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

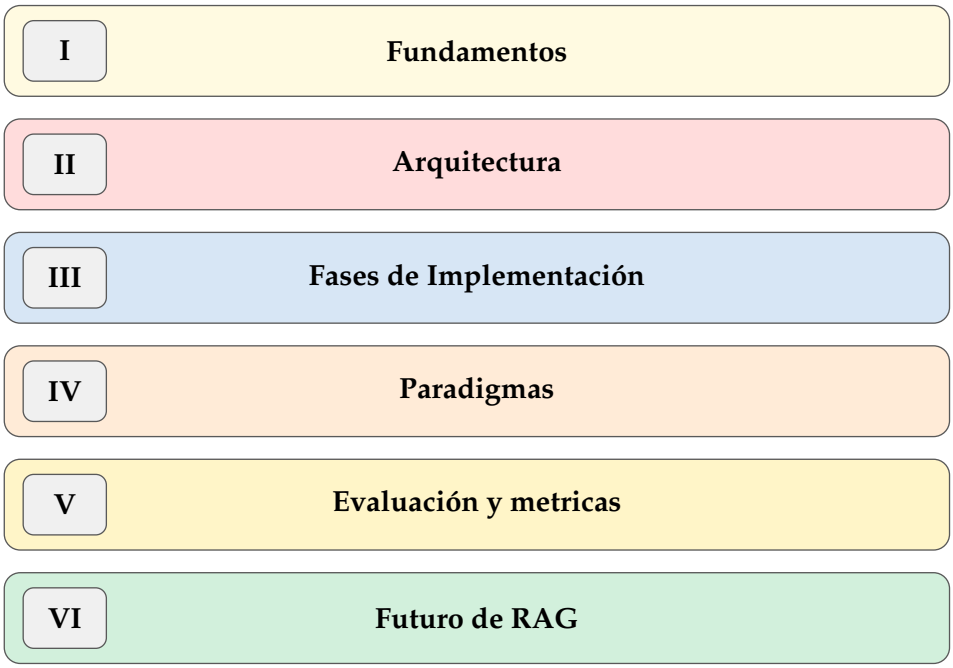

Figura 1.1: Resumen esquemático de RAG

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.

Fundamentos

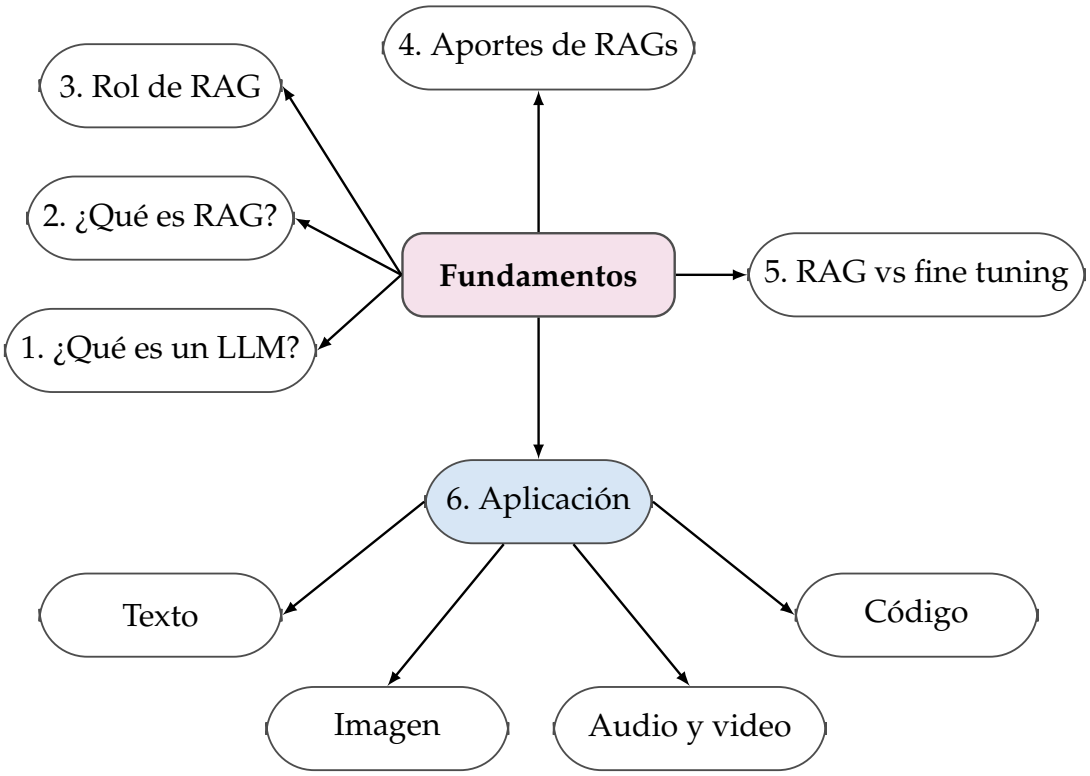

Figura 1.2: Fundamentos de RAG

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 [4], 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 [19], 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 [9], 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.

Arquitectura

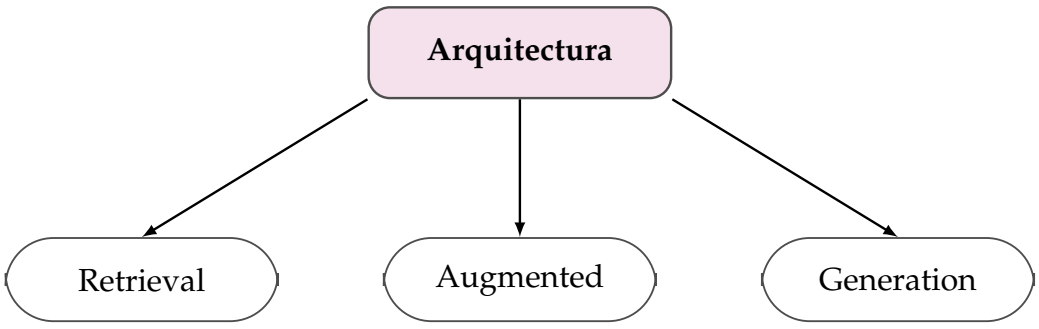

Figura 1.3: Componentes de RAG

RAG se compone de tres fases: recuperación, augmentation y generación (ver Figura 1.3). Como lo menciona Gao et al. [7], 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	FAISS, Annoy, Milvus, Weaviate, Pinecone, Qdrant
	Algoritmos de búsqueda	Approximate Nearest Neighbors, Locality-Sensitive Hashing
Enhancements	Reranking	Re2G, AceCoder, XRICL, monoT5
	Retriever Finetuning	REPLUG, APICoder, EDITSUM
	Hybrid Search	RAP-Gen, BlendedRAG, ReACC
	Chunk Optimization	RAPTOR, LlamaIndex
	Recursive Retrieval	ReAct, RATP
	Query Reformulation	HyDE
Tipos de retrievers	Sparse	BM25, TF-IDF
	Dense	Embeddings (ej. BERT, OpenAI, etc.)
	Others	Modelos híbridos u otros

Tabla 1.1: Componente Retriever

Indexing es encargado de organizar y representar la informacion de forma eficiente. En este sentido los vector databases (VDBs) han aumentado su popularidad ya que permiten alamacenar vectores de alta dimensionalidad y permiten realizar busquedas por similitud semantica. Como explica Joshi [13] estas bases de datos resultan esenciales para en aplicacion de inteligencia artificial generativa, ya que superan las limitacion de las bases realcionales en el manejo de datos no estructurados. Los VDBs integran mecanismos de busqueda aproximada de vecinos mas cercanos (ANNS) lo que posibilita consultas rapidas incluso sobre miles de objetos. Ademas, como lo menciona Ma et al. [15] incorporan técnicas de optimización como particionamiento, sharding, cachés y replicación para garantizar escalabilidad y baja

latencia en entornos distribuidos Locality-Sensitive Hashing (LSH) es una técnica de indexación ampliamente utilizada en VDBs para acelerar la búsqueda de vecinos aproximados en espacios de alta dimensionalidad. A diferencia de los esquemas de hashing tradicionales, cuyo objetivo es dispersar uniformemente los datos para minimizar colisiones, LSH está diseñado para maximizar la probabilidad de que vectores similares se asignen al mismo bucket de hash [15]. Según esta técnica, la preservación de la localidad se consigue mediante funciones de hash que reflejan la similitud entre vectores en colisiones más frecuentes dentro del espacio reducido. De esta manera, al realizar una consulta, el sistema solo necesita comparar el vector de entrada con aquellos almacenados en el mismo o en buckets cercanos, reduciendo drásticamente la complejidad computacional de la búsqueda [10]. Estas características explican por qué los VDBs se han consolidado como una infraestructura fundamental en el soporte de sistemas RAG.

Categoría	**Subcategoría / Tipo**	**Técnicas / Ejemplos**
Tipos	Pre-training	knowledge graph embeddings
	Fine-tuning	REPLUG,UPRISE
	Inference	Retrieve-then-Read-then-Revise (RARR), Lost in the Middle
Data	Structured	Knowledge Graphs
	Unstructured	open corpus (Common Crawl, PubMed, etc.)
	LLM generated content	self-retrieval
Process	Once	document augmentation
	Iterative	pseudo-relevance feedback (PRF)
	Adaptative	SKR (Selective Knowledge Retrieval)

Tabla 1.2: Componente Augmentation

Augmentation es el proceso el cual un modelo de lenguaje incorpora información adicional ya sea externa, como documentos, bases de conocimiento o corpus abiertos, o procesada internamente en diferentes etapas de su funcionamiento. Los autores coinciden en que esta integración cumple objetivos clave: mejorar la precisión de las respuestas al aportar evidencia relevante, reducir las alucinaciones al contrastar el conocimiento implícito del modelo con fuentes verificables, y actualizar el conocimiento de los LLMs sin necesidad de reentrenarlos desde cero, ya que el acceso a

información recuperada permite mantenerlos al día en dominios dinámicos como ciencia, medicina o derecho.

Los tipos de augmentation, Zhao et al. [24] señalan que se puede aplicar en 3 momentos distintos. Durante el pre-training, se integran representaciones estructuradas como knowledge graph embeddings que dotan al modelo de memoria explícita sobre entidades y relaciones. En la etapa de fine-tuning, técnicas como REPLUG y UPRISE permiten alinear mejor el recuperador y el generador en dominios específicos. Finalmente, en la inferencia, se aplican métodos sin reentrenamiento: por ejemplo, RARR (Retrieve-then-Read-then-Revise) refina las respuestas con evidencia recuperada, mientras que la mitigación de Lost in the Middle reorganiza documentos recuperados para que el modelo aproveche mejor la ventana de contexto.

Con respecto a los datos, Fan et al. [5] detancan tres clases. La información estructurada, como los Knowledge Graphs, es fundamental para tareas de razonamiento factual ya que permite modelar entidades y relaciones explícitas. La información no estructurada, como corpus abiertos (Common Crawl, PubMed, Wikipedia), se ha vuelto estándar en open-domain QA. Tal como explican Gao et al. [7] estos corpus aportan amplitud temática, pero también requieren mecanismos de filtrado, chunking y re-ranking para evitar ruido y mitigar el problema de Lost in the Middle. Finalmente, surge la categoría de contenido generado por LLM, donde el propio modelo actúa como fuente en esquemas de self-retrieval, generando y reutilizando conocimiento de manera autónoma. En este caso, los LLMs generan documentos intermedios, hipótesis o representaciones que se utilizan posteriormente como consultas o evidencia emplean un módulo crítico que evalúa si es necesario recuperar información externa o si el propio contenido generado basta para resolver la tarea. Gracias a este enfoque , como subraya Fan et al. [5] abre la posibilidad de que los modelos se autocomplementen y reutilicen su conocimiento previo sin depender exclusivamente de bases externas.

Los procesos de *augmentation* se los clasifica en tres modalidades según Zhao et al. [24]. Augmentation puede aplicarse una sola vez (Once), como en el caso de la document augmentation, donde se enriquece directamente la entrada con información adicional antes de la generación. Otra modalidad es la iterativa, que emplea técnicas como el pseudo-relevance feedback (PRF), mediante el cual la consulta inicial se reformula a partir de los resultados recuperados, repitiendo el ciclo para refinar la relevancia. La modalidad adaptativa es cuando el sistema decide dinámicamente si conviene recuperar información o no. Un ejemplo de esto es Selective Knowledge Retrieval (SKR), que evita búsquedas innecesarias cuando el modelo ya posee el

conocimiento suficiente, reduciendo costes y minimizando la incorporación de ruido.

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

Tabla 1.3: Componente Generator

El componente generador es el encargado de producir texto,imagenes u otro tipo de contenido. La capacidad del generador no depende unicamente de su arquitectura sino de un conjunto de estrategias que optmizan su rendimiento y controlan la calidad de las salidas. Estas mejoras incluyen prompt engineering, el fine-tuning especializado en dominios y ajustes en los métodos de decodificación para equilibrar coherencia y diversidad en la generación. En este sentido, comprender tanto los tipos de modelos como las técnicas de optimización resulta fundamental para evaluar el papel del *Generator* en sistemas avanzados como los de Retrieval-Augmented Generation (RAG), donde la combinación de arquitectura y optimización garantiza la generación de respuestas más fiables, contextualizadas y relevantes [4, 5, 23]

Los transformes son la arquitectura mas influyente en NLP ya que, según lo menciona Casola, Lauriola y Lavelli [4], su capacidad de manejar dependencias a largo plazo mediante mecanismos de auto-atención ha permitido el desarrollo de modelos como GPT, BART y T5, los cuales han superado ampliamente a enfoques previos y marcado un cambio de paradigma en la generación de lenguaje. os LSTM se empleaban en arquitecturas secuencia a secuencia, resolviendo problemas de memoria en redes recurrentes, aunque presentaban limitaciones en el escalamiento y en la captura de dependencias largas [12]. Por otro lado, introdujeron un enfoque basado en el enfrentamiento entre un generador y un discriminador, logrando avances en la creación de imágenes y, posteriormente, en la generación de texto. Mas recientemete los Diffusion Models se han consolidado en el ámbito multimodal (imagen, audio y video), gracias a su capacidad de producir datos de alta calidad a partir de procesos

de ruido inverso, ampliando las fronteras de la generación más allá del texto.

Con respecto a la mejora, el Prompt Engineering ha demostrado ser una técnica central para guiar los modelos hacia respuestas más precisas y controladas. Segun Zhang y Zhang [23] incluye el diseño de instrucciones específicas, así como métodos como chain of thought o step-back prompting, que mejoran la coherencia y reducen alucinaciones en los resultados Otra estrategia es el fine tuning que consiste en adaptar un modelo a un modelo concreto, optimizando su capacidad de manejar informacion especial especializada y aumentando su fiabilidad en tareas críticas, por ejemplo en contextos científicos o legales El Decoding Tuning se refiere al ajuste de los métodos utilizados por un modelo generativo para decidir qué palabra o token producir a continuación. En lugar de limitarse a elegir siempre la opción con mayor probabilidad, lo cual puede generar textos repetitivos y poco naturales, se aplican estrategias que permiten modular el balance entre coherencia y diversidad. Entre ellas, el beam search explora varias rutas posibles de generación en paralelo para seleccionar la más prometedora, garantizando coherencia aunque sacrificando creatividad. Por su parte, el nucleus sampling restringe las opciones a un conjunto dinámico de palabras que concentran la mayor parte de la probabilidad acumulada y selecciona una de ellas de manera aleatoria, lo que produce textos más naturales y variados. Finalmente, el top-k sampling limita las elecciones a las k palabras más probables y elige una de ellas de acuerdo con su distribución de probabilidad, lo que ofrece un equilibrio controlado entre precisión y diversidad.

> Estas estrategias permiten modular el equilibrio entre coherencia y diversidad en la generación de texto, adaptando el comportamiento del modelo sin necesidad de modificar su arquitectura [11].

Fases de Implementación

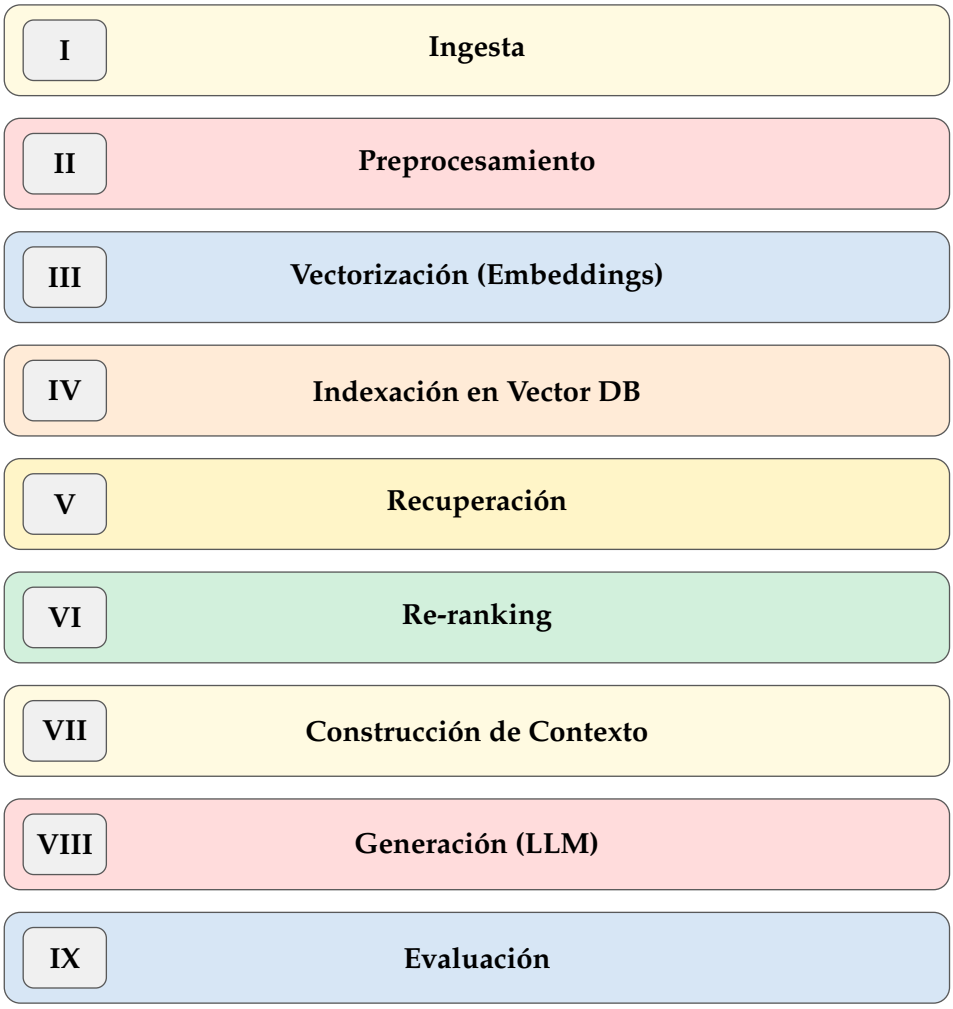

Figura 1.4: Pipeline de RAG

Las fases de implementacion son una secuencia de pasos diseñados para enriquecer la generación de respuestas con información externa y actualizada. Según Tabassum y Patil [21], todo inicia con la ingesta y el preprocesamiento de los datos, donde las técnicas de limpieza y segmentación garantizan que el texto sea utilizable para fases posteriores. Luego se vectoriza mediante embeddings los cuales transforman los fragmentos de texto en representaciones numericas para la recuperacion semantica tal como lo expresa Minaee et al. [16]. Estos vectores se almacenan en una base de datos vectorial que en conjunto con inidice tradicionales como bm25 permiten la recuperacion hibrida.[11]. Con los datos obtenidos, como lo señala Sarthi et al. [20], es necesario aplicar procesos de re-ranking y filtrado que permiten priorizar las partes más relevantes dando lugar la construcción de contexto, en la que

los fragmentos de texto seleccionados se organizan, se condensan para reducir su extensión y se adaptan al límite de entrada del modelo.para conformar un contexto optimizado y manejable que servirá de entrada al modelo generativo. En la fase final, Knollmeyer et al. [14] sostiene que el LLM debe complementarse con mecanismos de evaluación para aegurar la fidelidad y coherencia de las respuestas generadas.

Paradigmas

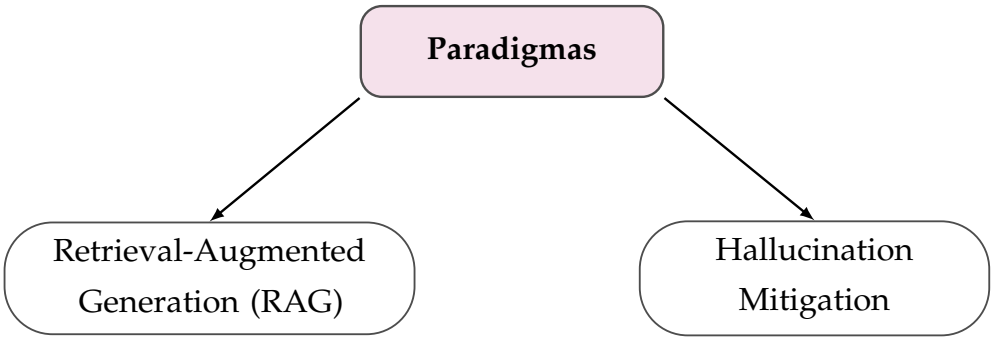

Figura 1.5: Paradigmas principales en RAG

Dentro del area de Retrieval Augmented Generation (RAG) Gao et al. [7] distiguen los paradigmas Naive RAG, Advanced RAG y Modular RAG, que representan un progreso desde enfoques básicos de recuperación y generación hasta arquitecturas modulares y flexibles. A su vez Zhang y Zhang [23] clasifican los problemas, en el area de Hallucination Mitigation, en dos ejes: retrieval failure (fallos en fuentes, consultas o recuperadores) y generation deficiency (ruido o conflicto contextual y límites de capacidad). FInalmente, Zhao et al. [24] proponen fundaciones de RAG segun la forma que el retriver complementa al generador se clasifican en Query-based RAG, Latent Representation-based RAG, Logit-based RAG y Speculative RAG, de esta forma se expande la aplicación de RAG a múltiples dominios y modalidades.

Tipos de Paradigmas

- **Retrieval-Augmented Generation (RAG) — Gao et al. (2023)**
    - Naive RAG: enfoque básico de recuperación y generación.
    - Advanced RAG: incorpora optimizaciones como segmentación fina, re-ranking y recuperación iterativa.

    - Modular RAG: paradigma flexible con módulos especializados para búsqueda, memoria, alineación y validación.

- **Tipos de RAG — Zhao et al. (2024)**
    - Query-based RAG: integra directamente la consulta y la información recuperada en el input del generador.
    - Latent Representation-based RAG: incorpora la información recuperada como representaciones latentes en el modelo generativo.
    - Logit-based RAG: combina la información de recuperación en la fase de decodificación a nivel de logits.
    - Speculative RAG: sustituye pasos de generación por recuperación para acelerar y reducir costes.

- **Hallucination Mitigation — Zhang & Zhang (2025)**
    - Retrieval failure: fallos en las fuentes, consultas, recuperadores o estrategias de recuperación.
    - Generation deficiency: deficiencias en la generación como ruido, conflictos contextuales, middle curse, problemas de alineación y límites de capacidad.

Evaluacion y Metricas

La evaluación de los sistemas de Recuperación de Información (IR) y de Retrieval-Augmented Generation (RAG) ha sido objeto de un creciente interés en la literatura reciente. En el ámbito de IR, Bernard y Balog [2] destacan que las nociones de equidad, transparencia y responsabilidad requieren enfoques diversos: mientras la *fairness* se mide mayormente con métricas automáticas, la *transparency* y la *accountability* suelen evaluarse mediante auditorías y estudios con usuarios. En contraste, la dimensión ética carece de métricas claras y se asocia más a aspectos como privacidad y seguridad.

En el área de RAG, Knollmeyer et al. [14] identifican cinco dimensiones clave de evaluación: relevancia del contexto, fidelidad (*faithfulness*), relevancia de la respuesta, corrección y calidad de las citas. Estas dimensiones permiten evaluar de manera más integral los sistemas que combinan recuperación y generación. Asimismo, Gao et al. [7] señalan que, además de las métricas automáticas tradicionales de IR (precisión,

recall, nDCG), resulta fundamental incluir evaluaciones humanas que capten aspectos como coherencia, transparencia y verificabilidad.

Categorías de Evaluación

- **Enfoques de Evaluación - Gao et al. (2023)**
  - Evaluación independiente por etapas: análisis separado de retrieval, augmentation y generation, con métricas específicas en cada módulo.
  - Evaluación End-to-End: valoración directa de la salida final del sistema RAG, con dos variantes:
    - Evaluación End-to-End automática: frameworks que miden habilidades/abilities clave como exactitud, fidelidad (*faithfulness*), atribución de fuentes, reducción de alucinaciones y transparencia.
    - Evaluación End-to-End con juicio humano: juicios expertos que valoran coherencia, verificabilidad, utilidad práctica y confianza.
  - Combinación de métricas: integración de métricas clásicas de recuperación (precisión, recall, nDCG), métricas de generación (coherencia, verificabilidad, calidad narrativa) y evaluación humana.
- **Evaluación en IR con FATE — Bernard & Balog (2025)**
  - Fairness: métricas automáticas como *top-k*, *exposure* y *pairwise metrics*.
  - Transparency y Accountability: auditorías y estudios de usuarios, centrados en interpretabilidad y trazabilidad.
  - Ethics: enfoques cualitativos vinculados a privacidad y seguridad.
  - Tensiones: conflictos entre fairness individual vs. grupal y entre transparencia excesiva vs. carga cognitiva.
- **Evaluación Clásica en IR — Hambarde & Proença (2023)**
  - Métricas de recuperación: precisión, recall, F1, MAP (Mean Average Precision), MRR (Mean Reciprocal Rank).
  - Métricas de ranking: nDCG (Normalized Discounted Cumulative Gain) y calidad del ordenamiento.
  - Evaluación de modelos neuronales: comparación con benchmarks tradicionales (TREC, MS MARCO).

    - Dimensión multi-modal: evaluación de IR en escenarios que integran texto, imágenes y audio.

- **Evaluación en RAG — Knollmeyer et al. (2024); Gao et al. (2023)**
    - *Fase de Recuperación*:
        - Context relevance: grado de pertinencia de los documentos recuperados.
        - Dataset quality: uso de bases de datos curadas y representativas (ej. Wikipedia, MS MARCO).
        - Métricas aplicadas: precisión, recall, nDCG, cobertura de conocimiento.
    - *Fase de Generación*:
        - Faithfulness: consistencia factual entre recuperación y generación.
        - Answer relevance: utilidad de la respuesta respecto a la consulta.
        - Correctness: exactitud de la información producida y reducción de alucinaciones.
        - Métricas aplicadas: BLEU, ROUGE, métricas de consistencia factual y evaluaciones humanas.
    - *Fase de Integración*:
        - Citation quality: precisión, trazabilidad y cobertura de las fuentes citadas.
    - *Evaluación Global*:
        - Evaluadores mixtos: combinación de métricas automáticas (similaridad semántica, BLEU, ROUGE) con juicios humanos (coherencia, verificabilidad, utilidad práctica).
        - Datasets: necesidad de colecciones específicas adaptadas a RAG, más allá de benchmarks tradicionales de IR.

Futuro de RAG

Desafios Actuales De acuerdo con Zhai [22], uno de los principales desafíos de los sistemas RAG es la generación de alucinaciones, que comprometen la confianza y limitan su uso en aplicaciones críticas. Asimismo, persisten limitaciones en la calidad de la recuperación y en la eficiencia computacional, especialmente en dominios

donde se requiere información actualizada y especializada (Hu & Lu, 2024). Además, Ramdurai [19] señala que la integración de RAG con otras arquitecturas, como redes neuronales convolucionales, enfrenta problemas de escalabilidad y de adaptación a contextos heterogéneos.

Direcciones Potenciales Las líneas de avance incluyen el desarrollo de retrievers más robustos, la optimización de las interacciones entre modelo y recuperación, así como mecanismos de evaluación que prioricen la relevancia y la coherencia.Ramdurai [19] enfatiza la sinergia con arquitecturas multimodales y con sistemas capaces de combinar información textual, visual y estructurada. Asimismo, enfoques como RAPTOR que introduce resúmenes recursivos y jerárquicos en la recuperación muestran cómo superar la fragmentación del contexto y mejorar la integración semántica (Sarthi et al., 2024).

Perspectivas A medio plazo, RAG y los LLMs convergerán con la IR clásica hacia sistemas de acceso conversacional al conocimiento: motores de búsqueda "con opinión" pero con respaldo documental, capaces de planificar, citar y aprender de la interacción. Incluso si los LLMs de contexto largo siguen mejorando, la recuperación seguirá siendo diferencial por costo, frescura, control y gobernanza de la evidencia; veremos stacks donde RAG, herramientas y memoria trabajan de forma coordinada, con evaluación centrada en utilidad y veracidad para tareas compuestas y empresariales

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 [3], 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.

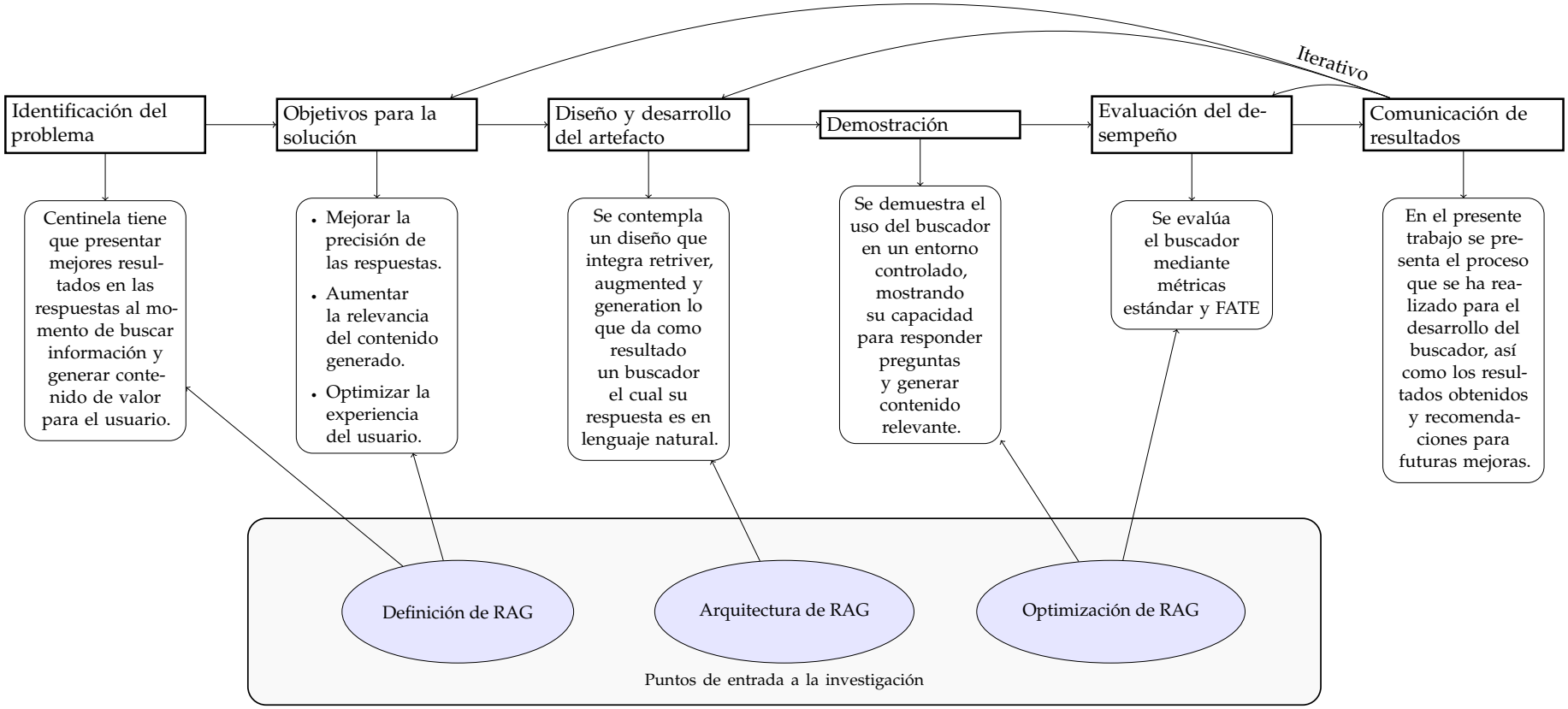

Figura 2.1: Proceso de Diseño de Investigación para el desarrollo de RAG en Centinela

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] Nolwenn Bernard y Krisztian Balog. «A Systematic Review of Fairness, Accountability, Transparency, and Ethics in Information Retrieval». En: *ACM Computing Surveys* 57.6 (feb. de 2025), 136:1-136:29. DOI: 10.1145/3637211.

[3] 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.

[4] 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.

[5] Wenqi Fan et al. «A Survey on RAG Meeting LLMs: Towards Retrieval-Augmented Large Language Models». En: *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining KDD '24*. Barcelona, Spain: ACM, 2024, págs. 6491-6501. DOI: 10.1145/3637528.3671470. URL: https://doi.org/10.1145/3637528.3671470.

[6] Yixing Fan et al. «Pre-training Methods in Information Retrieval». En: *arXiv preprint arXiv:2111.13853* (2021). URL: https://arxiv.org/abs/2111.13853.

[7] 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.

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

[9] 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.

[10] Yikun Han, Chunjiang Liu y Pengfei Wang. «A Comprehensive Survey on Vector Database: Storage and Retrieval Technique, Challenge». En: *arXiv preprint arXiv:2310.11703* (2023). URL: https://arxiv.org/abs/2310.11703.

[11] Yucheng Hu y Yuxing Lu. «RAG and RAU: A Survey on Retrieval-Augmented Language Models in Natural Language Processing». En: *arXiv preprint arXiv:2404.19543* (2024). URL: https://arxiv.org/abs/2404.19543.

[12] Zhi Jing, Yongye Su y Yikun Han. «When Large Language Models Meet Vector Databases: A Survey». En: *arXiv preprint arXiv:2402.01763* (2024). URL: https://arxiv.org/abs/2402.01763.

[13] Satyadhar Joshi. «Introduction to Vector Databases for Generative AI: Applications, Performance, Future Projections, and Cost Considerations». En: *International Advanced Research Journal in Science, Engineering and Technology (IARJSET)* 12.2 (2025), págs. 79-89. DOI: 10.17148/IARJSET.2025.12210. URL: https://doi.org/10.17148/IARJSET.2025.12210.

[14] Simon Knollmeyer et al. «Benchmarking of Retrieval Augmented Generation: A Comprehensive Systematic Literature Review on Evaluation Dimensions, Evaluation Metrics and Datasets». En: *Proceedings of the 16th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2024) - Volume 3: KMIS*. SCITEPRESS – Science y Technology Publications, Lda, 2024, págs. 137-148. ISBN: 978-989-758-716-0. DOI: 10.5220/0013065700003838. URL: https://doi.org/10.5220/0013065700003838.

[15] Le Ma et al. «A Comprehensive Survey on Vector Database: Storage and Retrieval Technique, Challenge». En: *arXiv preprint arXiv:2310.11703* (2025). URL: https://arxiv.org/abs/2310.11703v2.

[16] Shervin Minaee et al. «Deep Learning–based Text Classification: A Comprehensive Review». En: *ACM Computing Surveys* 54.3 (2021), 62:1-62:40. ISSN: 0360-0300. DOI: 10.1145/3439726.

[17] 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.

[18] 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.

[19] 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.

[20] Parth Sarthi et al. «RAPTOR: Recursive Abstractive Processing for Tree-Organized Retrieval». En: *Proceedings of the International Conference on Learning Representations (ICLR)*. Published as a conference paper at ICLR 2024. OpenReview, 2024. URL: https://openreview.net/forum?id=raptor2024.

[21] Ayisha Tabassum y Rajendra R. Patil. «A Survey on Text Pre-Processing & Feature Extraction Techniques in Natural Language Processing». En: *International Research Journal of Engineering and Technology (IRJET)* 7.6 (2020), págs. 4864-4870. ISSN: 2395-0056. URL: https://www.irjet.net/archives/V7/i6/IRJET-V7I6872.pdf.

[22] ChengXiang Zhai. «Large Language Models and Future of Information Retrieval: Opportunities and Challenges». En: *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*. Washington, DC, USA: ACM, 2024, págs. 1-10. DOI: 10.1145/3626772.3657848. URL: https://doi.org/10.1145/3626772.3657848.

[23] Wan Zhang y Jing Zhang. «Hallucination Mitigation for Retrieval-Augmented Large Language Models: A Review». En: *Mathematics* 13.5 (2025), pág. 856. DOI: 10.3390/math13050856. URL: https://doi.org/10.3390/math13050856.

[24] P. Zhao et al. «Retrieval-Augmented Generation for AI-Generated Content: A Survey». En: *arXiv* abs/2402.19473 (2024). APA style: Zhao, P., Zhang, H., Yu, Q., Wang, Z., Geng, Y., Fu, F., Yang, L., Zhang, W., Jiang, J., & Cui, B. (2024). Retrieval-augmented generation for AI-generated content: A survey. *arXiv.* https://doi.org/10.48550/arXiv.2402.19473. DOI: 10.48550/arXiv.2402.19473. URL: https://arxiv.org/abs/2402.19473.