ESCUELA POLITÉCNICA NACIONAL

FACULTAD DE CIENCIAS

INTEGRACIÓN DE MODELOS GENERATIVOS PARA LA RECUPERACIÓN ACADÉMICA

PROYECTO PREVIO A LA OBTENCIÓN DEL TÍTULO DE INGENIERO DE CIENCIAS DE LA COMPUTACION

ALEJANDRO SEBASTIAN CHAVEZ VEGA
chavezalejo85@gmail.com

Director: DRA. GABRIELA SUNTAXI
Gabriela.suntaxi@epn.edu.ec

QUITO, JULIO 2025

DECLARACIÓN

Yo ALEJANDRO SEBASTIAN CHAVEZ VEGA, declaro bajo juramento que el trabajo aquí escrito es de mi autoría; que no ha sido previamente presentado para ningún grado o calificación profesional; y que he consultado las referencias bibliográficas que se incluyen en este documento.

A través de la presente declaración cedo mis derechos de propiedad intelectual, correspondientes a este trabajo, a la Escuela Politécnica Nacional, según lo establecido por la Ley de Propiedad Intelectual, por su reglamento y por la normatividad institucional vigente.

Alejandro Sebastian Chavez Vega

CERTIFICACIÓN

Certifico que el presente trabajo fue desarrollado por ALEJANDRO SEBASTIAN CHAVEZ VEGA, bajo mi supervisión.

Dra. Gabriela Suntaxi
Director del Proyecto

AGRADECIMIENTOS

A todos.

DEDICATORIA

A Georg Ferdinand Ludwig Philipp Cantor,
pues nadie nos expulsará del paraíso que creó para nosotros.

Índice general

Resumen

En el presente trabajo...

Abstract

In this paper...

Capítulo 1

Introducción

1.1. Planteamiento del problema

1.2. Justificación

1.3. Justificación Metodológica

1.4. Objetivos

1.4.1. Objetivo general

Desarrollar e implementar un sistema RAG que mejore el desempeño del buscador de la plataforma Centinela, permitiendo recuperar información científica relevante y generar respuestas automáticas de valor para el usuario.

1.4.2. Objetivos específicos

- Realizar una revisión sistemática de la literatura sobre metodologías y/o frameworks para la implementación de RAG.
- Diseñar e implementar la arquitectura técnica del sistema RAG utilizando modelos de recuperación y generación de texto.
- Evaluar el sistema RAG desarrollado mediante métricas estándar.

1.5. Alcance

1.6. Marco Teórico

1.7. Revision de literatura

En los últimos años, la evolución de los modelos de lenguaje de gran escala (Large Language Models, LLM) han redefinido el procesamiento del lenguaje natural e impulsado nuevas líneas de investigación. Sin embargo, estos modelos dependen únicamente de los datos empleados durante su entrenamiento, lo que limita su capacidad para ofrecer respuestas actualizadas, verificables y contextualizadas. En respuesta a esta limitación surge el enfoque de Retrieval-Augmented Generation (RAG), el cual combina la recuperación de información con la generación de lenguaje natural, logrando mejorar la precisión, la coherencia y la actualidad de las respuestas producidas por los modelos.

Dada la creciente relevancia de los LLM, resulta necesario llevar a cabo una revisión exhaustiva de la literatura que permita consolidar los avances recientes y evaluar los desafíos aún presentes. En esta sección se presenta un análisis estructurado de la literatura disponible, considerando tanto los fundamentos conceptuales de RAG como sus fases de desarrollo, aplicaciones y el futuro. Para ello, el proceso de revisión se organiza en las fases que se presentan a continuación, las cuales buscan garantizar la consistencia, validez y pertinencia de la evidencia obtenida.

1.7.1. Propósito y objetivos de la revisión

El propósito de esta revisión es consolidar la información disponible sobre los RAG, abordando su estudio desde los fundamentos hasta las fases de desarrollo. Se inicia con su definición y arquitectura, para luego profundizar en las etapas clave del proceso: Extracción del corpus, preprocesamiento, vectorización, recuperación de información, evaluación, almacenamiento en bases vectoriales y generación de resultados. Asimismo, se examinan los paradigmas, las métricas de evaluación y el futuro de RAG. Durante esta revision se busca lograr el objetivo general de proporcionar un panorama global y actualizado sobre los RAG, exponiendo sus fundamentos, desarrollo y aplicación.

1.7.2. Criterios de inclusión y exclusión

Se incluyen únicamente revisiones sistemáticas y metaanálisis publicados entre 2018 y 2025, en inglés o español, dado que la producción científica en el área comenzó a incrementarse a partir de 2018, con base en información de Lens.org[1], este incremento coincide con la popularización de los modelos de lenguaje basados en transformers[2]. Los estudios deben provenir de fuentes confiables y ser, a su vez, revisados por un experto. Se da preferencia a aquellos que presenten una cobertura amplia de los temas más relevantes para el objeto de estudio.

Se excluyen las revisiones narrativas, los documentos que carezcan de transparencia en sus métodos de búsqueda o síntesis, así como las publicaciones que no estén directamente relacionadas con el objeto de estudio delimitado.

1.7.3. Identificación del estudio semilla y selección de revisiones relevantes

El proceso de búsqueda se inicia con la identificación de dos estudios semilla, extraídos de Google Scholar mediante los parámetros "retrieval information" y "retrieval augmented generation". Debido al análisis realizado en Lens.org, se estableció el filtro de 2018 a 2025, ya que se observa que a partir de 2018 el término retrieval-augmented generation comenzó a adquirir una relevancia en la literatura científica, mostrando interés de la comunidad investigadora hasta la actualidad.

El primer estudio seleccionado fue Information Retrieval: Recent Advances and Beyond (Hambarde & Proença, 2023), publicado en IEEE Access. Este trabajo constituye una revisión exhaustiva de la recuperación de información, abarcando desde los métodos tradicionales hasta los enfoques basados en deep learning y transformers, por lo que resulta un punto de partida principal para explorar la literatura reciente y relevante.

El segundo estudio semilla corresponde al artículo Retrieval-Augmented Generation for Large Language Models (Gao, Xiong, Gao, Jia, Pan, Bi, Dai, Sun & Wang, 2023), publicado en arXiv, el cual presenta un marco conceptual y aplicado sobre la integración de recuperación de información y modelos generativos de gran escala. Su incorporación permite establecer una base teórica para contextualizar el análisis

[1] Es una plataforma abierta para la búsqueda, análisis y visualización de literatura científica y patentes. Accesible en: Lens.org

[2] Se atribuye a hitos como BERT (2018), GPT-2 (2019) y T5 (2020), que impulsaron un avance en la investigación del procesamiento del Lenguaje Natural

de las revisiones seleccionadas.

A partir de estos dos estudios semilla, y aplicando los criterios de inclusión y exclusión previamente definidos, se identificaron 25 revisiones relevantes que cumplen con los criterios establecidos. Estas revisiones constituyen la base para el análisis y síntesis en el presente trabajo.

1.7.4. Valoracion de la evidencias y extracion de la infomacion

De los estudios seleccionados se procede a realizar un análisis, con el fin de excluir aquellos artículos que no cumplen con los criterios establecidos o que presentan un nivel de profundidad insuficiente para los objetivos de la revisión. La selección final de los estudios se realiza en consenso con expertos en el área, garantizando así la pertinencia y relevancia de la evidencia incluida.Para la organización, codificación y síntesis de la información se usa ATLAS.ti [3] que facilitará la estructuración de los hallazgos.

1.7.5. Síntesis y representación de resultados

Con la literatura seleccionada se identificó la hoja de ruta que se presenta a continuación en la Figura 1.1.

[3]Scientific Software Development GmbH. Disponible en: Atlas.ti

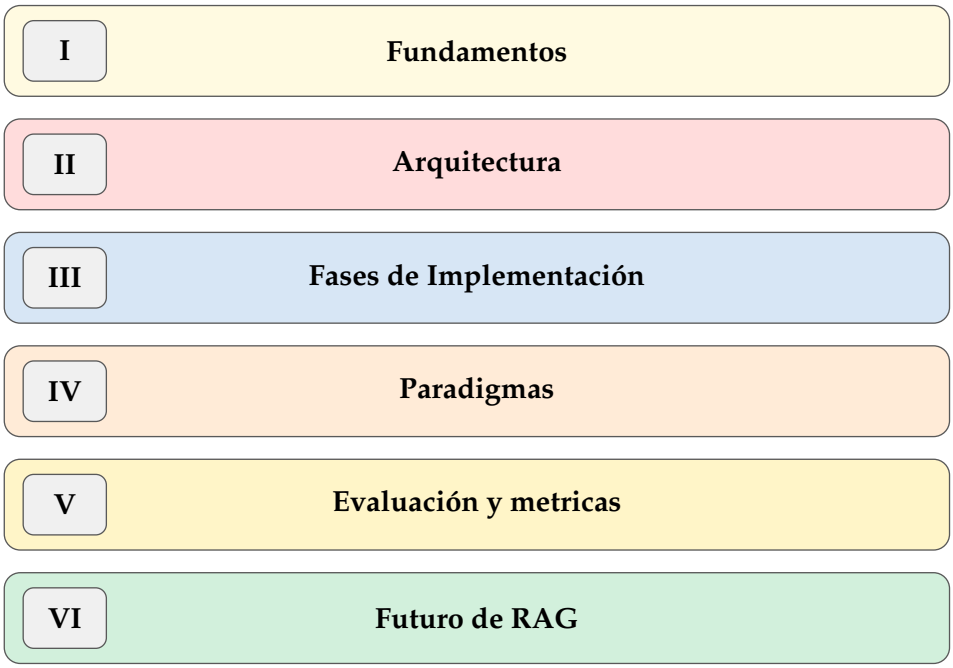

Figura 1.1: Resumen esquemático de RAG

A partir de esta hoja de ruta se desarrolla un esquema más detallado, en el que se expone primero exploraremos su teoría, características y aplicaciones, como se muestra en la Fig 1.2. Luego profundizamos en su arquitectura en la cual se describe cada uno de los componentes que lo conforman (retriever, augmented y generation) y las variantes y mejoras que existen de cada uno. Posteriormente, se detalla el proceso de implementación, desde la preparación de datos hasta el componente de generación, incluyendo las técnicas y herramientas más relevantes. A continuación en la Fig , se examinan los paradigmas de RAG, dando a conocer los tipos de paradigmas y sus clases, para luego en la Fig , se presentan las métricas y evaluadores automáticos utilizados en la evaluación de sistemas RAG, así como las consideraciones éticas y de equidad que deben tenerse en cuenta. Finalmente, se discuten las tendencias emergentes, los desafíos que actualmente se tienen y futuras direcciones que podrían tomar los sistemas RAG.

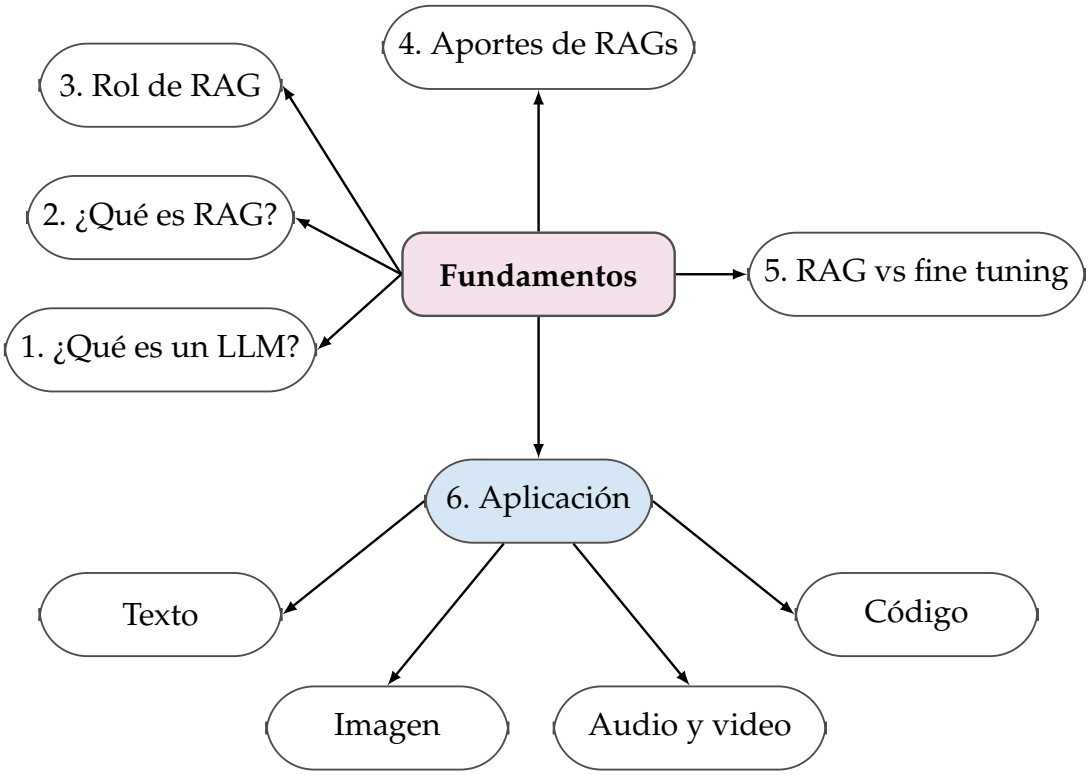

Figura 1.2: Fundamentos de RAG

Fundamentos

En esta subsección se presentan los fundamentos teóricos de Retrieval Augmented Generation (RAG), comenzando con la definición de los modelos de lenguaje de gran escala (LLMs) y su relación. Se expone también el papel que desempeña RAG, los principales aportes que ha generado en distintos ámbitos y su diferenciación frente al fine tuning. Finalmente, se introduce su aplicación práctica, lo que permite comprender la importancia y el impacto que RAG tiene en la actualidad.

Que es un LLM Son modelos de inteligencia artificial (IA) basados en la arquitectura transformer, entrenados con grandes volumnes de datos textuales con el objetivo de aprender representaciones contextuales del lenguaje. Según Casola, Lauriola y Lavelli [3], estos modelos utilizan técnicas de pre-entrenamiento no supervisado para captar patrones lingüísticos y semánticos, lo que permite que posteriormente puedan ajustarse a tareas específicas como clasificación de texto, análisis de sentimientos, traducción automática, reconocimiento de entidades o respuesta a preguntas. Ejemplos destacados son *BERT, RoBERTa, ALBERT, XLNet, DistilBERT y GPT-3*, que han mostrado rendimientos sobresalientes en diversas aplicaciones de procesamiento de

lenguaje natural (NLP).

De acuerdo con Ramdurai [9], los LLMs también se definen como una clase de modelos de IA capaces de procesar y generar texto de forma similar al lenguaje humano, gracias al uso de redes neuronales profundas y la capacidad de aprender no solo gramática y relaciones entre palabras, sino también aspectos más complejos como humor, tono emocional y contexto. Entrenados en enormes corpus de datos provenientes de libros, artículos y sitios web, estos modelos pueden responder preguntas, redactar ensayos, traducir, resumir y crear contenido de manera autónoma. Ejemplos recientes incluyen *GPT-4, T5, XLNet y PaLM*, los cuales demuestran su versatilidad en tareas avanzadas de NLP y en sistemas aplicados en diferentes industrias.

Que es un RAG Según Han, Susnjak y Mathrani [6], Retrieval-Augmented Generation (RAG) es una técnica que integra la capacidad generativa de los modelos de lenguaje con la precisión de la recuperación de información en tiempo real. En lugar de basarse únicamente en el conocimiento almacenado en los parámetros durante el entrenamiento, RAG permite consultar repositorios externos como bases de datos o motores de búsqueda para obtener documentos relevantes y actualizados. Estos se incorporan al prompt del usuario, lo que fundamenta la respuesta en fuentes verificables y disminuye los problemas de errores y alucinaciones que suelen presentarse en los modelos de lenguaje de gran escala.

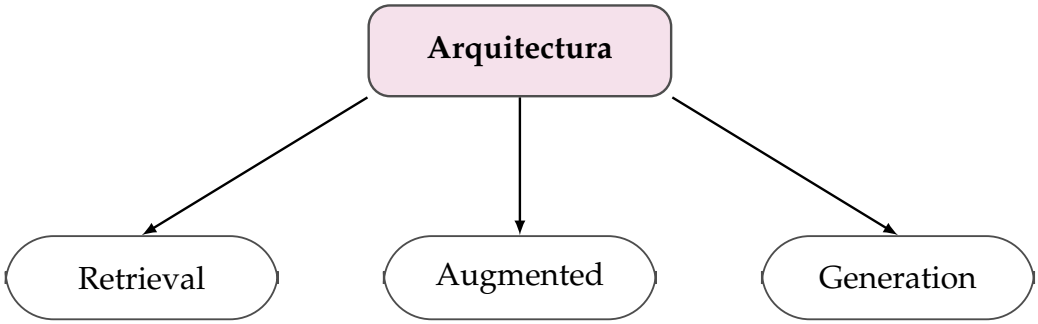

Figura 1.3: Componentes de RAG

Arquitectura

RAG se compone de tres fases: recuperación, augmentación y generación. Como lo menciona Gao et al. [4], primero se localizan documentos relevantes para la consulta; luego, se enriquece la entrada del usuario con esos textos; finalmente, el

modelo produce una respuesta basada tanto en su conocimiento interno como en la información recuperada. Gracias a este enfoque, RAG incrementa la exactitud de las respuestas, facilita la actualización del conocimiento sin necesidad de reentrenar el modelo y mejora la transparencia al permitir la cita de fuentes. Es una de las técnicas más relevantes en tareas que requieren gran cantidad de conocimiento, como el ámbito médico, legal o de investigación científica.

Categoría	**Subcategoría / Tipo**	**Técnicas**
Indexing	Vector DB	StepBack Prompt
	Algoritmos de búsqueda	Approximate Nearest Neighbors, Locality-Sensitive Hashing
Enhancements	Reranking	HyDE
	Retriever Finetuning	-
	Hybrid Search	-
	Chunk Optimization	-
	Recursive Retrieval	-
Tipos de retrievers	Sparse	BM25, TF-IDF
	Dense	Embeddings (ej. BERT, OpenAI, etc.)
	Others	Modelos híbridos u otros

Cuadro 1.1: Componente Retriever

Categoría	**Subcategoría / Tipo**	**Técnicas**
Tipos	Pre-training	-
	Fine-tuning	-
	Inference	-
Data	Structured	-
	Unstructured	-
	LLM generated content	-
Process	Once	-
	Iterative	-
	Adaptative	-

Cuadro 1.2: Componente Augmentation

Categoría	Subcategoría / Tipo	Ejemplos
Tipos	Transformers	GPT, BART, T5
	LSTM	Modelos secuencia a secuencia tradicionales
	GANs	Generación adversarial en imágenes y texto
	Diffusion Models	Imagen, audio y video
Enhancements	Prompt Engineering	Diseño de instrucciones, chain-of-thought, step-back prompts
	Generator Fine-tuning	Ajuste del modelo al dominio específico
	Decoding Tuning	Beam search, nucleus sampling, top-k sampling

Cuadro 1.3: Componente Generator

Paradigmas

Tipos de Paradigmas

Clases

Evaluacion

Metricas

Evaluadores automaticos

Futuro de RAG

Desafios Actuales

Direcciones potenciales

Perspectivas

Capítulo 2

Metodología

2.1. Revisión sistemática

Umbrella Review, según los lineamientos del Instituto Joanna Briggs (JBI), es un tipo de revisión sistemática que recopila y analiza evidencia secundaria, es decir, revisiones sistemáticas y metaanálisis ya publicados. Su propósito es consolidar el conocimiento disponible, identificar coincidencias y contradicciones en la literatura existente, así como señalar vacíos de evidencia. Para ello, requiere la elaboración de un protocolo previo que establezca criterios de inclusión y exclusión, estrategias de búsqueda y métodos de síntesis, garantizando un proceso transparente y riguroso.

Por otra parte, la estrategia de propagación de citaciones (Back-and-Forward Citation Propagating) complementa este enfoque al permitir encontrar dinámicamente la literatura. A través de la propagación de citaciones se amplía y actualiza la literatura encontrada en las bases de datos tradicionales. De este modo, se superan limitaciones como la indexación incompleta, las variaciones en el uso de palabras clave o la exclusión de ciertas publicaciones.

Metodología: Umbrella Review con Propagación de Citaciones

Como parte de la metodología Umbrella Review es necesario establecer un protocolo para ejecutar la revisión. Se han considerado las siguientes fases para dicho protocolo:

1. **Propósito de la revisión**

   La revisión se justifica en la necesidad de consolidar evidencia secundaria de

calidad, aprovechando el enfoque de propagación de citaciones para garantizar una búsqueda amplia, estructurada y actualizada.

2. **Objetivos específicos**

   Se definen los objetivos generales y específicos que guiarán la identificación de literatura mediante la propagación de citaciones, así como el proceso de síntesis de resultados.

3. **Criterios de inclusión y exclusión** Se definen de manera general como la incorporación de revisiones y metaanálisis que sean pertinentes, de calidad y relacionados con el tema de estudio, y la exclusión de aquellos trabajos que no cumplan con estos requisitos de relevancia.

4. **Identificación del estudio semilla y propagación de citaciones**

   La búsqueda se inicia en bases de datos académicas como *Scopus*, *Web of Science*, *IEEE Xplore* o *Google Scholar*, a fin de localizar un estudio semilla (revisión o resumen amplio) que ofrezca una cobertura representativa del tema. A partir de este estudio, se aplica la estrategia de Back-and-Forward Citation Propagation, que combina:

   - *Backward citation:* revisión de las referencias citadas en el estudio semilla.
   - *Forward citation:* identificación de trabajos más recientes que citan al estudio semilla.

   De este modo, el corpus de literatura se amplía progresivamente hasta alcanzar un punto de saturación en el que la propagación deja de aportar nueva evidencia relevante.

5. **Selección de revisiones relevantes**

   A partir de la propagación de citaciones, se aplican los criterios de inclusión - exclusión para determinar qué revisiones serán incorporadas al análisis.

6. **Valoración de la calidad de la evidencia**

   La calidad de los estudios se evalúa según los criterios definidos, garantizando su consistencia al tema de estudio. Para apoyar este proceso se usa una herramienta de análisis que facilite la organización y valoración sistemática de la evidencia.

7. **Extracción de información clave**

   De cada revisión seleccionada se extraerán datos esenciales, organizados en una tabla de extracción que incluirá:

   - Autor y año de publicación
   - Objetivo del estudio
   - Tipo de revisión
   - Número de estudios primarios incluidos
   - Principales hallazgos
   - Conclusiones generales
   - Limitaciones reportadas

8. **Síntesis y representación de resultados**

   Los hallazgos se organizarán en dos niveles complementarios:

   - **Tabular:** tablas comparativas de las revisiones incluidas.
   - **Narrativo:** síntesis descriptiva de los principales hallazgos.
   - **Temático y visual:** mapas de evidencia y esquemas que reflejen la propagación de citaciones, mostrando las conexiones entre estudios clave.

9. **Discusión y conclusiones**

   Los resultados se interpretan desde una perspectiva crítica, destacando fortalezas, limitaciones y la evolución de la evidencia en el tiempo. Se identifican coincidencias y divergencias entre revisiones, así como vacíos de conocimiento, y se proponen líneas de investigación futura.

En esta metodología, el Umbrella Review se utiliza como marco general para sintetizar evidencia secundaria a partir de revisiones de exhaustivas de la literatura, complementándose con la propagación de citaciones para integrar aportes recientes y reflejar la evolución del conocimiento disponible.

2.2. Enfoque Design Science Research (DSR)

De acuerdo con vom Brocke et al. Brocke, Hevner y Maedche [2], Design Science Research, desarrollada en 1969, es un paradigma de resolución de problemas que

busca mejorar el conocimiento humano mediante la creación de artefactos innovadores. En otras palabras, es una metodología que crea soluciones a problemas reales y, al mismo tiempo, genera conocimiento útil y aplicable sobre cómo diseñar estas soluciones. Las etapas que se aplicarán en el presente trabajo son las siguientes:

- **Identificación del problema y motivación** En esta etapa se precisa el problema y se justifica por qué es necesaria una solución. De acuerdo con Peffers et al. (2008), esta etapa exige analizar el problema en detalle, descomponiéndolo en sus partes clave para identificar sus causas, efectos y alcance. Además, es crucial justificar la relevancia del problema, tanto desde una perspectiva teórica (es decir, cómo contribuye al conocimiento académico) como desde una perspectiva práctica (cómo afecta a organizaciones, usuarios o sistemas reales). También implica explorar la literatura para verificar que el problema es relevante, desafiante y nuevo, lo que permite definir los límites del proyecto de investigación.
- **Definir los objetivos para la solución** Se plantean los criterios que debe cumplir una solución exitosa basándose en el conocimiento existente y en la factibilidad técnica y organizacional. Los objetivos deberán permitir construir algo efectivo y deseable, no solamente desde el ámbito académico sino también en el entorno en que se aplicará. Estos pueden expresarse en términos cualitativos o cuantitativos; el investigador establece aquí la meta hacia donde se dirigirá el artefacto.
- **Diseño y desarrollo del artefacto** En esta etapa se construye una solución concreta, como un modelo, software o sistema, que responde directamente a los objetivos planteados. Para ello, se utiliza el conocimiento existente que fundamenta las decisiones del diseño y la estructura del artefacto. No solo se trata de crear algo, sino de asegurar que pueda ser comprendido, evaluado y replicado por otros.
- **Demostración del uso del artefacto para resolver el problema** Se muestra cómo se usa el artefacto en un escenario real o simulado. Esta demostración no valida científicamente su efectividad, sino que muestra su aplicabilidad, evidenciando que el artefacto propuesto puede operar de forma efectiva. Por su parte, vom Brocke et al. (2020) destacan que esta etapa es fundamental para conectar el diseño teórico con la realidad del usuario o del entorno organizacional, permitiendo detectar oportunidades de mejora antes de una evaluación rigurosa.

- **Evaluación del desempeño del artefacto** Se busca medir su efectividad, eficiencia e impacto, aportando evidencia que justifique su valor y utilidad. Además, según vom Brocke et al. (2020), esta puede asumirse de forma continua mediante una evaluación formativa que permita ciclos iterativos de rediseño y mejora a lo largo del proceso de investigación.
- **Comunicación de los resultados al público académico y profesional** Finalmente, esta etapa consiste en difundir de forma clara los resultados del diseño y de la investigación realizada.

Estos pasos están basados en el modelo clásico de DSR de Peffers (2008), que vom Brocke adapta y expande en su guía.

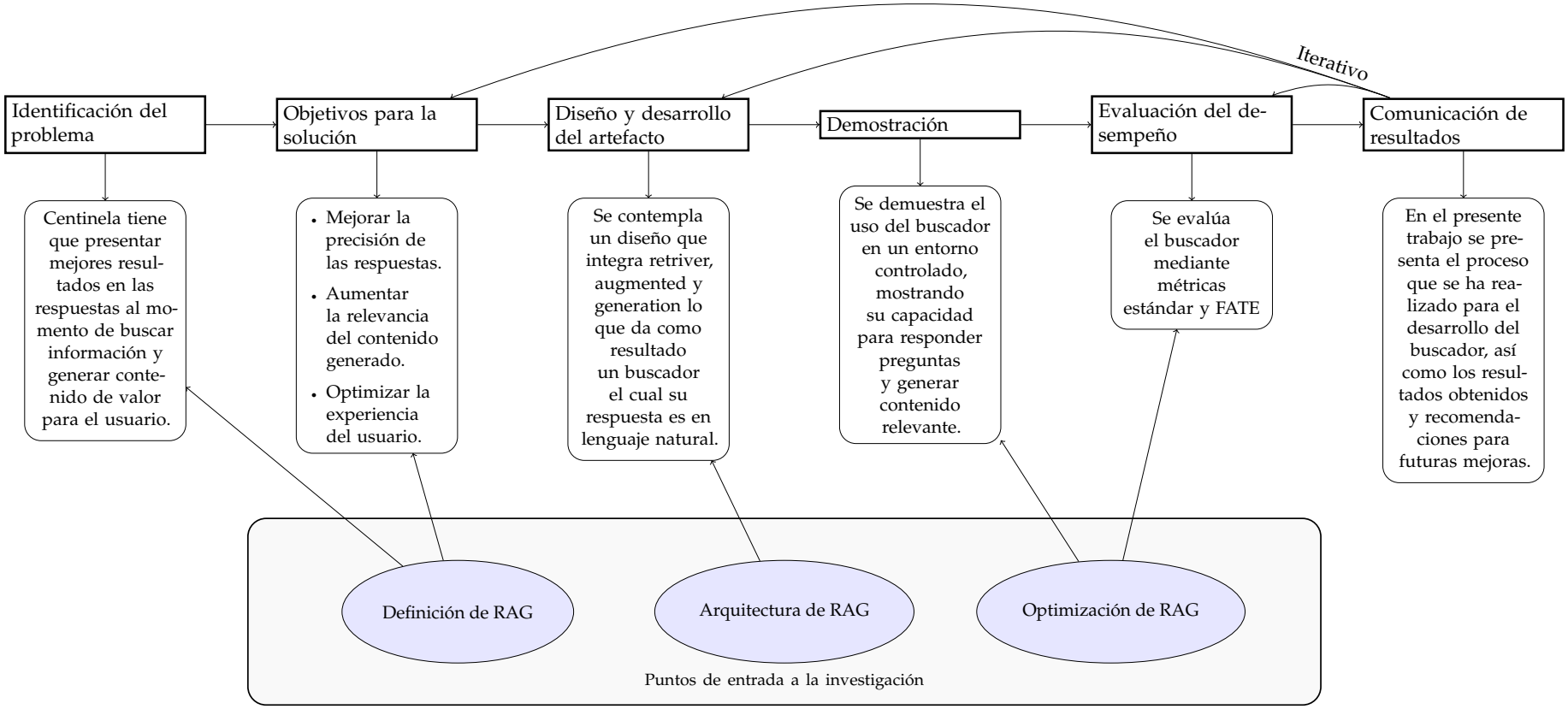

Figura 2.1: Proceso de Diseño de Investigación para el desarrollo de RAG en Centinela

2.3. Diseño y desarrollo del artefacto

Capítulo 3

Pruebas, Resultados, Conclusiones y Recomendaciones

3.1. Demostracion

3.2. Evaluacion del desempeño

3.3. Resultados

3.4. Conclusiones

3.5. Recomendaciones

Bibliografía

[1] Edoardo Aromataris et al. *Methodology for JBI Umbrella Reviews*. Disponible en Research Online de la University of Wollongong. Adelaide: Joanna Briggs Institute, 2014. URL: https://ro.uow.edu.au/smhpapers/3344.

[2] Jan vom Brocke, Alan Hevner y Alexander Maedche. «Introduction to Design Science Research». En: *Design Science Research. Cases*. Ed. por Jan vom Brocke, Alan Hevner y Alexander Maedche. Cham: Springer, 2020, págs. 1-13. DOI: 10.1007/978-3-030-46781-4_1.

[3] Silvia Casola, Ivano Lauriola y Alberto Lavelli. «Pre-trained transformers: An empirical comparison». En: *Machine Learning with Applications* 9 (2022). Acceso abierto, CC BY 4.0, pág. 100334. DOI: 10.1016/j.mlwa.2022.100334.

[4] Yunfan Gao et al. *Retrieval Augmented Generation for Large Language Models: A Survey*. 2023. arXiv: 2312.10997 [cs.CL]. URL: https://arxiv.org/abs/2312.10997.

[5] Kailash A. Hambarde y Hugo Proença. «Information Retrieval: Recent Advances and Beyond». En: *IEEE Access* 11 (2023), pág. 76581 76620. DOI: 10.1109/ACCESS.2023.3295776.

[6] Binglan Han, Teo Susnjak y Anuradha Mathrani. «Automating Systematic Literature Reviews with Retrieval Augmented Generation: A Comprehensive Overview». En: *Applied Sciences* 14.19 (2024), pág. 9103. DOI: 10.3390/app14199103.

[7] Stefania Papatheodorou. «Umbrella reviews: what they are and why we need them». En: *European Journal of Epidemiology* 34.6 (2019), págs. 543-546. DOI: 10.1007/s10654-019-00505-6.

[8] Ken Peffers et al. «A design science research methodology for information systems research». En: *Journal of Management Information Systems* 24.3 (2008), pág. 45 77. DOI: 10.2753/MIS0742-1222240302. URL: https://doi.org/10.2753/MIS0742-1222240302.

[9] Balagopal Ramdurai. «Large Language Models (LLMs), Retrieval-Augmented Generation (RAG) systems, and Convolutional Neural Networks (CNNs) in Application systems». En: *International Journal of Marketing and Technology* 15.01 (2025). Disponible en acceso abierto en ResearchGate. URL: https://www.researchgate.net/publication/387128512.