aplicó el pipeline de transformación de documentos de dominio de conocimiento. La ejecución está configurada para que cuando se ejecute persista los siguientes documentos: JSON Doctags, markdown, imágenes y tablas como CSV.

Para determinar la tasa de éxito de la conversión de documentos, analizamos cuantos documentos fueron convertidos de forma exitosa de acuerdo con la siguiente ecuación.

$$Exito_{procesamiento} = \frac{\Sigma\, Docs_{parseados}}{\Sigma\, Docs_{dominios}}$$

Se toma como parámetros el número de documentos convertidos sin errores sobre el número de documentos totales.

$$Exito_{procesamiento} = \frac{113}{120} = 94\%$$

En conclusión, la Tasa de Éxito de procesamiento del 94,2% evidencia que el pipeline de ingesta basado en Docling es efectivo para el procesamiento de corpus documentales heterogéneos.

Métricas de la Fase de recuperación:

1. **Precisión/ Precision@k:** esta métrica evalúa la calidad de los chunks obtenidos, es decir considera el porcentaje de chunks útiles respecto a la totalidad de chunks. La interpretación de esta métrica es que tiende a ser mejor en función de acercarse a 100%. Se calcula con la siguiente fórmula.

$$Precision@k = \frac{\sum chunks_{útiles}}{\sum chunks_{totales}}$$

Para cada pregunta, se ha establecido que se obtengan entre 15 y 20 chunks relacionados. Cada chunk ha sido evaluado para identificar si ayuda o no a responder la pregunta ingresada por el usuario en función de los distintos dominios de conocimiento planteados en la sección Población y Muestra.

Por ejemplo:

Pregunta s_015: ¿qué significa obstaculización para los empleados?

$$Precision@k_{s\_015} = \frac{17}{19} = 89\%$$

Tabla 13: Ejemplo tasa de precision@k

Id_chunk	***Chunk***	***Resultado***
274	Las renuncias o cualquier disposición con respecto a la notificación de renuncia por parte de un empleado dependerán de las condiciones de permanencia y continuidad del contrato que están vigentes […].	Útil
277	Los medios para la divulgación de vacantes disponibles se realizarán por medio de los colaboradores que refieran conocidos suyos, anuncios impresos, anuncios en medios electrónicos, bolsas de trabajo, eventos masivos, ferias de empleo, etc.	No útil

Analizando los chunks de las 50 preguntas se ha obtenido una tasa de precision@k de 84%.

$$Precision@k_{total} = \frac{771}{922} = 84\%$$

En conclusión, el agente obtuvo una Tasa de Precisión@20 del 84% sobre el conjunto de 50 preguntas de evaluación, lo que indica que, en promedio, 17 de cada 20 fragmentos recuperados son relevantes para responder la consulta del usuario. Este resultado es homogéneo entre dominios, lo que evidencia que el mecanismo de enrutamiento y recuperación híbrida opera de forma consistente independientemente del contexto semántico consultado. El 16% de fragmentos no relevantes es atribuible a la naturaleza probabilística de la búsqueda semántica y representa un margen esperado en sistemas RAG, sin comprometer la capacidad del agente para generar respuestas fundamentadas en evidencia documental

2. **Tasa de similitud semántica/cosine similarity:** evalúa la relevancia semántica de los chunks recuperados respecto a la consulta, midiendo la cercanía vectorial entre ambas representaciones en el espacio de embeddings. A diferencia de Precision@k, que determina si un fragmento contribuye a responder la pregunta, esta métrica evalúa únicamente si el fragmento recuperado pertenece al mismo dominio semántico que la consulta, independientemente de su utilidad para la generación de la respuesta.

Considera la siguiente fórmula:

$$Cosine\_sim_{chunks\_utiles} = \frac{\sum||cosine\_sim\_sem||}{n}$$

Donde $||cosine\_sim\_sem||$ es la similitud semántica (abreviada como sim_sem) normalizada al máximo y mínimo de su distribución.

Para poder calcular el cosine similarity es necesario analizar la distancia de similitud semántica que se obtiene al aplicar el operador <=> entre dos vectores en PostgreSQL pgvector , esto es parte del proceso de descrito en la capa de recuperación para el RAG, con este indicador se puede obtener el score de similitud semántica con la siguiente fórmula:

$$Score\_sim\_sem_{i=1:50,j=1:20} = \frac{1}{1 - distancia\_sim\_sem_{ij}}$$

Sin embargo, los scores de similitud coseno no son directamente comparables entre las preguntas planteadas, dado que la variabilidad semántica entre dominios introduce diferencias sistemáticas en la magnitud de los valores recuperados. Para hacer los scores comparables entre sí, se aplica una normalización min-max por pregunta mediante la siguiente fórmula.

$$Scores\_||cosine\_sim\_sem||_{\ i=1:50,j=1:20} = \frac{min_i - score\_sim\_sem_{ij}}{max_i - min_i}$$

Analizando los chunks de las 50 preguntas se ha obtenido una tasa de cosine sim de 82%.

$$Cosine\_sim_{chunks\_utiles} = 82\%$$

En conclusión, el agente obtuvo una Tasa de Similitud Semántica Coseno del 82% sobre el conjunto de 50 preguntas de evaluación, lo que indica que 8 de cada 10 fragmentos recuperados pertenecen al mismo espacio semántico que la consulta planteada. Este resultado, consistente con la Tasa de Precisión@20 obtenida, confirma que el agente recupera fragmentos temáticamente coherentes de forma sistemática a lo largo de todos los dominios de conocimiento evaluados. La relevancia de este indicador trasciende el ámbito técnico: una tasa de similitud semántica del 82% implica que el sistema es capaz de interpretar correctamente la intención de la consulta en lenguaje natural y dirigirla al conocimiento organizacional pertinente, lo que constituye una evidencia concreta de que el sistema agéntico propuesto contribuye a hacer accesible el conocimiento institucional almacenado en documentos heterogéneos, mejorando directamente los procesos de gestión documental de la organización.

Métricas de la Fase de generación:

1. **Tasa de factualidad/faithfulness:** este indicador evalúa si la respuesta entregada por el agente es correcta o incorrecta.

Considera la siguiente fórmula:

$$Tasa\ de\ factualidad = \frac{\sum respuestas\ correctas}{respuestas\ totales}$$

Por ejemplo:

CASO 1:

Pregunta s_004: *¿Cuánto tiempo tiene el Usuario del Centro de Cumplimiento para confirmar una solicitud resuelta?*

Respuesta esperada: *48 horas.*

Respuesta entregada: *El Usuario del Centro de Cumplimiento tiene un plazo de 48 horas para confirmar una solicitud que haya sido resuelta.*

Conclusión: *respuesta correcta*

CASO 2:

Pregunta s_043: *¿En qué casos el Banco puede cerrar de manera unilateral las cuentas de una persona jurídica según el contrato marco?*

Respuesta esperada: *El banco puede cerrar cuentas por falta de fondos en cheques, operaciones irregulares o fraudulentas, información falsa proporcionada por el cliente, inclusión en listas de la Superintendencia de Banca y Seguros, o por decisión de negocios con comunicación previa*

Respuesta entregada: *Según el Contrato Marco de Productos Pasivos y Servicios Transaccionales para Personas Jurídicas del Banco, el Banco puede cerrar de manera unilateral las cuentas de una persona jurídica principalmente en los siguientes casos:*

1.Incumplimiento contractual

Cuando el cliente incumple las obligaciones, condiciones o compromisos asumidos en el contrato marco, en la solicitud de productos o en cualquiera de los documentos que forman parte integrante del contrato.

2. Incumplimiento de requisitos legales o regulatorios

Si el cliente deja de cumplir los requisitos mínimos exigidos por la normativa vigente, incluidos los filtros regulatorios aplicables (por ejemplo, normas financieras, de prevención de lavado de activos y financiamiento del terrorismo).

3. Resultados adversos en filtros o controles regulatorios

Cuando, como resultado de evaluaciones, verificaciones o controles internos exigidos por la ley, el Banco determine que el cliente representa un riesgo legal, operativo, reputacional o regulatorio.

4. Falta de información o información inconsistente

Si el cliente no proporciona la información requerida, la proporciona de forma incompleta o inconsistente, o no la actualiza cuando es exigido por el Banco conforme a la ley.

5. Disposición legal o mandato de autoridad competente

Cuando una norma legal, regulación vigente o instrucción de una autoridad competente obligue al Banco a proceder con el cierre de la cuenta.

En estos supuestos, el Banco está facultado para dar por terminada la relación contractual y cerrar las cuentas de forma unilateral, en cumplimiento del contrato y de las obligaciones legales que rigen su actividad.

Conclusión: *respuesta incorrecta*

Analizando los chunks de las 50 preguntas se ha obtenido una tasa de faithfulness de 84%.

$$Tasa\ de\ factualidad = \frac{42}{50} = 84\%$$

En conclusión, el agente obtuvo una Tasa de Faithfulness del 84% sobre el conjunto de 50 preguntas de evaluación, lo que indica que 8 de cada 10 respuestas generadas están correctamente fundamentadas en la evidencia documental recuperada. Este resultado es el más significativo del conjunto de métricas evaluadas, dado que la faithfulness mide directamente la confiabilidad del sistema: una respuesta fiel no solo es temáticamente relevante, sino que su contenido puede ser trazado y verificado en las fuentes documentales del corpus. Alcanzar un 84% en esta dimensión demuestra que el agente supera el umbral mínimo de precisión establecido en los objetivos de la investigación y valida que la arquitectura RAG agéntica propuesta es una solución efectiva para hacer accesible el conocimiento organizacional mediante consultas en lenguaje natural, reduciendo la dependencia de procesos manuales de consulta documental y fortaleciendo la toma de decisiones basada en evidencia dentro de la organización."

2. **Tasa de coherencia de respuesta:** esta métrica evalúa si la respuesta entregada por el agente es útil y sin alucinaciones, es decir no basta con que responda a la pregunta, no debe

agregar información incorrecta. Para esta evaluación se emplea técnica modelos LLM como juez (LLM-as-a-Judge). Esta consiste en utilizar un modelo de lenguaje adicional con una configuración calibrada para evaluar automáticamente la calidad de las respuestas generadas por el sistema.

Considera la siguiente fórmula:

$$Tasa\ de\ factualidad = \frac{\sum calif\_respuesta_{4-5}}{respuestas\ totales}$$

Donde:

$calif\_respuesta_{4-5}$: se ha usado a un LLM como juez de calificación de coherencia, es decir ha otorgado un score del 1-5 sobre qué tan coherente es, donde 1 es poco coherente y 5 es muy coherente. El LLM juez tiene la siguiente calibración:

- Modelo juez: Gemini 3 Flash
- Instrucción: Evaluar un score de 1 a 5 de coherencia entre la respuesta esperada (ground truth) y la respuesta generada por el sistema.
- Temperatura = 0.2 – Evaluación ajustada a determinismo.
- Top P = 0.95 – Elige solo tokens muy probables que sumen 95%.

Se considera como respuesta coherente a todas las respuestas con un score del LLM juez de 4 o 5 puntos.

Por ejemplo:

CASO 1:

Pregunta s_050*: ¿Cuál es el plazo máximo establecido por Esperanza de Vida ONG para completar un proceso de reclutamiento externo?*

Respuesta esperada*: El proceso de reclutamiento externo tiene un plazo máximo de 30 días a partir de la requisición de personal para colocar a una persona en la vacante disponible*

Respuesta entregada: *Con base exclusiva en el contexto proporcionado, no se encuentra establecido ni mencionado un plazo máximo para completar un proceso de reclutamiento externo en la Organización Cristiana de Beneficio Social Esperanza de Vida ONG.*

LLM judge score = 1

CASO 2:

Pregunta s_001: *¿Qué puesto ocupa Ecuador en el Informe de Competitividad Global 2019 del Foro Económico Mundial?*

Respuesta esperada: *Puesto 90 de 141 países.*

Respuesta entregada*: Según el Informe de Competitividad Global 2019 del Foro Económico Mundial, Ecuador ocupa el puesto 90 entre 141 países evaluados.*

LLM judge score = 5

Analizando las 50 preguntas se ha obtenido una tasa de coherencia de respuesta de 86%

$$Tasa\ de\ factualidad = \frac{43}{50} = 86\%$$

En conclusión, el agente obtuvo una Tasa de Coherencia de Respuesta del 86% sobre el conjunto de 50 preguntas de evaluación, lo que indica que 9 de cada 10 respuestas generadas son coherentes, bien estructuradas y libres de alucinaciones. Este resultado es particularmente relevante en el contexto de sistemas basados en modelos de lenguaje de gran escala, cuya naturaleza probabilística representa uno de los principales riesgos de adopción en entornos organizacionales. Alcanzar un 86% de coherencia evidencia que el diseño arquitectónico del sistema, específicamente el anclaje de la generación en fragmentos documentales verificables mediante RAG, mitiga de forma efectiva la tendencia a la alucinación inherente a los LLM. Este indicador, analizado en conjunto con la Tasa de Faithfulness del 84%, confirma que el agente no solo recupera información relevante, sino que la articula de manera confiable, consolidando al sistema propuesto como una herramienta viable para la gestión del conocimiento organizacional.

En la Tabla 14, se muestran los Métricas técnicos de las 3 fases del agente RAG, donde se evidencia que se ha cumplido la meta esperada exitosamente, validando que el agente tiene un rendimiento óptimo y confiable.

Tabla 14. Resumen Métricas técnicos

Fase del Sistema	***Métrica Clave (KPI)***	***Meta Esperada***	***Justificación***	***Resultado***
Extracción	Tasa de éxito OCR/Parsing	> 80%	Garantiza que la "materia prima" (datos) sea de calidad.	**94%**

Recuperación	Precisión (Precision@k)	≥ 80%	Cumplimiento del Objetivo Específico 2.	**84%**
Recuperación	Similitud Semántica (Cosine Sim)	~ 0.70	Tu estimación inicial para asegurar relevancia contextual.	**82%**
Generación	Factualidad (Faithfulness)	≥ 70%	Cumplimiento del Objetivo Específico 3.	**84%**
Generación	Coherencia de Respuesta	> 80%	Garantiza que la respuesta sea legible y útil para el humano.	**86%**

En síntesis, los resultados obtenidos sobre el conjunto de 50 preguntas de evaluación demuestran que el sistema agéntico propuesto opera con niveles de desempeño consistentes a lo largo de todas las dimensiones evaluadas: Tasa de Éxito de Parsing 94,2%, Precisión@20 84%, Similitud Semántica Coseno 82%, Faithfulness 84% y Coherencia de Respuesta 86%. La homogeneidad de estos scores entre dominios confirma que la arquitectura RAG agéntica es robusta frente a la heterogeneidad semántica del corpus, validando su utilidad como solución tecnológica para hacer accesible el conocimiento organizacional mediante consultas en lenguaje natural.

Estos niveles de desempeño no solo validan la arquitectura desde una perspectiva técnica, sino que permiten confiar en el agente como intermediario confiable entre el usuario y el conocimiento institucional, lo que tiene una implicación directa sobre la gestión documental organizacional. Al centralizar documentos de múltiples formatos y dominios en una única arquitectura de conocimiento, el sistema aborda uno de los problemas estructurales más frecuentes en las organizaciones: el conocimiento disperso entre áreas, sistemas y repositorios heterogéneos sin interoperabilidad. En este sentido, la capacidad del agente para identificar el dominio al que pertenece una consulta en lenguaje natural y recuperar los fragmentos pertinentes con independencia del área de origen demuestra que es posible unificar el acceso al conocimiento organizacional sin requerir que el usuario conozca la ubicación, el formato o la estructura de los documentos subyacentes.

Análisis de fallos y robustez del sistema

La evaluación del sistema no se limita al reporte de métricas de desempeño satisfactorias, sino que requiere una caracterización rigurosa de los casos en que el sistema no alcanzó el comportamiento esperado. Con este propósito, el análisis de fallos se estructura en dos dimensiones

complementarias: el comportamiento del agente frente al conjunto de evaluación general y su robustez ante consultas adversariales diseñadas para inducir respuestas incorrectas.

Fallos en la dimensión de fidelidad

La Tasa de Faithfulness del 84% implica que en el 16% de los casos evaluados la respuesta generada no pudo ser completamente atribuida a la evidencia documental recuperada. El análisis cualitativo de estos casos permitió identificar dos patrones sistemáticos de fallo. El primero se asocia a consultas formuladas con un alto nivel de generalidad semántica, en las que la imprecisión de la pregunta impide que el mecanismo de recuperación delimite con exactitud los chunks relevantes, resultando en respuestas que abordan el tema de manera genérica en el conocimiento del LLM sin fundamentarse en documentos específicos. El segundo patrón corresponde a consultas cuyos conceptos centrales aparecen distribuidos transversalmente entre múltiples documentos de un mismo dominio, provocando que el agente recupere fragmentos temáticamente relacionados pero provenientes de fuentes heterogéneas, lo que deriva en respuestas que integran contextos distintos y deterioran la precisión factual de la respuesta. La identificación de estos patrones permite concluir que las limitaciones observadas no son deficiencias del modelo LLM empleado en el agente, sino a factores de diseño relacionados con la formulación de las consultas y la granularidad del mecanismo de recuperación semántica.

Robustez frente a consultas adversas

Con el propósito de evaluar la capacidad del sistema para controlar consultas que buscan inducir a las alucinaciones, comparaciones semánticamente incorrectas entre documentos o referencias a entidades inexistentes en el corpus, se diseñó un conjunto de 10 preguntas adversas clasificadas en cuatro categorías: comparación incorrecta entre documentos, corrección de premisas falsas, consultas fuera del alcance del corpus y atribución de contenido o versiones inexistentes.

Tabla 15. Métricas de fallo

Categoría	***Preguntas***	***Respuestas correctas***	***Tasa***
Comparación incorrecta entre documentos	3	3	100%
Corrección de premisa falsa	1	1	100%
Consulta fuera del corpus	1	1	100%
Atribución de contenido inexistente	3	3	100%
Referencia a versión inexistente en el corpus	2	0	0%
Total	**10**	**8**	**80%**

El sistema respondió correctamente en 8 de los 10 casos adversos, alcanzando una tasa de robustez del 80%. Los casos exitosos evidencian que el agente es capaz de discriminar entre documentos semánticamente relacionados pero temáticamente distintos, contrastar premisas falsas con la evidencia disponible en el corpus, abstenerse de responder ante consultas cuyo tema no está representado en la base de conocimiento, y reconocer la ausencia de información específica, como por ejemplo la identidad del firmante de un documento o la disponibilidad de una traducción, sin inferir ni fabricar datos no presentes en las fuentes.

Los dos fallos identificados corresponden a preguntas que hacen referencia a la "VII Edición de la Política de Reclutamiento de Esperanza de Vida", versión inexistente en el corpus. Este comportamiento es consistente con una limitación estructural documentada en sistemas RAG: cuando una consulta referencia una entidad con alta similitud semántica respecto a un documento existente en la base vectorial, el mecanismo de recuperación tiende a recuperar el documento más próximo en el espacio de embeddings y el modelo de lenguaje procede a generar una respuesta a partir de él, sin detectar que la entidad específica referenciada no corresponde a ningún documento indexado. En el caso analizado, la VI Edición disponible en el corpus actuó como ancla semántica, induciendo al agente a responder como si la versión solicitada existiera.

Considerados en conjunto, ambos análisis revelan que los fallos del sistema presentan un carácter acotado, concentrándose en los límites semánticos del corpus: consultas excesivamente generales, superposición de conceptos entre documentos de un mismo dominio y referencias a entidades semánticamente próximas a documentos existentes, pero no disponibles en la base de conocimiento (caso de VI Edición). Estos errores constituyen una contribución relevante del presente estudio, dado que provee un criterio para el diseño de mecanismos de mitigación en investigaciones futuras, entre los que se destacan la incorporación de un paso de verificación explícita de existencia de entidades previo a la generación de la respuesta y el establecimiento de criterios formales para la formulación de consultas en entornos organizacionales.

Interfaz de presentación

Para poder complementar la capa de presentación se ha implementado una interfaz gráfica de tipo web UI, que está construida en la tecnología Streamlit de Python. Se ha implementado múltiples features en la interfaz para mejorar la experiencia del usuario. A continuación, se presentan las pantallas en una secuencia de ejecución del agente ante una pregunta sencilla de autoconocimiento: ¿Quién eres?.

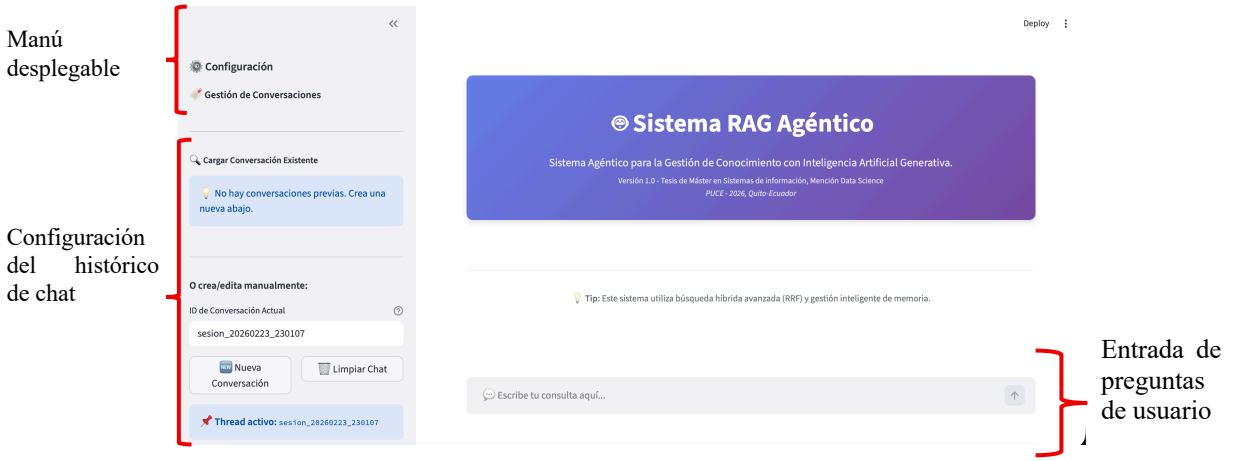


Figura 22. Interfaz de Sistema RAG, header

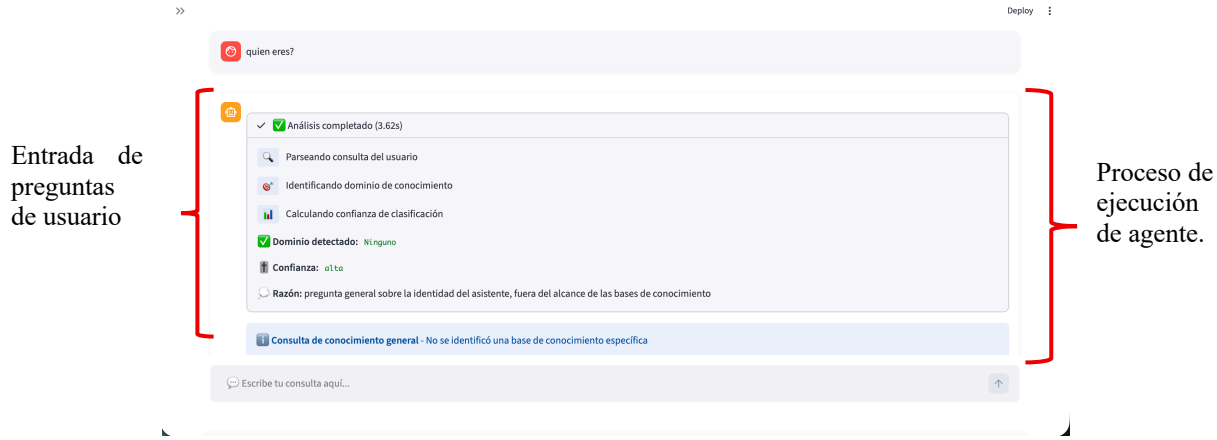


Figura 23. Interfaz módulo de ejecución del agente

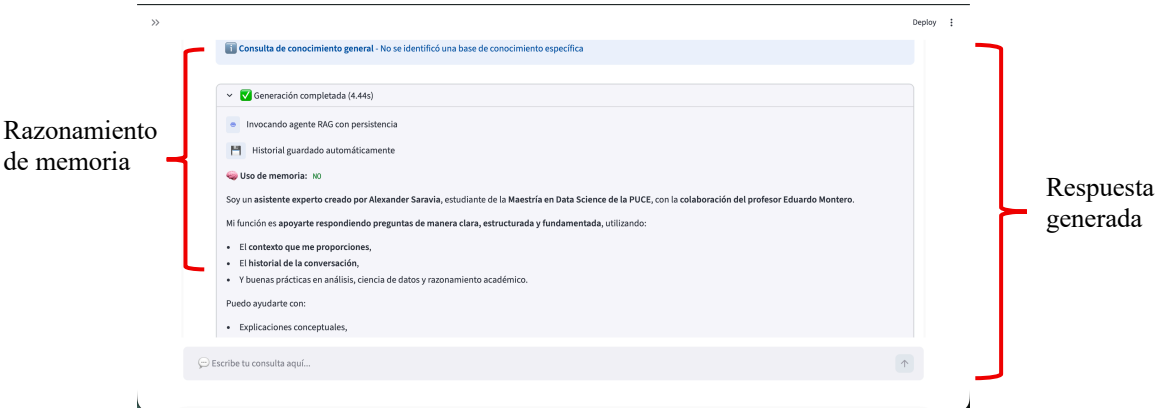


Figura 24. Interfaz, razonamiento de la memoria.

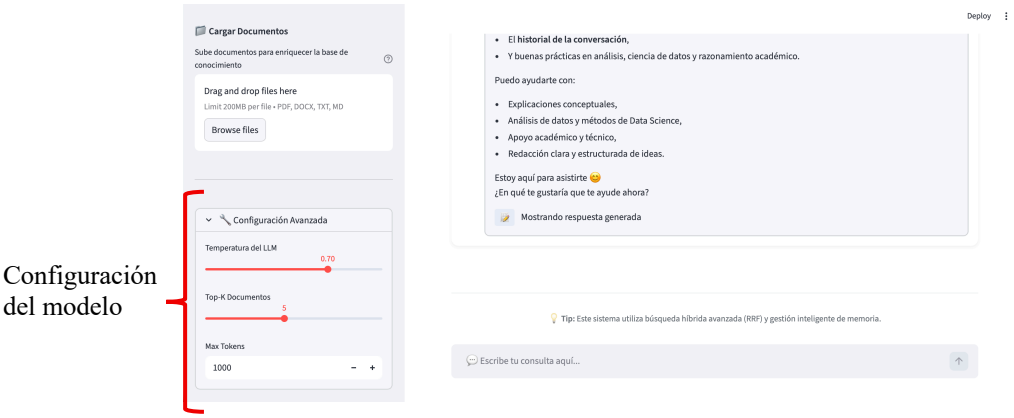


Figura 25. Interfaz, respuesta generada

En la interfaz se van a encontrar las preguntas, razonamiento, y respuesta referenciada de modo que el usuario pueda navegar por su historial de chat e interactuar de forma iterativa con el agente.

Conclusiones y Recomendaciones

Conclusiones

El presente trabajo partió de una problemática organizacional recurrente: el conocimiento institucional se encuentra disperso en repositorios heterogéneos, fragmentado entre áreas, formatos y sistemas sin interoperabilidad, lo que limita su acceso, recuperación y aprovechamiento eficiente. Frente a esta brecha, se diseñó, implementó y evaluó una arquitectura de un sistema agéntico basado en Retrieval-Augmented Generation (RAG), demostrando que es posible centralizar el acceso al conocimiento organizacional mediante consultas en lenguaje natural, con independencia del formato, estructura o dominio de origen de los documentos.

Desde una perspectiva de arquitectura de desarrollo, el ejercicio de implementación permite concluir que un sistema de gestión del conocimiento de esta naturaleza requiere como mínimo cuatro capas a nivel de tecnología: ingesta y procesamiento documental, particionamiento semántico, vectorización y almacenamiento, y agente de consulta RAG; acompañadas de tres roles operativos claramente definidos: dueño de dominio de conocimiento, administrador de base de conocimiento y usuario final. La ausencia de cualquiera de estos componentes comprometería la trazabilidad, mantenibilidad y confiabilidad del sistema en un entorno organizacional real. Adicionalmente, el análisis de lenguaje natural (NLP) realizado sobre el corpus confirmó que los dominios presentan vocabularios diferenciables con una riqueza léxica estándar aproximada de 0,4, condición adecuada para la aplicación de técnicas de recuperación semántica y léxica.

En la etapa de procesamiento documental, el pipeline basado en Docling obtuvo una Tasa de Éxito de Parsing del 94%, procesando exitosamente 113 de los 120 documentos de los corpus distribuidos en 6 dominios funcionales. En la etapa de recuperación, el sistema alcanzó una Precisión@20 del 84% y una Similitud Semántica Coseno del 82%, superando el umbral del 80% establecido en los objetivos de la investigación. La homogeneidad de estos indicadores entre dominios confirma que el mecanismo de enrutamiento y recuperación híbrida funciona de forma consistente frente a la heterogeneidad semántica del corpus. En la etapa de generación, el agente obtuvo una Tasa de Faithfulness del 84% y una Coherencia de Respuesta del 86%, superando el umbral del 70% definido como criterio de éxito, lo que valida que la arquitectura RAG agéntica mitiga de forma efectiva la tendencia a la alucinación de los modelos de lenguaje de gran escala (LLM).

El análisis de fallos reveló que los errores del sistema presentan un carácter acotado. El 16% de respuestas que no alcanzaron el umbral de fidelidad se concentra en dos patrones: consultas formuladas con excesiva generalidad semántica y casos de superposición conceptual entre dominios del corpus. Por su parte, la evaluación adversarial sobre 10 preguntas diseñadas para inducir alucinaciones obtuvo una tasa de robustez del 80%, con fallos exclusivamente asociados a referencias de versiones documentales inexistentes en la base de conocimiento, comportamiento atribuible a la proximidad semántica entre la entidad referenciada y documentos existentes en el corpus. En conjunto, estos hallazgos permiten concluir que las limitaciones del sistema no son

aleatorias ni atribuibles al modelo de lenguaje en sí, sino a factores estructurales relacionados con el diseño de las consultas y los límites semánticos del corpus, aspectos abordables mediante mejoras en el mecanismo de enrutamiento y en los criterios de formulación de preguntas.

Finalmente, es necesario señalar las limitaciones del presente estudio. La evaluación se realizó exclusivamente sobre métricas técnicas automatizadas mediante LLM-as-a-Judge, sin incluir una evaluación con usuarios reales que permita medir la percepción de utilidad y confianza en las respuestas generadas. Asimismo, la arquitectura propuesta no fue comparada experimentalmente frente a enfoques alternativos, como RAG sin componente agéntico o recuperación puramente léxica, lo que limita la posibilidad de cuantificar el aporte específico de cada componente del sistema. Estas limitaciones definen con precisión el alcance del presente trabajo y constituyen las líneas prioritarias para investigaciones futuras, sin comprometer la validez de los resultados obtenidos ni la pertinencia de la solución propuesta frente al problema organizacional identificado.

Recomendaciones

Las recomendaciones que se presentan a continuación se derivan directamente de los resultados empíricos obtenidos y de las limitaciones identificadas en el presente estudio, con el propósito de orientar tanto la mejora del sistema propuesto como el desarrollo de investigaciones futuras en el ámbito de los sistemas agénticos para la gestión del conocimiento organizacional.

En primer lugar, la Tasa de Éxito de Parsing del 94,2% demostró que los fallos del pipeline de ingesta. En consecuencia, se recomienda usar alternativas comerciales a Docling, así como técnicas de preprocesamiento de imagen orientadas a la reducción de ruido visual, con el propósito de reducir el margen de fallo identificado y ampliar la cobertura del sistema frente a formatos de documento de alta complejidad.

En segundo lugar, el análisis de fallos de faithfulness evidenció que la superposición semántica entre documentos de un mismo dominio constituye una fuente sistemática de error en la recuperación. Se recomienda explorar técnicas de reranking post-recuperación, que permitan refinar la selección final de chunks una vez realizada la búsqueda híbrida, reduciendo la incorporación de contexto proveniente de documentos adyacentes. Complementariamente, se recomienda establecer criterios formales para la formulación de consultas en entornos organizacionales, de modo que los usuarios finales cuenten con pautas que minimicen la ambigüedad semántica de sus preguntas.

En tercer lugar, la evaluación frente consultas adversas identificó que el sistema falla ante referencias a versiones documentales inexistentes en el corpus, debido a la proximidad semántica con documentos disponibles. Se recomienda incorporar un paso de verificación explícita de existencia de entidades previo a la generación de la respuesta, que contraste la versión o edición referenciada en la consulta contra los metadatos indexados en la base vectorial antes de proceder con la recuperación y generación.

En cuarto lugar, dado que el presente trabajo fue evaluado mediante un conjunto de preguntas valoradas a partir de los criterios generales de evaluación, se recomienda diseñar un experimento de evaluación con un grupo de entre 3 a 5 expertos de dominios funcionales relevantes (por ejemplo, áreas de Legal o Finanzas) y solicitarles que evalúen 10 respuestas generadas por el sistema, calificándolas según sus criterios de utilidad y confianza. Además, se recomienda complementar estos resultados mediante análisis de "Percepción del Usuario", que permitirá robustecer la validación técnica con una validación sociotécnica, evidenciando significativamente el impacto en el acceso al conocimiento organizacional.

En quinto lugar, la ausencia de una comparación experimental frente a arquitecturas alternativas constituye una limitación reconocida del presente estudio. Se recomienda que investigaciones futuras establezcan una línea base comparativa que contraste la arquitectura RAG agéntica propuesta frente a enfoques de menor complejidad, como RAG sin componente agéntico o recuperación puramente léxica, con el fin de cuantificar el aporte específico del enrutamiento de dominio y del razonamiento agéntico sobre los indicadores de recuperación y generación.

En sexto lugar, la homogeneidad de scores entre dominios valida la arquitectura actual para consultas de recuperación semántica directa, sin embargo, consultas que requieran razonamiento relacional entre documentos de distintos dominios representan un límite estructural del enfoque RAG vectorial. Se recomienda explorar la incorporación de técnicas basadas en grafos de conocimiento, como GraphRAG, para habilitar la navegación de relaciones semánticas complejas entre entidades distribuidas en múltiples dominios, capacidad identificada en la literatura como tendencia relevante para la siguiente generación de sistemas agénticos de gestión del conocimiento.

Finalmente, para garantizar la viabilidad del sistema en un entorno productivo organizacional, se recomienda acompañar la implementación tecnológica con un marco de gobernanza documental estructurado sobre los estándares ISO 30401, DAMA-DMBOK2 y TOGAF, que defina formalmente los roles de dueño de dominio, administrador de base de conocimiento y criterios de calidad documental. Esta dimensión organizacional es condición necesaria para que los niveles de desempeño técnico alcanzados en el presente estudio se traduzcan en una mejora efectiva y sostenible de los procesos de gestión del conocimiento en la organización.

Referencias

Al Ahbabi, A. & others. (2024). The effects of knowledge management processes on service sector performance: Evidence from Saudi Arabia. *Humanities and Social Sciences Communications*. https://doi.org/10.1057/s41599-024-02876-y

Alavi, M., & Leidner, D. E. (2001). Knowledge Management and Knowledge Management Systems: Conceptual Foundations and Research Issues. *MIS Quarterly*, *25*(1), 107–136.

Autio, E. & others. (2024). *NIST AI 600-1: The NIST Generative AI Profile*. National Institute of Standards and Technology. https://doi.org/10.6028/NIST.AI.600-1

Bai, Z., Wang, P., Xiao, T., He, T., Han, Z., Zhang, Z., & Shou, M. Z. (2024a). Hallucination of Multimodal Large Language Models: A Survey. *CoRR*, *abs/2404.18930*. https://arxiv.org/abs/2404.18930

Bai, Z., Wang, P., Xiao, T., He, T., Han, Z., Zhang, Z., & Shou, M. Z. (2024b). Hallucination of Multimodal Large Language Models: A Survey. *CoRR*, *abs/2404.18930*. https://arxiv.org/abs/2404.18930

Becerra-Fernandez, I., & Sabherwal, R. (2015). *Knowledge Management: Systems and Processes*. Routledge. https://doi.org/10.4324/9781315716991

Chen, Y., Zhang, E., Yan, L., Wang, S., Huang, J., Yin, D., & Mao, J. (2025). MAO-ARAG: Multi-Agent Orchestration for Adaptive Retrieval-Augmented Generation. *arXiv preprint arXiv:2508.01005*. https://arxiv.org/abs/2508.01005

Chui, M., Hazan, E., Roberts, R., Singla, A., Smaje, K., Sukharevsky, A., Yee, L., & Zemmel, R. (2023). *The economic potential of generative AI*.

Chui, M., Yee, L., & Hall, B. (2023, agosto). The state of AI in 2023: Generative AI's breakout year. *Sukharevsky Alexander*.

Cook, T., Osuagwu, R., Tsatiashvili, L., Vrynsia, V., Ghosal, K., Masoud, M., & Mattivi, R. (2025a). Retrieval Augmented Generation (RAG) for Fintech: Agentic Design and Evaluation. *arXiv preprint arXiv:2510.25518*. https://arxiv.org/abs/2510.25518

Cook, T., Osuagwu, R., Tsatiashvili, L., Vrynsia, V., Ghosal, K., Masoud, M., & Mattivi, R. (2025b). Retrieval Augmented Generation (RAG) for Fintech: Agentic Design and Evaluation. *arXiv preprint arXiv:2510.25518*. https://arxiv.org/abs/2510.25518

Cormack, G. V., Clarke, C. L. A., & Buettcher, S. (2009). Reciprocal rank fusion outperforms condorcet and individual rank learning methods. *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 758–759. https://doi.org/10.1145/1571941.1572114

Davenport, T. H., & Prusak, L. (1998). *Working Knowledge: How Organizations Manage What They Know*. Harvard Business School Press.

Docling. (2026, febrero). *Advanced options—Docling* [Documentationn]. Advanced options. https://docling-project.github.io/docling/usage/advanced_options/

Es, S., James, J., Espinosa-Anke, L., & Schockaert, S. (2025). *Ragas: Automated Evaluation of Retrieval Augmented Generation* (arXiv:2309.15217). arXiv. https://doi.org/10.48550/arXiv.2309.15217

Forrester. (2023). *Use Knowledge Graphs To Manage Knowledge Instead Of Data* [Use Knowledge Graphs To Manage Knowledge Instead Of Data]. Forrester. https://www.forrester.com/report/use-knowledge-graphs-to-manage-knowledge-instead-of-data/RES179193

Gao, Y. & others. (2024). Retrieval-Augmented Generation for Large Language Models: A Survey. *CoRR*. https://arxiv.org/abs/2312.10997

Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, M., & Wang, H. (2024). *Retrieval-Augmented Generation for Large Language Models: A Survey* (arXiv:2312.10997). arXiv. https://doi.org/10.48550/arXiv.2312.10997

Gartner. (2025). *El Hype Cycle para la IA de 2025 va más allá de la IA generativa*. Gartner. https://www.gartner.es/es/articulos/hype-cycle-para-la-inteligencia-artificial

Gelashvili-Luik, T., Vihma, P., & Pappel, I. (2025). Navigating the AI revolution: Challenges and opportunities for integrating emerging technologies into knowledge management systems. Systematic literature review. *Frontiers in Artificial Intelligence*, *8*, 1595930. https://doi.org/10.3389/frai.2025.1595930

Grant, R. M. (1996). Toward a Knowledge-Based Theory of the Firm. *Strategic Management Journal*, *17*(S2), 109–122.

Hevner, A., & Salvatore, T. M. (2004, marzo). *Design Science in Information Systems Research*.

Hugging Face Community. (2025, marzo 9). *Open LLM Leaderboard*. https://huggingface.co/open-llm-leaderboard

International Organization for Standardization. (2018a). *ISO 30401:2018 Knowledge management systems—Requirements*. https://www.iso.org/standard/68683.html

International Organization for Standardization. (2018b). *ISO 30401:2018 Knowledge management systems—Requirements*. https://www.iso.org/standard/68683.html

Isa, M., & Rahmah, F. A. (2023). Knowledge Management and Organizational Performance: The Mediating Role of Dynamic Capabilities. *JBTI: Jurnal Bisnis: Teori dan Implementasi*, *14*(3), 478–492. https://doi.org/10.18196/jbti.v14i3.20404

James, A., Trovati, M., & Bolton, S. (2025). Retrieval-Augmented Generation to Generate Knowledge Assets and Creation of Action Drivers. *Applied Sciences*, *15*(11), 6247. https://doi.org/10.3390/app15116247

Jennex, M. E., & Olfman, L. (2005). Assessing knowledge management success effectiveness models. *International Journal of Knowledge Management*, *1*(2), 33–49. https://doi.org/10.4018/jkm.2005040103

Kassa, F., & Ning, X. (2023a). A systematic review on the roles of knowledge management in public sectors: Synthesis and way forwards. *Heliyon*, *9*(11), e22293. https://doi.org/10.1016/j.heliyon.2023.e22293

Kassa, F., & Ning, X. (2023b). A systematic review on the roles of knowledge management in public sectors: Synthesis and way forwards. *Heliyon*, *9*(11), e22293. https://doi.org/10.1016/j.heliyon.2023.e22293

Liu, X., Yu, Z., Zhang, Y., Zhang, N., & Xiao, C. (2024). Automatic and Universal Prompt Injection Attacks against Large Language Models. *CoRR*, *abs/2403.04957*. https://arxiv.org/abs/2403.04957

Liu, Y., Zhang, Y., Wang, Y., Hou, F., Yuan, J., Tian, J., Zhang, Y., Shi, Z., Fan, J., & He, Z. (2022). *A Survey of Visual Transformers* (arXiv:2111.06091). arXiv. http://arxiv.org/abs/2111.06091

Livathinos, N., Auer, C., Lysak, M., Nassar, A., Dolfi, M., Vagenas, P., Ramis, C. B., Omenetti, M., Dinkla, K., Kim, Y., Gupta, S., Lima, R. T. de, Weber, V., Morin, L., Meijer, I., Kuropiatnyk, V., & Staar, P. W. J. (2025). *Docling: An Efficient Open-Source Toolkit for AI-driven Document Conversion* (arXiv:2501.17887). arXiv. https://doi.org/10.48550/arXiv.2501.17887

Mahadevkar, S. V., Patil, S., Kotecha, K., Soong, L. W., & Choudhury, T. (2024a). Exploring AI-driven approaches for unstructured document analysis and future horizons. *Journal of Big Data*, *11*(92). https://doi.org/10.1186/s40537-024-00948-z

Mahadevkar, S. V., Patil, S., Kotecha, K., Soong, L. W., & Choudhury, T. (2024b). Exploring AI-driven approaches for unstructured document analysis and future horizons. *Journal of Big Data*, *11*(92). https://doi.org/10.1186/s40537-024-00948-z

Matsumoto, N., Moran, J., Choi, H., Hernandez, M. E., Venkatesan, M., Wang, P., & Moore, J. H. (2024). KRAGEN: A knowledge graph-enhanced RAG framework for biomedical problem solving using large language models. *Bioinformatics*, *40*(6), btae353. https://doi.org/10.1093/bioinformatics/btae353

Minaee, S., Mikolov, T., Nikzad, N., Chenaghlu, M., Socher, R., Amatriain, X., & Gao, J. (2024a). Large Language Models: A Survey. *CoRR*, *abs/2402.06196*. https://arxiv.org/abs/2402.06196

Minaee, S., Mikolov, T., Nikzad, N., Chenaghlu, M., Socher, R., Amatriain, X., & Gao, J. (2024b). Large Language Models: A Survey. *CoRR*, *abs/2402.06196*. https://arxiv.org/abs/2402.06196

Nonaka, I. (1994). A Dynamic Theory of Organizational Knowledge Creation. *Organization Science*, *5*(1), 14–37.

Nonaka, I., Toyama, R., & Konno, N. (2000). SECI, Ba and Leadership: A Unified Model of Dynamic Knowledge Creation. *Long Range Planning*, *33*(1), 5–34.

Peng, B., Zhu, Y., Liu, Y., Bo, X., Shi, H., Hong, C., Zhang, Y., & Tang, S. (2024). Graph Retrieval-Augmented Generation: A Survey. *Proceedings of the ACM Web Conference 2024 (WWW 2024)*. https://doi.org/10.1145/3777378

Polanyi, M. (1966). *The Tacit Dimension*. Routledge.

Qin, Y., Hu, S., Lin, Y., Chen, W., Ding, N., & others. (2024). Tool Learning with Foundation Models. *ACM Computing Surveys*. https://doi.org/10.1145/3704435

Ristoski, P., & Paulheim, H. (2016). Semantic technologies in knowledge management. *European Journal of Information Systems*, *25*(6), 483–500. https://doi.org/10.1057/ejis.2015.7

Saad-Falcon, J., Khattab, O., Potts, C., & Zaharia, M. (2024a). ARES: An Automated Evaluation Framework for Retrieval-Augmented Generation Systems. *Proceedings of NAACL*. https://doi.org/10.48550/arXiv.2311.09476

Saad-Falcon, J., Khattab, O., Potts, C., & Zaharia, M. (2024b). ARES: An Automated Evaluation Framework for Retrieval-Augmented Generation Systems. *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics*. https://doi.org/10.48550/arXiv.2311.09476

Scuotto, V. & others. (2024). The influence of knowledge management on innovation and organizational performance. *Journal of Innovation & Knowledge*. https://doi.org/10.1016/j.jik.2024.100538

Ştefan, S. C., Popa, I., Olariu, A. A., Popa, Ş. C., & Popa, C.-F. (2024). Knowledge management–performance nexus: Mediating effect of motivation and innovation. *Business Process Management Journal*, *30*(8), 27–48. https://doi.org/10.1108/BPMJ-07-2023-0537

Sternad Zabukovšek, S., Jaklič, J., & Bobek, S. (2023a). Managing Document Management Systems' Life Cycle in Relation to an Organization's Maturity for Digital Transformation. *Sustainability*, *15*(21), 15212. https://doi.org/10.3390/su152115212

Sternad Zabukovšek, S., Jaklič, J., & Bobek, S. (2023b). Managing Document Management Systems' Life Cycle in Relation to an Organization's Maturity for Digital Transformation. *Sustainability*, *15*(21), 15212. https://doi.org/10.3390/su152115212

Stewart, I. A., & Buehler, M. J. (2026). *Higher-Order Knowledge Representations for Agentic Scientific Reasoning* (arXiv:2601.04878). arXiv. https://doi.org/10.48550/arXiv.2601.04878

Teece, D. J. (2007). Explicating Dynamic Capabilities: The Nature and Microfoundations of Sustainable Enterprise Performance. *Strategic Management Journal*, *28*(13), 1319–1350.

The Open Group. (2022a). *The TOGAF® Standard, 10th Edition*. The Open Group.

The Open Group. (2022b). *The TOGAF® Standard, 10th Edition*. https://www.opengroup.org/togaf

Vision AI Leaderboard—Best Image & Multimodal Models. (2025). Vision AI Leaderboard - Best Image & Multimodal Models. https://arena.ai

Walsh, J. P., & Ungson, G. R. (1991). Organizational memory. *Academy of Management Review*, *16*(1), 57–91. https://doi.org/10.2307/258607

Wu, Z. & others. (2023). Multimodal Large Language Models: A Survey. *CoRR*, *abs/2311.13165*. https://arxiv.org/abs/2311.13165

Xi, Z., Chen, W., Guo, X., He, W., Ding, Y., Hong, B., Zhang, M., Wang, J., Jin, S., Zhou, E., & others. (2024). The Rise and Potential of Large Language Model Based Agents: A Survey. *arXiv preprint arXiv:2309.07864*. https://doi.org/10.48550/arXiv.2309.07864

Xu, Y., Li, M., Cui, L., Huang, S., Wei, F., & Zhou, M. (2020). LayoutLM: Pre-training of Text and Layout for Document Image Understanding. *Proceedings of the 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. https://doi.org/10.1145/3394486.3403172

Yang, R., Yang, B., Zhao, X., Gao, F., Feng, A., Ouyang, S., Blum, M., She, T., Jiang, Y., Lecue, F., Lu, J., & Li, I. (2025). Graphusion: A RAG Framework for Scientific Knowledge Graph Construction with a Global Perspective. *Companion Proceedings of the ACM on Web Conference 2025, WWW '25*, 2579–2588. https://doi.org/10.1145/3701716.3717821

Yao, Y., Duan, J., Xu, K., Cai, Y., Sun, Z., & Zhang, Y. (2024a). A Survey on Large Language Model Security and Privacy: The Good, The Bad, and The Ugly. *High-Confidence Computing*, *4*, 100211. https://doi.org/10.1016/j.hcc.2024.100211

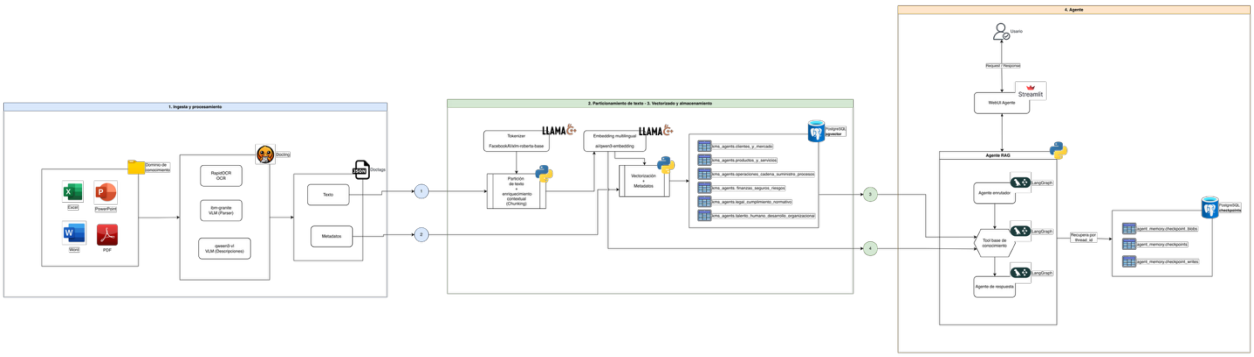
Yao, Y., Duan, J., Xu, K., Cai, Y., Sun, Z., & Zhang, Y. (2024b). A Survey on Large Language Model Security and Privacy: The Good, The Bad, and The Ugly. *High-Confidence Computing*, *4*, 100211. https://doi.org/10.1016/j.hcc.2024.100211

Yu, H., Gan, A., Zhang, K., Tong, S., Liu, Q., & Liu, Z. (2024a). Evaluation of Retrieval-Augmented Generation: A Survey. *CoRR*, *abs/2405.07437*. https://arxiv.org/abs/2405.07437

Yu, H., Gan, A., Zhang, K., Tong, S., Liu, Q., & Liu, Z. (2024b). Evaluation of Retrieval-Augmented Generation: A Survey. *CoRR*, *abs/2405.07437*. https://arxiv.org/abs/2405.07437

Anexos

Anexo 1. Arquitectura C3, end to end.



Anexo 2. Set de preguntas para evaluar Métricas técnicas

N°	***Pregunta***
1	¿Qué puesto ocupa Ecuador en el Informe de Competitividad Global 2019 del Foro Económico Mundial?
2	¿Qué porcentaje de los GAD municipales cuentan con una unidad responsable de la competencia de vialidad?
3	¿Cuántos proyectos ejecutaron las juntas parroquiales entre 2016 y 2021 en vialidad?
4	¿Cuánto tiempo tiene el Usuario del Centro de Cumplimiento para confirmar una solicitud resuelta?
5	¿Qué pasa si el usuario crea una solicitud en una categoría incorrecta?
6	¿Cuál es el tamaño máximo permitido para adjuntar archivos en el sistema?
7	¿Qué es la Menor Cuantía?
8	¿Cuánto tiempo tiene el proveedor para enviar la Manifestación de Interés?
9	¿Cuántos parámetros se califican en un procedimiento de Menor Cuantía de Bienes y Servicios?
10	cuales son las razones para mantener un inventario segun el SECAP?
11	cual es la definicion de Bodega para sustancias peligrosas segun MUTUAL de sseguridad
12	Cual es el proposito del codigo de etica de machalilla?
13	a que productos corresponde el código–título 10100000

14	que codigo tienen los Artículos de talabartería y arreos
15	que significa obstaculizacion apra los empleaados?
16	quien es Daniel Lastra
17	¿Qué es el Ojoche
18	¿Cuántos pasajeros transporta mensualmente el tranvía de Tenerife?
19	¿Cuántas paradas tiene el tranvía de Tenerife?
20	¿Cuántas descargas tiene actualmente la App Vía Móvil?
21	¿Cuál es el objetivo principal del Plan de acción Go To Market?
22	¿Qué público objetivo se define en el Plan de acción Go To Market?
23	¿Qué canales de venta se establecen en el Plan de acción Go To Market?
24	¿Cuánto tiempo de garantía tienen los equipos Kärcher para uso doméstico?
25	¿Qué debo presentar para hacer válida la garantía?
26	¿Qué situaciones no están cubiertas por la garantía?
27	¿Qué temperatura debe mantenerse en el refrigerador de medicamentos?
28	¿Qué documentos deben estar disponibles para consulta pública en el almacén farmacéutico?
29	¿Qué señaléticas son obligatorias en la sala de ventas?
30	¿Cuáles son las variables que determinan el tamaño del inventario?
31	¿Qué métodos básicos se usan en el análisis de inventarios?
32	¿Cuál es el objetivo fundamental de la logística de inventarios?
33	¿Quién es el responsable de aplicar el Manual de Tesorería?
34	¿Qué tipos de ingresos se registran en la Tesorería?
35	¿Qué modelo centraliza los recursos públicos para optimizar su administración?
36	¿Qué debilidad se identificó en el control de prevención de daño antijurídico dentro del proceso de gestión contractual?
37	¿Qué riesgo reputacional se asoció al incumplimiento del plan anual de adquisiciones en el proceso de gestión contractual?

38	¿Qué hallazgo se relacionó con la etapa de selección de contratistas en el proceso de gestión contractual?
39	¿Qué principios éticos deben guiar la actuación de los funcionarios del GAD Parroquial Rural de Puerto Machalilla?
40	¿Qué conductas están prohibidas según el Código de Ética y Buena Conducta?
41	¿Cuál es el objetivo principal del Código de Ética y Buena Conducta del GAD Parroquial Rural de Puerto Machalilla?
42	¿Qué establece el contrato del Banco respecto al plazo de vigencia de los productos pasivos y servicios transaccionales para personas jurídicas?
43	¿En qué casos el Banco puede cerrar de manera unilateral las cuentas de una persona jurídica según el contrato marco?
44	¿Qué obligaciones principales asume la persona jurídica al firmar el contrato marco con el Banco ?
45	¿Qué valores corporativos establece el Manual del Colaborador de NOVA Latin America?
46	¿Qué derechos reconoce NOVA a sus colaboradores según el manual?
47	¿Cómo regula el manual las vacaciones de los colaboradores de NOVA?
48	¿Qué garantiza la Política de Reclutamiento y Selección de Personal de Esperanza de Vida ONG en relación con la igualdad de oportunidades?
49	¿Qué requisitos de documentación deben presentar los aspirantes para integrar un expediente interno en Esperanza de Vida ONG?
50	¿Cuál es el plazo máximo establecido por Esperanza de Vida ONG para completar un proceso de reclutamiento externo?

Anexo 3. Código fuente del proyecto.

Todo el código fuente desarrollado ha sido versionado y registrado en un repositorio de Github https://github.com/basaravia/ms-ds-puce2026, a continuación se muestra una tabla resumen de las ramas del proyecto.

Nombre de la rama	***Features clave***	***Descripción***
master	Repositorio base y estabilización	Rama principal, con el agente final y su Web UI
feature/rag-eval-mac	Evaluaciones RAG y limpieza	Evaluaciones del agente RAG y pruebas end to end.
feature/vector-db	Esquemas de Base de Datos	Definición y validación del módulo de la base de datos vectorial kms_agents.

feature/agent-evaluation	Metadatos LLM y Checkpoints LangGraph	Soluciones de recuperación de checkpoints integradas con pandas. Implementación de extracción y captura de metadatos de LLM.
feature/agentic-rag-v2	Búsqueda Híbrida RRF y consultas	Implementación de búsqueda híbrida basada en RRF (Reciprocal Rank Fusion) y refactorización de los componentes del agente RAG. Incluye salidas de ejecución para consultas sobre almacenes de sustancias peligrosas.
feature/agent-rag	Enrutamiento (Routing)	Implementación de análisis de base de conocimiento y enrutamiento (routing) de agentes, usando ingeniería de prompts.
feature/document-analysis	Análisis de dominios	Implementación de análisis de documentos y dominios.
feature/pipeline-batch	RAG Pipeline Core, PGVector, Batch	Implementación central del pipeline de ingesta y transformación para parsing de documentos, indexación y recuperación. Añade PGVector como almacén vectorial , scripts de procesamiento de documentos por lotes/dominio (con conversión y exportación de tablas) y Jupyter notebooks de análisis descriptivo.
feature/pipeline-oneshot	VLM Inferencia (Qwen3-VL/Vertex)	Desarrollo del pipeline versión 1 utilizando Qwen3-VL y Vertex para inferencia VLM. Implementa utilidades para la descripción de imágenes y extracción de tablas, así como procesamiento y exportación por documento.
feature/parser-pii-obfuscator	Ofuscación PII y NER	Implementación de la primera versión del proceso de parsing con Docling, incluye ofuscación de Información de Identificación Personal (PII) usando GLiNER y un pipeline NER en español, junto con salidas estructuradas en markdown e imágenes.
desing/solutions-architecture	Arquitectura de soluciones	Documentación de diseño y actualizaciones de la arquitectura (nivel c1, c2 y c3).