



EDUCACIÓN

SECRETARÍA DE EDUCACIÓN PÚBLICA



TECNOLÓGICO
NACIONAL DE MÉXICO

Tecnológico Nacional de México

**Centro Nacional de Investigación
y Desarrollo Tecnológico**

Tesis de Maestría

**Desarrollo de un modelo basado en lenguaje natural
para la recuperación de información histórica y
cultural mediante técnicas de retrieval augmented
generation**

presentado por

Ing. Johan Argenis Franco Rogel

como requisito para la obtención del grado de
Maestro en Ciencias de la Computación

Director de tesis

Dr. Noé Alejandro Castro Sánchez

Codirector de tesis

Dr. Máximo López Sánchez

Cuernavaca, Morelos, México. Septiembre de 2026.



Centro Nacional de Investigación y Desarrollo Tecnológico
Departamento de Ciencias Computacionales

Cuernavaca, Mor., 14/AGOSTO/2026

OFICIO No. DCC/157/2026

Asunto: Aceptación de documento de tesis
CENIDET-AC-004-M14-OFICIO

CARLOS MANUEL ASTORGA ZARAGOZA
SUBDIRECTOR ACADÉMICO
PRESENTE

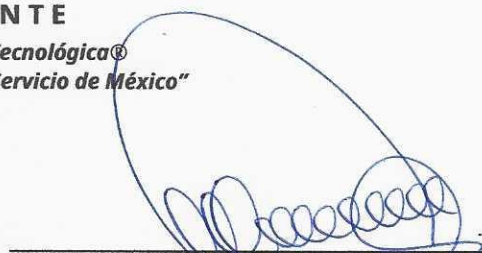
Por este conducto, los integrantes de Comité Tutorial de **Johan Argenis Franco Rogel** con número de control **M24CE052**, de la Maestría en Ciencias de la Computación, le informamos que hemos revisado el trabajo de tesis de grado titulado **"Desarrollo de un modelo basado en lenguaje natural para la recuperación de información histórica y cultural mediante técnicas de Retrieval-Augmented generation"** y hemos encontrado que se han atendido todas las observaciones que se le indicaron, por lo que hemos acordado aceptar el documento de tesis y le solicitamos la autorización de impresión definitiva.

ATENTAMENTE

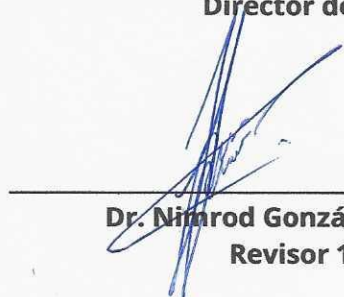
Excelencia en Educación Tecnológica®
"Conocimiento y Tecnología al Servicio de México"



Dr. Noé Alejandro Castro Sánchez
Director de tesis



Dr. Máximo López Sánchez
Codirector de tesis



Dr. Nimrod González Franco
Revisor 1



Dr. Juan Gabriel González Serna
Revisor 2

C.c.p. Depto. Servicios Escolares.
Expediente / Estudiante



2026
año de
Margarita
Maza





Centro Nacional de Investigación y Desarrollo Tecnológico
Subdirección Académica

Cuernavaca Mor, 19/agosto/2026

Oficio No. SAC/262/2026

Asunto: Autorización de impresión de tesis

JOHAN ARGENIS FRANCO ROGEL
CANDIDATO AL GRADO DE MAESTRO EN CIENCIAS
DE LA COMPUTACIÓN
P R E S E N T E

Por este conducto, tengo el agrado de comunicarle que el Comité Tutorial asignado a su trabajo de tesis titulado **"Desarrollo de un modelo basado en lenguaje natural para la recuperación de información histórica y cultural mediante técnicas de Retrieval-Augmented generation"**, ha informado a esta Subdirección Académica, que están de acuerdo con el trabajo presentado. Por lo anterior, se le autoriza a que proceda con la impresión definitiva de su trabajo de tesis.

Esperando que el logro del mismo sea acorde con sus aspiraciones profesionales, reciba un cordial saludo.

ATENTAMENTE

Excelencia en Educación Tecnológica®

"Conocimiento y Tecnología al servicio de México"


CARLOS MANUEL ASTORGA ZARAGOZA
SUBDIRECTOR ACADÉMICO



c.c.p. Departamento de Ciencias Computacionales
Departamento de Servicios Escolares

CMAZ/lmz



2026
año de
Margarita
Maza



Dedicatoria

A mi madre, por su apoyo constante a lo largo de mi formación académica y por acompañarme durante cada etapa de este camino.

A mi familia, por estar presente y brindarme su apoyo durante la realización de este trabajo.

Con gratitud, les dedico este logro.

Agradecimientos

Quiero expresar mi más sincera gratitud a todas las personas e instituciones que hicieron posible este trabajo.

Agradezco a la Secretaría de Ciencia, Humanidades, Tecnología e Innovación (SECIHTI) por su apoyo económico, que abrió las puertas para que esta tesis pudiera realizarse.

Mi reconocimiento al Tecnológico Nacional de México, campus Centro Nacional de Investigación y Desarrollo Tecnológico (TecNM/CENIDET), y a todo su personal, por ofrecer un entorno que fomentó mi aprendizaje, crecimiento y desarrollo académico.

A mi director de tesis, el Dr. Noé Alejandro Castro Sánchez, y a mi comité tutorial, por sus orientaciones, observaciones y consejos constantes. Su acompañamiento y experiencia fueron clave para avanzar y fortalecer cada etapa de este trabajo.

A mis maestros y compañeros de maestría, por las ideas compartidas, el apoyo mutuo y los momentos de aprendizaje que hicieron de esta etapa una experiencia enriquecedora y memorable.

Finalmente, agradezco a todos quienes, con su apoyo, confianza y cariño, me acompañaron a lo largo de este camino. Cada gesto, consejo y palabra de aliento contribuyó a que este logro se hiciera realidad.

Resumen

Los modelos de lenguaje de gran tamaño han demostrado un desempeño sobresaliente en tareas de generación de texto; sin embargo, presentan limitaciones para responder consultas sobre dominios especializados debido a la posibilidad de generar información no respaldada por fuentes documentales. Una estrategia ampliamente utilizada para reducir este problema consiste en emplear arquitecturas de Generación Aumentada por Recuperación (RAG por sus siglas en inglés), las cuales incorporan información externa durante el proceso de generación.

En esta investigación se desarrolló y evaluó un sistema RAG orientado al dominio museográfico, empleando un modelo de lenguaje pequeño (SLM por sus siglas en inglés) y una estrategia de recuperación híbrida basada en la combinación lineal de evidencia léxica y evidencia semántica. Como parte del trabajo se realizaron cinco experimentos: evaluación de la robustez del recuperador ante errores ortográficos; selección del modelo generativo mediante el banco de pruebas *Multilingual Massive Multitask Language Understanding* (MMMLU); comparación de estrategias de recuperación sobre el conjunto MessIRve; evaluación integral del sistema RAG mediante métricas de recuperación, generación y evaluadores automáticos; y evaluación con usuarios para analizar el desempeño y la interacción con el sistema.

Los resultados muestran que la estrategia híbrida propuesta supera a los métodos de recuperación convencionales, obteniendo mejores valores en *MRR* y exhaustividad (*Recall*). Asimismo, la implementación desarrollada obtuvo un mejor desempeño que una implementación equivalente basada en *LangChain* tanto en recuperación como en generación de respuestas. Finalmente, la evaluación con usuarios evidenció que el sistema facilita la consulta de información histórica mediante lenguaje natural y favorece la exploración del contenido museográfico.

Palabras clave: *Retrieval-Augmented Generation*, RAG, modelos de lenguaje pequeños, recuperación de información, sistemas de preguntas y respuestas, museos, procesamiento de lenguaje natural.

Abstract

Large Language Models have achieved remarkable performance in natural language generation tasks; however, they may produce responses that are not supported by reliable sources when operating in specialized domains. Retrieval-Augmented Generation (RAG) has emerged as an effective approach to mitigate this limitation by incorporating external knowledge during the generation process.

This research presents the development and evaluation of a RAG system for the museum domain based on a Small Language Model (SLM) and a hybrid retrieval strategy that combines lexical and semantic evidence through a linear weighting scheme. Five experiments were conducted: robustness evaluation under spelling errors, selection of the generative model using the MMMLU benchmark, comparison of retrieval strategies on the MessIRve dataset, end-to-end evaluation of the RAG system using retrieval, generation and automatic evaluation metrics, and a user study to analyze system performance and interaction.

The results show that the proposed hybrid retrieval strategy outperforms conventional retrieval methods, achieving higher MRR and Recall values. In addition, the proposed implementation obtained better retrieval and generation performance than an equivalent LangChain-based implementation. Finally, the user evaluation indicates that the system supports natural language access to historical information and encourages a more interactive exploration of museum content.

Keywords: Retrieval-Augmented Generation, RAG, Small Language Models, Information Retrieval, Question Answering Systems, Museums, Natural Language Processing.

Contenido

Resumen	2
Abstract	3
Capítulo I. Introducción	9
1.1 Planteamiento del problema.....	9
1.2 Pregunta de investigación.....	10
1.3 Objetivos	10
1.4 Justificación	11
1.5 Alcances y limitaciones	11
1.6 Contribución	12
1.7 Organización del documento.....	12
Capítulo II. Marco conceptual.....	14
2.1 Procesamiento de lenguaje natural	14
2.2 <i>Transformers</i>	14
2.3 Modelos de lenguaje largos	15
2.4 Modelos de lenguaje pequeños.....	16
2.5 Recuperación de información.....	16
2.6 <i>Prompt</i>	16
2.7 <i>Embeddings</i> y representación vectorial	16
Capítulo III. Estado del arte.....	18
3.1 Sistemas RAG y NLP en entornos culturales	19
3.2 Mitigación de alucinaciones en modelos de lenguaje	30
3.3 Discusión.....	37
Capítulo IV. Generación aumentada por recuperación.....	39
4.1 Introducción a la recuperación aumentada por generación	39
4.2 Arquitectura general de un sistema RAG	40
4.3 Conocimiento paramétrico y conocimiento no paramétrico.....	41
4.4 El fenómeno de las alucinaciones en modelos de lenguaje.....	41
4.5 Estrategias para reducir alucinaciones	43
4.6 Recuperación de información dentro de RAG.....	44

4.7	Recuperación dispersa	44
4.8	Recuperación densa.....	45
4.9	Recuperación híbrida.....	47
4.10	Flujo completo de un sistema RAG.....	48
4.11	Métricas para evaluación RAG.....	49
Capítulo V. Método de solución		55
5.1	Etapa 1. Investigación y selección de estrategias	55
5.2	Etapa 2. Diseño e integración del sistema	58
5.3	Etapa 3. Validación experimental.....	62
Capítulo VI. Resultados y discusión		71
6.1	Resultados de la revisión contextual sobre comunicación museográfica.....	71
6.2	Experimento 1. Robustez ante errores ortográficos	72
6.3	Experimento 2. Evaluación de modelos generativos	74
6.4	Experimento 3. Evaluación de estrategias de recuperación	75
6.5	Experimento 4. Evaluación integral del sistema RAG	80
6.6	Experimento 5. Evaluación con usuarios	84
6.7	Discusión de resultados.....	86
Capítulo VII. Conclusiones		89
7.1	Conclusiones.....	89
7.2	Productos y aportaciones.....	90
7.3	Trabajo futuro	91
Referencias		92
Anexos.....		97
Anexo A. <i>Prompt</i> utilizado en la evaluación automática		97
Anexo B. Constancia de presentación del artículo Postprocesamiento Geométrico de Espacios de <i>Embeddings</i> para Clasificación de Textos con Voronoi en COMIA 2026.		100
Anexo C. Reconocimiento como ponente de la conferencia <i>La magia detrás de los LLM</i>		100
Anexo D. Reconocimiento por impartición del taller Sistemas de Lenguaje Natural en Inteligencia Artificial.		101

Anexo E. Constancia de participación en el certamen InnovaTec 2025 con el proyecto SanEEG.....	101
--	-----

Índice de tablas

Tabla 3.1. Comparación de sistemas RAG y NLP aplicados a entornos culturales y museográficos.	29
Tabla 3.2. Comparación de enfoques para la mitigación de alucinaciones en modelos de lenguaje y sistemas RAG.	37
Tabla 5.1. Configuración de parámetros del sistema RAG.....	62
Tabla 5.2. Modelos de embeddings usados en experimento 1.....	64
Tabla 5.3. Modelos de lenguaje pequeños comparados para seleccionar el generador del sistema RAG (Experimento 2).....	64
Tabla 5.4. Configuraciones del experimento 2.....	65
Tabla 5.5. Especificaciones de los entornos utilizados.....	65
Tabla 6.1. Valor óptimo de alfa por familia que indica el mejor ajuste para la métrica MRR.....	72
Tabla 6.2. Resultados promedio para consultas con 2 errores ortográficos.....	73
Tabla 6.3. Resultados promedio para consultas con 4 errores ortográficos.....	73
Tabla 6.4. Modelos con mejor y peor desempeño global medido mediante MRR promedio en alfa.....	74
Tabla 6.5. Desempeño global por configuración medido con MCRC.	74
Tabla 6.6. Total de aciertos y MCRC global de los modelos por configuración de prompting.....	75
Tabla 6.7. Desempeño promedio por método en el contraste principal.....	76
Tabla 6.8. Estabilidad del peso seleccionado y cambio de MRR respecto a RSF.....	77
Tabla 6.9. Ganancia promedio respecto a la referencia léxica pura (BM25).....	79
Tabla 6.10. Top 5 modelos según su mejor MRR.	79
Tabla 6.11. Resultados de recuperación a $k = 10$ en el subconjunto de evaluación ($n = 118$).....	81
Tabla 6.12. Resultados de generación (ROUGE) en el subconjunto de evaluación ($n = 118$).....	81
Tabla 6.13. Evaluación del sistema de recuperación sobre el conjunto completo de consultas.	81
Tabla 6.14. Resultados agregados reportados por los evaluadores LLM.....	82
Tabla 6.15. Acuerdo interevaluador medido mediante kappa de Fleiss en las variables de evaluación.....	82
Tabla 6.16. Distribución de participantes por grupo experimental.....	84

Tabla 6.17. Pruebas estadísticas y valores p ajustados mediante Holm	84
Tabla 6.18. Correlaciones y valores p ajustados mediante Holm.....	84
Tabla 6.19. Distribución de categorías de consultas realizadas al sistema RAG	85
Tabla 6.20. Frecuencia de lemas identificados en las consultas	85
Tabla 6.21. Tópicos identificados mediante modelado LDA	85
Tabla 6.22. Comparación de métodos para clustering semántico	85

Índice de figuras

Figura 5.1. Etapas del método de solución	55
Figura 5.2. Arquitectura RAG propuesta.....	59
Figura 5.3. Módulo de recuperación propuesto	60
Figura 6.1. Distribución de MRR por método y nivel de ruido ortográfico	73
Figura 6.2. MRR promedio por método con dispersión; la línea punteada indica la referencia BM25.....	76
Figura 6.3. Cambio de MRR de la ponderación ajustada respecto a RSF por modelo.....	78
Figura 6.4. Matriz MRR por modelo y método para identificar estabilidad y casos extremos	80
Figura B.1. Anexo B. Constancia de presentación de artículo.....	100
Figura C.1. Anexo C. Reconocimiento como ponente.....	100
Figura D.1. Anexo D. Reconocimiento por impartir un taller	101
Figura E.1. Anexo E. Constancia de participación en el certamen InnovaTec 2025.....	101

Capítulo I. Introducción

En los últimos años, los modelos de lenguaje largos (LLM por sus siglas en inglés) han transformado la generación de texto y el procesamiento de lenguaje natural. No obstante, al aplicarse a dominios especializados, presentan limitaciones críticas como las «alucinaciones» o la generación de información no respaldada por fuentes documentales. Para mitigar esto, las arquitecturas de RAG han emergido como una solución eficiente al integrar bases de conocimiento externas. En este contexto, la presente investigación explora la viabilidad de estas tecnologías en el ámbito cultural, desarrollando un sistema RAG optimizado para el dominio museográfico que combina un SLM con una estrategia de recuperación híbrida.

Los museos desempeñan un papel importante en la preservación y difusión del patrimonio histórico y cultural. Sin embargo, datos recientes del Instituto Nacional de Estadística y Geografía (INEGI) indican que únicamente el 39.3 % de la población asistió a un museo o exposición durante el último año. Entre las principales razones señaladas por los propios visitantes sobre la baja asistencia destacan la falta de difusión de las actividades culturales (18.2 %), la percepción de que los museos son poco atractivos o alejados de la vida cotidiana (16.4 %) y la falta de tiempo (15.9 %) [1].

1.1 Planteamiento del problema

En respuesta a la baja afluencia y a las demandas de la era digital, diversas instituciones han impulsado la digitalización de colecciones, archivos y repositorios históricos. Este esfuerzo ha permitido salvaguardar grandes volúmenes de datos; sin embargo, el almacenamiento digital no garantiza por sí mismo el acceso democrático al patrimonio por parte del público general. A pesar de la disponibilidad de estos repositorios, la literatura y la información se continúan presentando principalmente mediante cédulas físicas, paneles informativos tradicionales o catálogos digitales. Si bien estos recursos facilitan la consulta del contenido, mantienen un esquema de interacción esencialmente estático, en el que los visitantes únicamente pueden acceder a la información previamente estructurada, sin la posibilidad de realizar consultas específicas, resolver dudas espontáneas o profundizar en los temas de su interés mediante una interacción natural [2].

Por otra parte, los sistemas de búsqueda disponibles suelen depender de palabras clave o estructuras predefinidas, restringiendo el acceso a la información mediante lenguaje natural. Como resultado, existe una brecha entre la cantidad de información histórica y cultural disponible y las herramientas utilizadas para consultarla. Esta situación limita el potencial de las tecnologías actuales para apoyar formas más interactivas de exploración y acceso al patrimonio cultural en contextos museográficos.

Asimismo, la literatura disponible al inicio de esta investigación mostraba que las aplicaciones existentes se centraban en la digitalización de contenidos o en mecanismos convencionales de búsqueda, sin reportar implementaciones basadas en modelos de lenguaje pequeños (SLM) que integraran recuperación semántica y generación de respuestas para el

dominio museográfico. Esta situación representa una oportunidad para investigar estrategias de recuperación de información y su integración en arquitecturas RAG, con el propósito de facilitar el acceso a la información histórica y cultural mediante consultas en lenguaje natural.

1.2 Pregunta de investigación

A partir de la problemática, surge la siguiente pregunta de investigación:

¿Es viable utilizar un SLM para generar respuestas contextualizadas a consultas sobre información museográfica que mantengan la fidelidad fáctica con respecto a los documentos de referencia?

1.3 Objetivos

1.3.1 Objetivo general

Desarrollar un sistema de recuperación y generación de información para el dominio museográfico basado en modelos de lenguaje pequeños, mediante la adaptación y mejora de una arquitectura RAG.

1.3.2 Objetivos específicos

- Analizar las limitaciones actuales en la presentación de información en museos.
- Integrar los componentes de recuperación y generación en una arquitectura RAG para responder consultas sobre información museográfica.
- Evaluar la robustez y efectividad del módulo de recuperación de forma aislada ante ruido ortográfico y variabilidad textual léxica y semántica, y estrategias de fusión de resultados.
- Validar el funcionamiento integral del modelo RAG mediante similitud textual automatizada y análisis estadístico-lingüístico de la interacción de usuarios.

1.4 Justificación

El interés por incorporar herramientas de inteligencia artificial en actividades de consulta y acceso a la información ha crecido considerablemente durante los últimos años. Sin embargo, gran parte de estas aplicaciones se han desarrollado para dominios generales, existiendo menos trabajos orientados a información histórica y cultural.

Ante la necesidad de crear interfaces más dinámicas, el Procesamiento de Lenguaje Natural (PLN) y la Inteligencia Artificial Generativa se presentan como alternativas para habilitar sistemas capaces de interactuar con los usuarios mediante lenguaje natural. Entre las distintas aproximaciones propuestas en la literatura, la Generación Aumentada por Recuperación (RAG en inglés) combina mecanismos de recuperación de información con modelos generativos, permitiendo construir sistemas de consulta basados en colecciones documentales específicas [2][3].

Aunque existen avances en sistemas de preguntas y respuestas (*QA Systems*) para distintos dominios, en el contexto mexicano aún son escasos los trabajos orientados a la consulta interactiva de información histórica y cultural mediante técnicas RAG. Asimismo, los repositorios digitales existentes suelen depender de mecanismos tradicionales de búsqueda y navegación, limitando las posibilidades de exploración mediante lenguaje natural.

Por ello, la presente investigación busca analizar la aplicación de técnicas RAG en el dominio histórico-cultural, así como evaluar el desempeño de distintos métodos de recuperación y el funcionamiento integral del sistema propuesto como herramienta de acceso y exploración de información en contextos museográficos.

1.5 Alcances y limitaciones

1.5.1 Alcances

- Desarrollar un sistema basado en técnicas de RAG para la consulta de información histórica y cultural mediante lenguaje natural.
- Implementar un módulo de recuperación de información basado en enfoques léxicos, semánticos y de fusión de rango.
- Evaluar la robustez del módulo de recuperación ante escenarios de ruido ortográfico y variabilidad textual.
- Integrar un modelo generativo para la construcción de respuestas a partir de la información recuperada.
- Validar el funcionamiento del sistema mediante métricas automáticas de recuperación, similitud textual y análisis de la interacción de los usuarios.
- Implementar la propuesta mediante una arquitectura de microservicios accesible a través de una aplicación web.

1.5.2 Limitaciones

- El sistema estará limitado a información contenida en el corpus documental utilizado durante la investigación.
- El estudio se enfocará exclusivamente en información histórica y cultural en idioma español.
- La calidad de las respuestas dependerá de la cobertura y calidad de los documentos incorporados al corpus.
- La evaluación experimental se realizará sobre los conjuntos de datos y escenarios definidos en esta investigación, por lo que los resultados podrían variar en otros dominios o colecciones documentales.
- La investigación no contempla la integración directa con bases de datos institucionales de museos ni la actualización automática de contenidos externos.
- El sistema se desarrollará como un prototipo de investigación y no como una plataforma de producción para uso institucional.

1.6 Contribución

La principal contribución de esta investigación consiste en la propuesta de un método de recuperación híbrida basado en una combinación lineal ponderada de evidencia léxica y evidencia semántica mediante puntajes normalizados. La estrategia propuesta permite integrar simultáneamente la coincidencia textual y la similitud semántica dentro de un mismo esquema de recuperación.

Los resultados experimentales muestran que la formulación ponderada generaliza la fusión relativa de puntuaciones (*Relative Score Fusion*, RSF), estrategia que normaliza y combina las puntuaciones de los recuperadores. RSF corresponde al caso particular $\alpha = 0.50$ cuando ambas señales normalizadas reciben el mismo peso. El desempeño se cuantificó mediante *Mean Reciprocal Rank* (MRR), que promedia el recíproco de la posición del primer documento relevante. Por ello, la contribución no consiste en superar a RSF con esa misma configuración, sino en seleccionar el peso de fusión sin utilizar los datos de prueba. En la evaluación complementaria con 16 modelos, tres semillas y validación cruzada anidada, la ponderación ajustada produjo una variación positiva de MRR en 43 de 48 escenarios y una mejora robusta, definida por un intervalo *bootstrap* del 95 % completamente positivo, en 37 de 48 escenarios. El beneficio dependió del modelo y no se observó de forma universal.

1.7 Organización del documento

El presente documento se organiza en siete capítulos.

El **Capítulo I** presenta la introducción de la investigación, incluyendo el planteamiento del problema, la pregunta de investigación, los objetivos, la justificación, los alcances, las limitaciones, la contribución y la organización del documento.

El **Capítulo II** expone el marco conceptual que sustenta la investigación. Se describen los principales conceptos relacionados con el procesamiento de lenguaje natural, los modelos *transformer*, los modelos de lenguaje, la recuperación de información, los *prompts* y los *embeddings*.

El **Capítulo III** presenta el estado del arte, analizando trabajos recientes sobre sistemas RAG aplicados a entornos culturales y museográficos, así como investigaciones enfocadas en la mitigación de alucinaciones en modelos de lenguaje.

El **Capítulo IV** desarrolla los fundamentos de la arquitectura RAG, describiendo sus componentes, estrategias de recuperación, técnicas de reducción de alucinaciones y métricas de evaluación.

El **Capítulo V** describe el método de solución propuesto. Se presentan las etapas de investigación, diseño e integración del sistema, así como la validación experimental mediante los cinco experimentos desarrollados.

El **Capítulo VI** presenta los resultados obtenidos y su discusión. Se reportan los resultados experimentales de los cinco experimentos realizados, incluyendo la evaluación de recuperación, generación y usuarios, además de la interpretación de los principales hallazgos.

Finalmente, el **Capítulo VII** expone las conclusiones de la investigación, las principales aportaciones obtenidas y las líneas de trabajo futuro.

Capítulo II. Marco conceptual

2.1 Procesamiento de lenguaje natural

El Procesamiento de Lenguaje Natural (PLN) es una rama de la Inteligencia Artificial que estudia el desarrollo de métodos computacionales para analizar, procesar y generar lenguaje humano. Su objetivo es permitir que los sistemas informáticos trabajen con información expresada en lenguaje natural, facilitando tareas como clasificación de texto, traducción automática, recuperación de información, análisis semántico y sistemas de preguntas y respuestas.

El PLN integra conocimientos provenientes de la lingüística, las ciencias de la computación y el aprendizaje automático. A través de estas disciplinas, los sistemas pueden identificar patrones lingüísticos y representar información textual de forma que resulte procesable por algoritmos computacionales.

2.2 *Transformers*

Con la publicación de Vaswani et al. [4], la arquitectura *transformer* introdujo un cambio de paradigma en el procesamiento de lenguaje natural al prescindir por completo de las conexiones recurrentes o convolucionales tradicionales, basando su funcionamiento exclusivamente en mecanismos de atención. Esta estructura permite un procesamiento de la información totalmente en paralelo a lo largo de secuencias de texto complejas, eliminando las restricciones de dependencia lineal paso a paso. El diseño del modelo se fundamenta en una estructura clásica de codificador (*encoder*) y decodificador (*decoder*), donde el codificador mapea una secuencia de representaciones continuas de entrada a una secuencia de vectores intermedios, y el decodificador genera la secuencia de salida elemento por elemento.

Los autores destacan que el componente central y novedoso de esta arquitectura es el mecanismo de atención de producto punto escalado (*Scaled Dot-Product Attention*), junto con su extensión denominada atención de múltiples cabezas (*Multi-Head Attention*). Matemáticamente, la atención se describe como el mapeo de un vector de consulta (*Query*, Q) y un conjunto de pares de vectores de clave (*Key*, k) y valor (*Value*, V) hacia una salida, donde las matrices operan bajo la siguiente ecuación:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (1)$$

Aquí, representa la dimensión de las claves, actuando como un factor de escalamiento que evita que el producto punto crezca excesivamente en dimensiones altas, lo que desestabilizaría los gradientes durante el entrenamiento. La atención de múltiples cabezas permite al modelo proyectar linealmente las consultas, claves y valores en diferentes

subespacios de representación de manera simultánea, logrando que el sistema atienda a información de diversos contextos y posiciones al mismo tiempo.

Finalmente, el estudio señala que, debido a la ausencia de recurrencia en la red neuronal, el modelo carece de una noción intrínseca sobre el orden secuencial de los elementos de entrada. Para resolver esta limitación, se incorporan las codificaciones posicionales (*Positional Encodings*) a las representaciones vectoriales (*embeddings*) de entrada en la base de las pilas del codificador y decodificador. Estas codificaciones utilizan funciones senoidales y cosenoidales de diferentes frecuencias para inyectar información sobre la posición relativa o absoluta de los *tokens* en la secuencia, lo que dota a los *transformers* de la capacidad de comprender la estructura ordenada del lenguaje natural conservando las ventajas de la computación paralela [4].

2.3 Modelos de lenguaje largos

De acuerdo con D. Naik [5], un Modelo de Lenguaje (LM por sus siglas en inglés) es un modelo de aprendizaje automático diseñado para comprender, procesar y generar lenguaje natural mediante un enfoque estadístico probabilístico. Este sistema codifica información lingüística para determinar la probabilidad de que una secuencia de fichas (*tokens*) ocurra en un orden determinado en función de los elementos precedentes. Dependiendo de la arquitectura, un *token* puede representar un carácter, un subfragmento de palabra o una palabra completa, componiendo el vocabulario base que permite al modelo realizar tareas tanto de clasificación como de generación de texto.

El autor destaca que los Modelos de Lenguaje Grande (LLMs por sus siglas en inglés) representan una evolución sofisticada de este concepto, caracterizados por su entrenamiento con volúmenes masivos de datos textuales procedentes de libros, artículos y archivos de internet. Este proceso, denominado preentrenamiento, permite al modelo capturar estructuras sintácticas, reglas gramaticales y vínculos semánticos. Mediante este aprendizaje previo, el modelo se especializa en la predicción del elemento subsecuente en una cadena de texto a partir de los patrones lingüísticos asimilados.

Finalmente, el estudio señala que, para descubrir estas correlaciones estadísticas a gran escala, los LLMs emplean redes neuronales sustentadas en la arquitectura *transformer*. Esta tecnología procesa la información de forma paralela y utiliza mecanismos de atención para evaluar la relevancia relativa de cada *token* dentro de una secuencia. Tras el preentrenamiento, estos modelos adquieren la capacidad de adaptarse dinámicamente a partir de la información recibida en tiempo real durante la interacción con los usuarios, estableciendo el fundamento operativo de la inteligencia artificial generativa actual [5].

2.4 Modelos de lenguaje pequeños

Los modelos de lenguaje pequeños (SLM por sus siglas en inglés) conservan los principios fundamentales de los modelos basados en *transformer*, pero reducen el número de parámetros y los requerimientos de memoria y cómputo. No existe un umbral universal para clasificarlos; su denominación es relativa al contexto de despliegue y a los recursos disponibles. Su menor tamaño facilita la ejecución local, disminuye la latencia y permite mayor control sobre los datos, aunque puede limitar la cobertura de conocimiento y el desempeño en tareas complejas.

Un ejemplo representativo es Phi-3-mini, modelo de 3.8 mil millones de parámetros diseñado para ejecutarse en dispositivos con recursos restringidos y que mantiene resultados competitivos en pruebas de razonamiento y comprensión [3].

2.5 Recuperación de información

La recuperación de información (*Information Retrieval*, IR) es la disciplina que estudia los métodos y sistemas para localizar material relevante típicamente documentos o fragmentos de texto en respuesta a una necesidad de información expresada por un usuario [6]. Su historia se remonta a los años cincuenta del siglo pasado, con los trabajos pioneros de Cranfield y las propuestas de Salton sobre el modelo de espacio vectorial [7], y ha evolucionado profundamente con la llegada de la web, los motores de búsqueda y, más recientemente, los modelos de lenguaje profundo.

2.6 Prompt

Un *prompt* es la entrada textual, instrucción o contexto que se proporciona a un modelo de lenguaje para guiar su salida hacia una tarea específica; en la literatura reciente también se describe como una interfaz entre el usuario y el LLM, y como un recurso para orientar generación, clasificación, razonamiento o extracción de información sin cambiar los parámetros del modelo [8].

2.7 Embeddings y representación vectorial

El concepto de *embedding* es central en el PLN moderno y constituye la base de los métodos de recuperación densa utilizados en los sistemas RAG. Un *embedding* es una función que mapea un objeto discreto (una palabra, una oración, un párrafo o un documento) a un vector de números reales en un espacio de alta dimensión, de manera que objetos semánticamente similares se mapean a vectores geoméricamente próximos [9][10].

La historia de los *embeddings* en procesamiento de lenguaje natural comienza con los modelos de espacio vectorial clásicos, como la Descomposición en Valores Singulares

(*Singular Value Decomposition*, SVD) aplicada a matrices de coocurrencia de palabras [11]. Sin embargo, el hito que marcó el inicio de la era moderna de los *embeddings* fue la publicación de Word2Vec por Mikolov et al [9], que demostró que era posible aprender representaciones vectoriales de palabras de alta calidad mediante redes neuronales superficiales entrenadas sobre grandes corpus de texto. Estas representaciones capturan relaciones semánticas y sintácticas de manera notable: la famosa analogía *rey - hombre + mujer $\approx$ reina* ilustra que las operaciones algebraicas en el espacio de *embeddings* corresponden a relaciones semánticas significativas.

La limitación fundamental de los *embeddings* estáticos como Word2Vec es que asignan un único vector a cada palabra, independientemente del contexto en el que aparece. Esto implica que palabras polisémicas (con múltiples significados) reciben la misma representación en todos sus usos, lo que es una simplificación excesiva. Esta limitación fue superada por los modelos de *embeddings* contextuales, que calculan representaciones diferentes para una misma palabra dependiendo del contexto en el que aparece. El modelo ELMo fue uno de los primeros en proporcionar *embeddings* contextuales de alta calidad, utilizando redes de memoria a corto y largo plazo bidireccionales (LSTM por sus siglas en inglés) [12]. Posteriormente, BERT revolucionó el campo al utilizar la arquitectura *transformer* y el mecanismo de atención para calcular representaciones contextuales de manera más eficiente y expresiva [10].

Capítulo III. Estado del arte

Una vez presentados en el capítulo anterior los conceptos básicos sobre recuperación de información, modelos generativos y arquitecturas RAG, es necesario analizar cómo se han usado estos enfoques en trabajos recientes y qué problemas siguen existiendo en contextos de patrimonio cultural y exploración histórica. En este sentido, el estado del arte se organiza en dos líneas que se relacionan con el problema central de esta investigación. La primera corresponde a los sistemas basados en RAG y PLN aplicados a entornos culturales, museográficos e históricos, los cuales muestran el potencial de estas arquitecturas para ofrecer acceso interactivo a colecciones especializadas. La segunda se relaciona con la precisión de las respuestas generadas y con las estrategias para reducir alucinaciones, donde la literatura señala que la calidad de la respuesta depende en gran medida de la evidencia recuperada durante el proceso de generación. En conjunto, ambas líneas permiten identificar una oportunidad de investigación: aunque existen múltiples propuestas de sistemas RAG en patrimonio cultural y diversos estudios sobre precisión en modelos generativos, aún hay poca evidencia experimental sobre cómo la calidad de la recuperación de información influye en la calidad de las respuestas generadas en contextos museográficos.

Con la finalidad de recopilar la evidencia científica más relevante y asegurar la replicabilidad de esta revisión, se diseñó una estrategia de búsqueda sistemática en bases de datos indexadas de alto impacto (tales como Scopus, Web of Science, IEEE Xplore, ACM Digital Library, Springer Link y ScienceDirect de Elsevier). Las cadenas de búsqueda utilizadas para identificar la literatura pertinente combinaron operadores booleanos (AND, OR) y términos en inglés relacionados con el núcleo tecnológico y el dominio de aplicación de este trabajo:

- ("retrieval-augmented generation" OR "RAG") AND ("cultural heritage" OR "museum" OR "archive" OR "digital library")
- ("NLP" OR "natural language processing") AND ("cultural heritage" OR "historical archives" OR "cultural tourism")
- ("hallucination mitigation" OR "factuality") AND ("large language model" OR "LLM")
- ("knowledge graph" OR "ontology") AND ("RAG" OR "question answering") AND "cultural"

Asimismo, el proceso de depuración y análisis se rigió bajo estrictos criterios de inclusión, limitando la selección a artículos arbitrados en inglés o español publicados en el periodo 2020–2025. Se dio prioridad a investigaciones con una metodología explícita sobre arquitecturas RAG o procesamiento de lenguaje natural aplicado al patrimonio cultural, así como a estudios con técnicas de mitigación de alucinaciones evaluadas empíricamente.

Como resultado de esta estrategia, de un total de 1,478 registros inicialmente identificados en las plataformas, se eliminaron duplicados (n=312), se excluyeron artículos por título y

resumen (n=1,073) tras verificar que no se alineaban con los objetivos, y se evaluaron 93 trabajos a texto completo. Finalmente, este proceso de filtrado resultó en la inclusión definitiva de 26 artículos, los cuales se analizan a detalle en las siguientes secciones.

3.1 Sistemas RAG y NLP en entornos culturales

3.1.1 Enhancement Large Language Models Domain Through Ontology-Based Retrieval-Augmented Generation

CAMPO	DETALLE
AUTORES	Fabio Clarizia, Massimo De Santo, Rosario Gaeta y Rocco Loffredo
AÑO	2025
PUBLICACIÓN	International Journal on Semantic Web and Information Systems, vol. 21, núm. 1, pp. 1–29
DOI	https://doi.org/10.4018/IJSWIS.392507

Se presenta el marco OB-RAG (*Ontology-Based Retrieval-Augmented Generation*), que incrusta triples sujeto-predicado-objeto de ontologías de dominio en un espacio vectorial continuo. Las ontologías se generan tanto manualmente por expertos del dominio como de forma semiautomática mediante técnicas de extracción de información. El marco incluye un módulo de conversión de triples ontológicos a representaciones vectoriales, un mecanismo de búsqueda semántica sobre este espacio vectorial, y un componente de generación condicionada por los triples recuperados. Se implementa un caso de estudio completo en patrimonio cultural con evaluación cuantitativa mediante métricas estándar y validación cualitativa por expertos del dominio [13].

3.1.2 RAG-Optimized Tibetan Tourism LLMs: Enhancing Accuracy and Personalization

CAMPO	DETALLE
AUTORES	Jinhu Qi, Shuai Yan, Yibo Zhang, Wentao Zhang, Rong Jin, Yuwei Hu y Ke Wang
AÑO	2024
PUBLICACIÓN	arXiv:2408.12003 (aceptado en AIPR 2024)
DOI / ENLACE	https://doi.org/10.48550/arXiv.2408.12003

Se propone un esquema de optimización RAG para LLMs de turismo tibetano. Se construye una base de datos de 563 puntos de vista turísticos curados manualmente de Ctrip Travel (fuente primaria) y Wikipedia (complementaria). Se evalúan sistemáticamente múltiples métodos de vectorización (TF-IDF, BERT) y algoritmos de indexación vectorial (Flat, HNSW, IVFFlat, SQ, IVFSQ, NSG, LSH). Se compara el enfoque RAG con *fine-tuning* en cuatro LLMs diferentes (ChatGLM3-6b, Baichuan2-7b, Qwen-7b-chat, Llama3-8b). El

diseño experimental es exhaustivo y sistemático, evaluando 14 combinaciones de método de vectorización $\times$ algoritmo de indexación [14].

3.1.3 An LLM-Based QA System for Chinese Painting and Calligraphy with Knowledge Graphs and External Documents

CAMPO	DETALLE
AUTORES	Jing Wan, Xinrong Li, Hao Zhang, Ao Zou y Rumei Wang
AÑO	2025
PUBLICACIÓN	npj Heritage Science
DOI / ENLACE	https://doi.org/10.1038/s40494-025-02219-3

Se desarrolla un sistema *QA* que combina grafos de conocimiento con documentos externos textuales para el dominio de pintura y caligrafía china. El sistema integra: (1) un módulo de construcción de grafo de conocimiento artístico chino con entidades (artistas, obras, estilos, períodos dinásticos, técnicas, materiales); (2) un corpus de documentos textuales especializados sobre pintura y caligrafía; (3) recuperación híbrida sobre ambas fuentes; (4) generación condicionada por el conocimiento recuperado. Se evalúa contra líneas base de LLMs sin RAG y sistemas RAG sin grafo de conocimiento [15].

3.1.4 Question Answering System Based on Retrieval-Augmented Generation

CAMPO	DETALLE
AUTORES	Yuqian Lin, Wenhui Zhang y Jiayi Lin
AÑO	2026
PUBLICACIÓN	Proc. ICCDC 2025
DOI / ENLACE	https://doi.org/10.1145/3783669.3783745

Se diseña e implementa un sistema de preguntas y respuestas para un humano digital (digital human) que integra un grafo de conocimiento con recuperación vectorial, desplegado en un entorno de museo virtual construido con *Unreal Engine*. El sistema utiliza un humano digital de alta fidelidad 3D como interfaz conversacional, creando una experiencia museística inmersiva. La metodología combina construcción de grafo de conocimiento artístico, indexación vectorial de colecciones, recuperación híbrida y generación condicionada [16].

3.1.5 A Personalized, Multimodal AI Assistant for Enhancing Museum Visitor Experience

CAMPO	DETALLE
AUTORES	Anmol Guragain, Jaime Bellver-Soler, Samuel Ramos-Varela, Long Lin, David Aragón Díaz y Luis Fernando D'Haro
AÑO	2025
PUBLICACIÓN	Proc. Workshop on Awareness in Learning Agents (ALA), CEUR-WS, vol. 4136
DOI / ENLACE	https://ceur-ws.org/Vol-4136/ala3.pdf

Se desarrolla un asistente de IA personalizado y multimodal para museos que integra RAG con capacidades de visión computacional. El sistema procesa tanto consultas textuales como visuales (imágenes de obras capturadas por el visitante), personaliza las respuestas según el perfil del usuario, y adapta el nivel de detalle y el estilo de comunicación a las preferencias del visitante. La evaluación incluye métricas de calidad de respuesta y satisfacción del usuario [17].

3.1.6 MuseRAG++: A Deep Retrieval-Augmented Generation Framework for Semantic Interaction and Multi-Modal Reasoning in Virtual Museums

CAMPO	DETALLE
AUTORES	Yifan Hu
AÑO	2026
PUBLICACIÓN	Scientific Reports
DOI / ENLACE	https://doi.org/10.1038/s41598-026-55700-9

Se propone MuseRAG++, un marco unificado de generación aumentada por recuperación que integra modelado profundo de la intención del usuario, un pipeline de recuperación híbrida disperso-densa, *re-ranking* semántico guiado por LLM, y un módulo de generación consciente de proveniencia. El sistema aborda los tres desafíos principales de los sistemas de diálogo museístico: comprensión de intención, recuperación efectiva y generación verificable [18].

3.1.7 Vision-Augmented RAG System for Interactive Local Heritage Exploration

CAMPO	DETALLE
AUTORES	Ritu Ram Ojha
AÑO	2025
PUBLICACIÓN	TechRxiv (preprint)
DOI / ENLACE	https://doi.org/10.36227/techrxiv.175691179.92356541/v1

Para consultas visuales, el asistente utiliza YOLO (*You Only Look Once*) para detección de objetos, identificando elementos patrimoniales en las imágenes capturadas por el usuario. Los objetos detectados se convierten en consultas textuales enriquecidas que alimentan el pipeline RAG. La base de datos vectorial contiene descripciones, metadatos históricos e información cultural del patrimonio local. El LLM genera respuestas que integran la información visual detectada con el conocimiento recuperado [19].

3.1.8 Evaluating Cultural Knowledge Processing in Large Language Models: A Cognitive Benchmarking Framework Integrating Retrieval-Augmented Generation

CAMPO	DETALLE
AUTORES	Hung-Shin Lee, Chen-Chi Chang, Ching-Yuan Chen y Yun-Hsiang Hsu
AÑO	2025
PUBLICACIÓN	arXiv:2511.01649
DOI / ENLACE	https://doi.org/10.48550/arXiv.2511.01649

Se propone un marco de benchmarking cognitivo que integra la Taxonomía de Bloom con RAG para evaluar el rendimiento de los LLMs en seis dominios cognitivos jerárquicos: Recordar, Comprender, Aplicar, Analizar, Evaluar y Crear. El marco utiliza un archivo digital de cultura Hakka taiwanesa como base de prueba primaria. Se evalúan múltiples LLMs con y sin RAG en las seis dimensiones cognitivas, midiendo precisión semántica y relevancia cultural de las respuestas [20].

3.1.9 How to Make Museums More Interactive? Case Study of Artistic Chatbot

CAMPO	DETALLE
AUTORES	Filip J. Kucia, Bartosz Grabek, Szymon D. Trochimiak y Anna Wróblewska
AÑO	2025
PUBLICACIÓN	Proc. ACM CIKM 2025, pp. 6654–6658
DOI / ENLACE	https://doi.org/10.48550/arXiv.2509.00572

Se presenta *Artistic Chatbot*, un sistema de chat RAG de voz a voz para apoyar el aprendizaje informal y mejorar el *engagement* del visitante durante una exposición de arte en vivo. El sistema fue desplegado durante la celebración del 15º aniversario de la Facultad de Arte Mediático de la Academia de Bellas Artes de Varsovia (Polonia). El estudio documenta el diseño, implementación y evaluación del sistema en un entorno de exposición real con visitantes reales [21].

3.1.10 RiverEcho: Real-Time Interactive Digital System for Ancient Yellow River Culture

CAMPO	DETALLE
AUTORES	Haofeng Wang, Yilin Guo, Zehao Li, Tong Yue, Yizong Wang, Enci Zhang, Rongqun Lin, Feng Gao, Shiqi Wang y Siwei Ma
AÑO	2025
PUBLICACIÓN	IEEE ICME Workshop 2025
DOI / ENLACE	https://doi.org/10.48550/arXiv.2506.21865

Se diseña RiverEcho, un sistema interactivo en tiempo real que responde a consultas de voz utilizando un LLM y una base de datos de conocimiento cultural, proporcionando explicaciones a través de un humano digital parlante (*talking-head digital human*). Se construye una base de datos de conocimiento enfocada en la cultura del Río Amarillo antiguo, incluyendo textos históricos y un pipeline de procesamiento. El sistema integra reconocimiento de voz, recuperación de conocimiento, generación de respuestas y síntesis de voz con animación facial sincronizada [22].

3.1.11Is AIGC Technology Useful? A Comparative Analysis of Search Technologies for Online Museums

CAMPO	DETALLE
AUTORES	Yurong Zeng, Chengxi Yan, Jiayi Li, Jingxuan Yao y Zhihan Sun
AÑO	2025
PUBLICACIÓN	Proc. Association for Information Science and Technology, vol. 62, pp. 1155–1160
DOI / ENLACE	https://doi.org/10.1002/pa2.1362

Se construye un sistema de búsqueda prototipo utilizando 800 pinturas de la plataforma de colección digital del Museo del Palacio de Beijing. Dieciséis usuarios realizaron tareas de búsqueda factual, explicativa y exploratoria, evaluando cinco métodos de búsqueda: búsqueda por palabras clave (KS) y cuatro optimizaciones de búsqueda generativa (RAG, COT, RAT y combinaciones). El diseño experimental es *within-subjects*, con los mismos usuarios evaluando todos los métodos, controlando por diferencias individuales [23].

3.1.12Humanities-in-the-Loop: Using Close Reading as a Method for Retrieval-Augmented Generation (RAG)

CAMPO	DETALLE
AUTORES	Jing Zhou, Li Si y Wenjun Hou
AÑO	2025
PUBLICACIÓN	Proc. Association for Information Science and Technology, vol. 62, núm. 1, pp. 1185–1189
DOI / ENLACE	https://doi.org/10.1002/pa2.1355

Se propone *Humanities-in-the-Loop*, un marco metodológico que incorpora la lectura cercana en cada etapa del pipeline RAG para mejorar el procesamiento de materiales de archivo digital. El marco incluye: anotación manual por humanistas, mantenimiento de conocimiento experto, validación por revisores, ingeniería de *prompts* con perspectiva humanística, e interpretación humana de los resultados. Como caso de estudio, se procesan los diarios de Coching Chu, destacado científico y educador chino del siglo XX [24].

3.1.13 Interactive and Provenance-Aware Search and QA over Documents Using LLMs, RAG and Knowledge Graph Verbalization

CAMPO	DETALLE
AUTORES	Iordanis Sapidis, Valantis Zervos, Michalis Mountantonakis y Yannis Tzitzikas
AÑO	2025
PUBLICACIÓN	New Trends in Theory and Practice of Digital Libraries, pp. 248–257
DOI / ENLACE	https://doi.org/10.1007/978-3-032-06136-2_24

Se desarrolla un sistema integrado de búsqueda interactiva y QA con conciencia de procedencia sobre documentos, combinando LLMs, RAG y grafos de conocimiento. El sistema permite a los usuarios rastrear cada afirmación en las respuestas generadas hasta su documento fuente específico, proporcionando trazabilidad completa. La integración de grafos de conocimiento enriquece la recuperación con relaciones estructuradas entre entidades documentales [25].

3.1.14 FolkRAG: A Retrieval-Augmented Generation System for Cultural Heritage Materials

CAMPO	DETALLE
AUTORES	Paul Kelly, Jonathan Schild y Amir Jafari
AÑO	2025
PUBLICACIÓN	Neural Computing and Applications, vol. 37, núm. 24, pp. 20281–20297
DOI / ENLACE	https://doi.org/10.1007/s00521-025-11455-4

Se desarrolla FolkRAG, un sistema RAG especializado para materiales de patrimonio cultural, con énfasis en folklore y tradiciones culturales. El sistema integra corpus de folklore digitalizado, recuperación semántica especializada y generación condicionada por el conocimiento folklórico recuperado. Se evalúa en múltiples tipos de consultas sobre folklore, desde preguntas factuales hasta análisis de variantes regionales [26].

3.1.15 An Intelligent Docent System with a Small Large Language Model (sLLM) Based on Retrieval-Augmented Generation (RAG)

CAMPO	DETALLE
AUTORES	Taemoon Jung e Inwhee Joe
AÑO	2025
PUBLICACIÓN	Applied Sciences, vol. 15, núm. 17, art. 9398
DOI / ENLACE	https://doi.org/10.3390/app15179398

Se diseña y evalúa empíricamente DocentGemma, un modelo de lenguaje especializado para museos basado en RAG y ajustado mediante LoRA sobre Gemma2-9B de Google. Se recopilan y refinan más de 199,000 descripciones de exposiciones, y se construye un *dataset* de QA de 102,000 pares que refleja personas de usuario reales. El sistema integra FAISS y OpenSearch con un Recuperador Espacio-Contextual (SC-Retriever) que combina similitud semántica y proximidad espacial. Se evalúa en tres museos de Seúl con 110 participantes reales [27].

3.1.16 Tabla comparativa

Con el propósito de sintetizar los hallazgos de la literatura y ofrecer una visión panorámica de las soluciones tecnológicas desarrolladas recientemente, en la Tabla 3.1 se recopilan y caracterizan los principales trabajos identificados en el estado del arte. Esta comparación se estructura en función de variables críticas para el diseño de sistemas de información, tales como el dominio específico de aplicación, la arquitectura del recuperador, el tipo de modelo generativo empleado, los conjuntos de datos utilizados para las pruebas, así como sus principales métricas, resultados clave y limitaciones identificadas.

Núm.	Trabajo	Síntesis comparativa
C1-01	Enhancement LLMs via OB-RAG	Dominio: Patrimonio Cultural Recuperador: Búsqueda semántica vectorial (triples ontológicos) Generador: LLM GPT-compatible Dataset: Ontologías culturales ad hoc Métricas: Faithfulness Answer Relevancy Resultados Clave: Faithfulness y Relevancy vs RAG estándar Limitaciones: Limitado a caso de estudio; escalabilidad no evaluada
C1-02	RAG-Optimized Tibetan Tourism LLMs	Dominio: Turismo Cultural (Tíbet) Recuperador: TF-IDF + BERT; HNSW, IVFlat, LSH Generador: Mistral-7B, LLaMA3-8b, Qwen Dataset: 563 puntos de vista turísticos Métricas: Tasa de aciertos; BERTScore; Puntuación compuesta Resultados Clave: Llama3-8b RAG: 92.25% (+24.18% sobre FT) Limitaciones: Dataset pequeño; Baichuan2-7b no mejoró
C1-03	LLM-Based QA Chinese Painting & Calligraphy	Dominio: Arte Chino Recuperador: Grafo de conocimiento + búsqueda semántica híbrida Generador: LLM de gran escala Dataset: Corpus pintura/caligrafía china Métricas: F1, EM, evaluación experta Resultados Clave: ↑ comprensión semántica; razonamiento multi-hop Limitaciones: Limitado al dominio de arte chino
C1-04	QA System Based on RAG (Digital Human Museum)	Dominio: Museos/Galerías Recuperador: Grafo de conocimiento + vectorial Generador: LLM generativo + TTS Dataset: Colección arte + grafo conocimiento Métricas: Precisión, satisfacción usuario Resultados Clave: ↑ búsqueda semántica vs palabras clave; inmersión 3D Limitaciones: Requiere infraestructura VR
C1-05	Multimodal AI Assistant Museum Visitor	Dominio: Museos (Experiencia Visitante) Recuperador: RAG multimodal (texto + imagen) Generador: LLM multimodal Dataset: Colección museo + imágenes obras Métricas: Recall contexto, Fidelidad, Corrección factual Resultados Clave: Recall@ctx: 0.96; Fidelidad: 0.98; Factual: 0.88 Limitaciones: Evaluado en colección específica
C1-06	MuseRAG++: Deep RAG for Semantic Museum Interaction	Dominio: Museos Virtuales Recuperador: Híbrido disperso-denso + re-ranking LLM Generador: GPT-4 / LLM avanzado Dataset: Colección museo multimodal Métricas: Recall@10, Precisión, F1 Resultados Clave: Recall@10: 83.1; Precisión: 76.4; F1: 79.6 Limitaciones: Escalabilidad a colecciones grandes no probada
C1-07	Vision-Augmented RAG for Local Heritage	Dominio: Patrimonio Local Recuperador: YOLO + RAG semántico Generador: LLM multimodal

Núm.	Trabajo	Síntesis comparativa
		Dataset: Corpus patrimonio local con imágenes Métricas: Precisión QA, relevancia semántica Resultados Clave: ↑ precisión con contexto visual integrado Limitaciones: Limitado a patrimonio local específico
C1-08	Evaluating Cultural Knowledge in LLMs (Hakka)	Dominio: Cultura Hakka (Taiwán) Recuperador: RAG sobre corpus cultural Hakka Generador: Múltiples LLMs evaluados Dataset: Benchmark cultura Hakka (Taxonomía Bloom) Métricas: 6 niveles cognitivos de Bloom Resultados Clave: RAG ↑ recuperación/aplicación; límites en tareas creativas Limitaciones: Benchmark específico; no generalizable directamente
C1-09	Artistic Chatbot for Museums	Dominio: Museos (Arte) Recuperador: Búsqueda semántica sobre colección artística Generador: LLM conversacional + TTS Dataset: Colección arte exposición Varsovia Métricas: Satisfacción usuario; calidad diálogo Resultados Clave: Chatbot voz-a-voz mejora interactividad museística Limitaciones: Evaluación cualitativa; sin métricas cuantitativas
C1-10	RiverEcho: Ancient Yellow River Culture	Dominio: Cultura Río Amarillo (China) Recuperador: RAG en tiempo real sobre corpus cultural fluvial Generador: LLM + TTS + Digital Human Dataset: Corpus cultura Río Amarillo Métricas: Precisión, relevancia, satisfacción usuario Resultados Clave: Sistema interactivo en tiempo real funcional Limitaciones: Dominio cultural muy específico
C1-11	AIGC: Comparative Analysis Search Museum	Dominio: Museos Digitales Recuperador: RAG, COT, RAT vs KS Generador: LLM con chain-of-thought Dataset: 800 pinturas Museo del Palacio Beijing Métricas: Satisfacción usuario; precisión búsqueda Resultados Clave: RAT mejor rendimiento general; RAG ↓ alucinaciones Limitaciones: Alucinaciones ocasionales; solo 16 usuarios
C1-12	Humanities-in-the-Loop RAG	Dominio: Archivos Históricos (Humanidades) Recuperador: RAG con lectura cercana humanística Generador: LLM + experto humanidades Dataset: Diarios Coching Chu (China) Métricas: Precisión interpretativa; evaluación experta Resultados Clave: ↑ precisión interpretativa con metodología humanística Limitaciones: Requiere expertise humanístico; no escalable fácilmente
C1-13	Provenance-Aware Search QA	Dominio: Archivos/Documentos Recuperador: RAG + grafo conocimiento con trazabilidad Generador: LLM con módulo proveniencia Dataset: Corpus documental con metadatos Métricas: Faithfulness, precisión proveniencia Resultados Clave: Respuestas con trazabilidad explícita a fuentes Limitaciones: Requiere metadatos de procedencia completos
C1-14	FolkRAG: Cultural Heritage Folklore	Dominio: Folklore/Patrimonio Intangible Recuperador: RAG especializado en folklore Generador: LLM generativo Dataset: Corpus folklore cultural Métricas: Precisión QA, relevancia cultural Resultados Clave: ↑ precisión en QA de folklore vs LLMs genéricos Limitaciones: Limitado al dominio de folklore
C1-15	DocentGemma: Intelligent Docent sLLM	Dominio: Museos (Seúl) Recuperador: SC-Retriever (semántico + espacial); FAISS + OpenSearch Generador: Gemma2-9B (LoRA) Dataset: 199K+ descripciones; 3 museos Seúl; 110 participantes Métricas: Recall@1 (53.45%), MRR (68.12%), satisfacción (4.24/5)

Núm.	Trabajo	Síntesis comparativa
		Resultados Clave: ↑ 17.5-20% precisión vs modelos base; satisfacción alta Limitaciones: Evaluación en 3 museos específicos; sin evaluación longitudinal

Tabla 3.1. Comparación de sistemas RAG y NLP aplicados a entornos culturales y museográficos.

Los trabajos revisados en esta sección muestran que las arquitecturas RAG son una opción viable para el acceso interactivo a información cultural, histórica y museográfica. Sin embargo, la mayoría de estas propuestas se enfoca en el diseño de asistentes, la personalización de la interacción o la incorporación de grafos de conocimiento y recursos multimodales, mientras que el análisis del módulo de recuperación suele quedar en segundo plano. Esta observación es importante porque, en sistemas RAG, la calidad de la respuesta generada no depende solo del modelo generativo, sino también de qué tan útil, completa y confiable sea la evidencia recuperada. Por ello, para comprender mejor las limitaciones actuales de estos sistemas, resulta necesario revisar la literatura sobre precisión de respuesta y reducción de alucinaciones en modelos de lenguaje.

3.2 Mitigación de alucinaciones en modelos de lenguaje

3.2.1 ChatTf: A Knowledge Graph-Enhanced Intelligent Q&A System for Mitigating Factuality Hallucinations in Traditional Folklore

CAMPO	DETALLE
AUTORES	Jun Xu, Hao Zhang, Haijing Zhang, Jiawei Lu y Gang Xiao
AÑO	2024
PUBLICACIÓN	IEEE Access, vol. 12, pp. 162638–162650
DOI / ENLACE	https://doi.org/10.1109/ACCESS.2024.3485877

Se presenta ChatTf, un sistema QA inteligente mejorado con grafos de conocimiento para mitigar alucinaciones de factuality en folklore tradicional chino. Se construye la Ontología de Folklore Tradicional (TFOnto) y se propone el marco de Extracción de Triples de Folklore *Zero-shot* (ZFTE). Impulsado por TFOnto, ZFTE construye el Grafo de Conocimiento de Folklore Tradicional (TFKG). Se implementa un sistema RAG de doble etapa que utiliza TFKG para recuperación estructurada y generación condicionada [28].

3.2.2 Reducing Hallucinations in LLMs: A Semantic Approach for Cultural Heritage

CAMPO	DETALLE
AUTORES	Fabio Clarizia, Massimo De Santo, Rosario Gaeta, Roberta Gimigliano y Rocco Loffredo
AÑO	2025
PUBLICACIÓN	Proc. i-Cities 2025
DOI / ENLACE	https://icities25.unicas.it/papers/43.pdf

Se propone un enfoque semántico para reducir alucinaciones en LLMs aplicados al patrimonio cultural. El enfoque integra representaciones semánticas enriquecidas con conocimiento de dominio cultural, mecanismos de verificación de factuality y estrategias de *prompting* diseñadas para minimizar la generación de información no fundamentada. Se evalúa en un caso de estudio de patrimonio cultural con métricas de *Faithfulness* y *Answer Relevancy* [29].

3.2.3 TravelRAG: A Tourist Attraction Retrieval Framework Based on Multi-Layer Knowledge Graph

CAMPO	DETALLE
AUTORES	Sihan Song, Chuncheng Yang, Li Xu, Haibin Shang, Zhuo Li y Yinghui Chang
AÑO	2024
PUBLICACIÓN	ISPRS International Journal of Geo-Information, vol. 13, núm. 11, art. 414
DOI / ENLACE	https://doi.org/10.3390/ijgi13110414

Se introduce TravelRAG, un marco de recuperación de atracciones turísticas basado en un grafo de conocimiento multicapa. El sistema extrae información de contenido generado por usuarios en redes sociales y la organiza en un grafo de conocimiento multicapa. RAG se integra para mejorar la precisión de las respuestas del LLM sobre atracciones turísticas, reduciendo alucinaciones y mejorando la personalización [30].

3.2.4 Hallucination Mitigation Techniques in Large Language Models

CAMPO	DETALLE
AUTORES	Mohamed Ali, Mohamed Mabrouk y Zaki Taha
AÑO	2024
PUBLICACIÓN	International Journal of Intelligent Computing and Information Sciences, vol. 24, núm. 4, pp. 73–81
DOI / ENLACE	https://doi.org/10.21608/ijicis.2024.336135.1365

Se realiza una revisión sistemática de las técnicas de mitigación de alucinaciones en LLMs, proporcionando una taxonomía comprehensiva y comparación de efectividad. El estudio analiza técnicas pre-generación (RAG, ajuste fino), durante la generación (*prompting*, *beam search* modificado) y post-generación (verificación, *fact-checking*). Se evalúa la efectividad de cada técnica en múltiples *benchmarks* [31].

3.2.5 Factuality of Large Language Models: A Survey

CAMPO	DETALLE
AUTORES	Yuxia Wang, Minghan Wang, Muhammad Arslan Manzoor, Fei Liu, Georgi Georgiev, Rocktim Jyoti Das y Preslav Nakov
AÑO	2024
PUBLICACIÓN	arXiv:2402.02420
DOI / ENLACE	https://doi.org/10.48550/arXiv.2402.02420

Se realiza una revisión comprehensiva de la factuality de los LLMs en 2024, analizando múltiples dimensiones: tipos de alucinaciones, *benchmarks* de evaluación, estrategias de mitigación (RAG, ajuste fino, verificación) y rendimiento de los principales modelos. El estudio proporciona un panorama completo del estado del arte en factuality de LLMs con énfasis en estrategias de mejora [32].

3.2.6 Towards Reliable Medical Question Answering: Techniques and Challenges in Mitigating Hallucinations in Language Models

CAMPO	DETALLE
AUTORES	Duy Khoa Pham y Bao Quoc Vo
AÑO	2024
PUBLICACIÓN	arXiv:2408.13808
DOI / ENLACE	https://doi.org/10.48550/arXiv.2408.13808

Se revisan sistemáticamente las técnicas y desafíos para la mitigación de alucinaciones en LLMs aplicados a QA médico. El estudio analiza estrategias de RAG médico, ajuste fino con datos de alta calidad, verificación de factuality mediante bases de conocimiento médico, y mecanismos de calibración de confianza. Se evalúa el rendimiento de estas estrategias en *benchmarks* médicos estándar [33].

3.2.7 Mitigating Large Language Model Hallucination with Faithful Finetuning

CAMPO	DETALLE
AUTORES	Minda Hu, Bowei He, Yufei Wang, Liangyou Li, Chen Ma e Irwin King
AÑO	2024
PUBLICACIÓN	arXiv:2406.11267
DOI / ENLACE	https://doi.org/10.48550/arXiv.2406.11267

Se propone un método de ajuste fino fiel que incorpora restricciones de factualidad en el proceso de entrenamiento de LLMs. El método incluye: (1) selección de datos de entrenamiento de alta calidad factual; (2) función de pérdida que penaliza explícitamente las alucinaciones; (3) regularización que preserva el conocimiento factual aprendido durante el preentrenamiento; (4) evaluación en *benchmarks* de factualidad y comparación con ajuste fino estándar [34].

3.2.8 What Are Models Thinking About? Understanding Large Language Model Hallucinations ‘Psychology’ Through Model Inner State Analysis

CAMPO	DETALLE
AUTORES	Peiran Wang, Yang Liu, Yunfei Lu, Jue Hong y Ye Wu
AÑO	2025
PUBLICACIÓN DOI / ENLACE	arXiv:2502.13490 https://doi.org/10.48550/arXiv.2502.13490

Se propone un marco de análisis del estado interno de los LLMs durante la generación para comprender los mecanismos psicológicos de las alucinaciones. El estudio analiza representaciones internas, activaciones de atención y patrones de distribución de probabilidad en tokens asociados con alucinaciones versus respuestas factuales correctas. Los hallazgos informan el diseño de nuevas estrategias de mitigación [35].

3.2.9 Survey on Factuality in Large Language Models: Knowledge, Retrieval and Domain-Specificity

CAMPO	DETALLE
AUTORES	Cunxiang Wang, Xiaoze Liu, Yuanhao Yue, Xiangru Tang, Tianhang Zhang et al.
AÑO	2023
PUBLICACIÓN DOI / ENLACE	arXiv:2310.07521 https://doi.org/10.48550/arXiv.2310.07521

Se realiza una revisión sistemática de 150+ trabajos sobre factualidad en LLMs, organizados en tres dimensiones analíticas: (1) tipos de conocimiento que los LLMs necesitan para ser factuales (conocimiento paramétrico vs. no paramétrico, conocimiento de sentido común vs. especializado, conocimiento estático vs. dinámico); (2) estrategias de recuperación y acceso al conocimiento externo (RAG, búsqueda web, bases de datos estructuradas, grafos de conocimiento); (3) especificidad de dominio y sus implicaciones para la factualidad (dominios abiertos vs. cerrados, dominios con verdad verificable vs. subjetiva). El análisis

incluye *benchmarks* existentes, métricas de evaluación, y una comparación cuantitativa de la efectividad de diferentes estrategias de mejora de factualidad [36].

3.2.10Hallucination Reduction in Large Language Models with Retrieval-Augmented Generation Using Wikipedia Knowledge

CAMPO	DETALLE
AUTORES	Jason Kirchenbauer y Caleb Barns
AÑO	2024
PUBLICACIÓN	OSF Preprints
DOI / ENLACE	https://doi.org/10.31219/osf.io/pv7r5

Se implementa un sistema RAG que recupera pasajes relevantes de Wikipedia en tiempo de inferencia y los proporciona como contexto al LLM antes de la generación. El diseño experimental compara cuatro condiciones: (1) LLM base sin RAG, (2) LLM con RAG sobre Wikipedia (recuperación densa), (3) LLM con RAG sobre Wikipedia (recuperación dispersa BM25), y (4) LLM con RAG sobre Wikipedia (recuperación híbrida). La evaluación se realiza en múltiples *benchmarks* de factualidad con evaluación automática y humana. El análisis incluye estratificación por tipo de consulta (factual directa, comparativa, temporal, sobre entidades de larga cola) para identificar en qué contextos RAG es más beneficioso [37].

3.2.11A Survey of Hallucination Problems Based on Large Language Models

CAMPO	DETALLE
AUTORES	Xinxin Liu
AÑO	2024
PUBLICACIÓN	Applied and Computational Engineering, vol. 97, pp. 24–30
DOI / ENLACE	https://doi.org/10.54254/2755-2721/97/2024.17851

Revisión sistemática de la literatura sobre alucinaciones en LLMs, cubriendo trabajos desde 2020 hasta 2024. La metodología incluye: (1) identificación y clasificación de tipos de alucinaciones mediante análisis de casos documentados en la literatura; (2) análisis de causas raíz agrupadas en categorías de datos de entrenamiento, arquitectura del modelo y proceso de inferencia; (3) revisión y comparación de técnicas de mitigación con análisis de efectividad reportada; (4) propuesta de una taxonomía unificada que integra perspectivas de

diferentes comunidades de investigación. El *survey* cubre 80+ trabajos sobre tipos de alucinaciones, 60+ trabajos sobre causas y 100+ trabajos sobre técnicas de mitigación [38]

.

3.2.12 Tabla comparativa

Con el propósito de examinar las soluciones propuestas para abordar una de las limitaciones más críticas en los sistemas de generación de texto, en la Tabla 3.2 se recopilan y contrastan los trabajos de la literatura orientados específicamente a la reducción del error fáctico. Esta selección incluye tanto implementaciones en dominios especializados como estudios globales (*surveys*), comparándolos a partir del enfoque de su recuperador, el modelo generativo base, los conjuntos de datos de prueba implementados, sus métricas de evaluación, así como los hallazgos más significativos y las limitantes técnicas documentadas por los autores.

Núm.	Trabajo	Síntesis comparativa
1	ChatTf: KG-Enhanced QA (Folklore)	Dominio: Folklore Chino Recuperador: Navegación TFKG (grafo folklore) Generador: LLM condicionado por TFKG Dataset: TFCQD (folklore tradicional) Métricas: Precisión QA; tasa alucinación Resultados Clave: Precisión 96.7% en TFCQD; ↓ alucinaciones Limitaciones: Limitado a folklore chino; KG costoso de construir
2	Reducing Hallucinations: Semantic Approach Cultural Heritage	Dominio: Patrimonio Cultural Recuperador: Recuperación semántica enriquecida con ontologías Generador: LLM con prompting semántico Dataset: Corpus patrimonio cultural Métricas: Faithfulness, Answer Relevancy Resultados Clave: ↑ Faithfulness; ↓ alucinaciones vs LLM base Limitaciones: Dominio cultural específico; evaluación dependiente de expertos
3	TravelRAG: Tourist Attraction Retrieval (KG Multicapa)	Dominio: Turismo Cultural Recuperador: Grafo de conocimiento multicapa (redes sociales) Generador: LLM condicionado por KG multicapa Dataset: Datos turísticos de redes sociales Métricas: Eficiencia recuperación, precisión recomendaciones Resultados Clave: ↑ eficiencia y precisión; ↓ costo vs fine-tuning Limitaciones: Sesgos de redes sociales; actualización periódica requerida
4	Factuality of LLMs in 2024 (Survey)	Dominio: NLP General Recuperador: Múltiples estrategias RAG Generador: Múltiples LLMs Dataset: TruthfulQA, FactScore, FEVER Métricas: TruthfulQA, FactScore, FEVER Resultados Clave: RAG es la estrategia más efectiva para factuality Limitaciones: Survey; sin evaluación experimental propia
5	Reliable Medical QA: Hallucination Mitigation	Dominio: Medicina/Salud Recuperador: RAG médico (PubMed, UMLS, SNOMED CT) Generador: LLM médico (BioGPT, MedPaLM) Dataset: MedQA, PubMedQA, MedMCQA Métricas: Accuracy médica, Faithfulness, calibración Resultados Clave: RAG médico ↑ fiabilidad; técnicas de verificación factual Limitaciones: Dominio médico; consecuencias directas en salud
6	Faithful Finetuning for Hallucination Mitigation	Dominio: NLP General Recuperador: RAG como estrategia complementaria Generador: LLM con ajuste fino fiel Dataset: TruthfulQA, FEVER, HaluEval Métricas: Faithfulness, Hallucination Rate Resultados Clave: Ajuste fino fiel ↓ alucinaciones; RAG complementario Limitaciones: Requiere datos de FT de alta calidad factual

Núm.	Trabajo	Síntesis comparativa
7	What are Models Thinking? LLM Hallucination Psychology	Dominio: NLP General Recuperador: RAG activado por señales de incertidumbre interna Generador: Múltiples LLMs analizados Dataset: Benchmarks factualidad + análisis internos Métricas: Tasa alucinación; análisis estados internos Resultados Clave: ↑ comprensión mecanismos de alucinación; RAG condicional Limitaciones: Análisis costoso; patrones varían entre arquitecturas
8	Hallucination Mitigation Techniques in LLMs (Survey)	Dominio: NLP General Recuperador: RAG + técnicas de grounding Generador: Múltiples LLMs Dataset: TruthfulQA, FEVER, HaluEval, FactScore Métricas: Hallucination Rate, Faithfulness, ROUGE Resultados Clave: RAG más efectivo; CoVe complementario; taxonomía completa Limitaciones: Survey; sin técnica nueva propuesta
9	Survey on Factuality in LLMs (Knowledge & Retrieval)	Dominio: NLP General Recuperador: RAG como estrategia principal Generador: Múltiples LLMs Dataset: FactScore, TruthfulQA, FEVER Métricas: FactScore, TruthfulQA, FEVER Resultados Clave: RAG es la estrategia más efectiva para factualidad Limitaciones: Survey; sin evaluación experimental propia
10	Hallucination Reduction with RAG (Wikipedia)	Dominio: NLP General Recuperador: RAG sobre Wikipedia Generador: LLM generativo Dataset: Benchmarks factualidad Métricas: Hallucination Rate, Faithfulness Resultados Clave: RAG ↓ alucinaciones 35-45% vs LLM base Limitaciones: Efectividad depende de calidad del corpus recuperado
11	Survey of Hallucination Problems in LLMs	Dominio: NLP General Recuperador: Múltiples estrategias RAG Generador: Múltiples LLMs Dataset: Benchmarks factualidad múltiples Métricas: Hallucination Rate, F1, Faithfulness Resultados Clave: Taxonomía comprehensiva tipos de alucinaciones; RAG principal Limitaciones: Survey; sin evaluación experimental propia
12	Hallucination Mitigation for RAG LLMs (Math)	Dominio: NLP General / Matemáticas Recuperador: RAG con verificación formal Generador: LLM con verificación matemática Dataset: Benchmarks matemáticos y factualidad Métricas: Hallucination Rate, Faithfulness Resultados Clave: ↑ Faithfulness 28% con marco matemático propuesto Limitaciones: Limitado a dominios con verdad formal verificable
13	HALO: Hallucination Analysis and Learning Optimization	Dominio: NLP General Recuperador: RAG + aprendizaje optimizado anti-alucinación Generador: LLM con HALO Dataset: Benchmarks factualidad múltiples Métricas: Hallucination Rate, ROUGE, BERTScore Resultados Clave: HALO ↓ alucinaciones 31% en benchmarks estándar Limitaciones: Requiere datos de entrenamiento específicos para HALO

Tabla 3.2. Comparación de enfoques para la mitigación de alucinaciones en modelos de lenguaje y sistemas RAG.

3.3 Discusión

La revisión realizada permite identificar dos tendencias principales en la literatura reciente. Por una parte, los sistemas basados en RAG se han consolidado como una alternativa para

facilitar el acceso a información especializada en dominios como el patrimonio cultural, los museos, los archivos históricos y el turismo cultural. Estos trabajos muestran la capacidad de combinar modelos de lenguaje con mecanismos de recuperación para ofrecer respuestas contextualizadas a partir de colecciones documentales específicas.

Por otra parte, la literatura sobre mitigación de alucinaciones coincide en que la incorporación de evidencia externa durante el proceso de generación constituye una de las estrategias más efectivas para mejorar la fidelidad fáctica de las respuestas. Diversos estudios señalan que la calidad de la información recuperada influye directamente en el desempeño de los sistemas RAG, por lo que la selección de estrategias de recuperación representa un elemento relevante dentro de estas arquitecturas.

En conjunto, ambos núcleos temáticos muestran que las arquitecturas RAG han evolucionado hacia sistemas capaces de integrar diferentes mecanismos de recuperación y modelos generativos para responder consultas sobre información especializada. Asimismo, evidencian que la recuperación de información y la generación de respuestas no deben considerarse procesos independientes, sino componentes complementarios cuyo funcionamiento conjunto determina la calidad de las respuestas producidas por el sistema.

Capítulo IV. Generación aumentada por recuperación

4.1 Introducción a la recuperación aumentada por generación

El surgimiento de los LLM ha transformado profundamente las posibilidades de interacción entre los sistemas computacionales y el lenguaje humano. Sin embargo, a pesar de su notable capacidad para generar texto fluido y contextualmente coherente, estos modelos enfrentan una limitación estructural fundamental: su conocimiento queda congelado en el momento del entrenamiento. Esto significa que no pueden incorporar información nueva sin un proceso de reentrenamiento costoso, y que su desempeño tiende a degradarse cuando se les consulta sobre eventos recientes, datos especializados o información de largo alcance que no estuvo bien representada en los datos de preentrenamiento.

Frente a esta limitación, se propuso una arquitectura llamada RAG que en español puede traducirse como Generación Aumentada por Recuperación[2]. La propuesta central de este trabajo consiste en combinar, dentro de un mismo marco unificado, las capacidades generativas de un modelo de lenguaje con un componente de recuperación de información capaz de acceder a una base de conocimiento externa en tiempo de inferencia. De este modo, el modelo no depende exclusivamente de lo que aprendió durante el entrenamiento, sino que puede consultar documentos externos relevantes antes de producir una respuesta, lo que le permite fundamentar sus salidas en evidencia textual concreta y actualizable.

Desde su publicación, el trabajo de Lewis et al.[2] ha dado lugar a una extensa línea de investigación. Gao et al. [39] presentaron un *survey* comprehensivo que documenta la evolución del paradigma RAG en tres etapas: RAG *naive*, RAG avanzado y RAG modular. En la etapa *naive*, el sistema simplemente recupera documentos relevantes y los concatena al *prompt* del modelo generativo. En la etapa avanzada se incorporan mecanismos de preprocesamiento de la consulta, reordenamiento de resultados y filtrado de evidencia. En la etapa modular, los distintos componentes del sistema (recuperación, reordenamiento, generación y memoria) se diseñan como módulos intercambiables que pueden optimizarse de forma independiente.

La relevancia del paradigma RAG no se limita a resolver el problema de la actualización del conocimiento. También aborda, de manera directa, el fenómeno de las alucinaciones, que se describirá con detalle en secciones posteriores. Al anclar la generación a evidencia documental recuperada, el modelo tiene menos necesidad de recurrir a inferencias probabilísticas no fundamentadas. Este principio ha sido validado empíricamente en numerosos trabajos posteriores, que muestran que los sistemas RAG producen respuestas más precisas y verificables que los modelos generativos aislados en una amplia variedad de dominios [40][41].

4.2 Arquitectura general de un sistema RAG

Un sistema RAG típico se organiza en torno a tres procesos centrales e interdependientes: la recuperación (*retrieval*), la aumentación (*augmentation*) y la generación (*generation*). Esta estructura tripartita fue descrita originalmente por Lewis et al. [2] y ha sido adoptada como marco de referencia en la literatura subsecuente [39][42].

El proceso de **recuperación** tiene como objetivo localizar, dentro de una base de conocimiento externa, los fragmentos de texto más relevantes para una consulta dada. Esta base de conocimiento puede adoptar distintas formas: una colección de documentos en texto libre, un conjunto de artículos de enciclopedia, un repositorio de registros estructurados o cualquier corpus documental que haya sido previamente indexado. La consulta formulada por el usuario se transforma en una representación computacional ya sea una bolsa de términos ponderados o un vector denso en un espacio de alta dimensión que se compara contra las representaciones de los fragmentos almacenados en el índice. Los fragmentos cuya representación es más similar a la de la consulta son seleccionados como evidencia candidata.

El proceso de **aumentación** consiste en integrar la evidencia recuperada dentro del contexto de entrada (*prompt*) que se proporcionará al modelo generativo. En su forma más directa, los fragmentos recuperados se concatenan a la consulta original antes de ser enviados al modelo. En implementaciones más sofisticadas, se aplican mecanismos de reordenamiento, filtrado y síntesis para construir un contexto más coherente y relevante. La calidad de esta etapa es crítica, puesto que el modelo generativo solo puede aprovechar la información que se le proporciona explícitamente en su ventana de contexto.

El proceso de **generación** corresponde al modelo de lenguaje propiamente dicho, que recibe la consulta aumentada con la evidencia recuperada y produce una respuesta en lenguaje natural. En la formulación original de Lewis et al. [2], se utilizó BART como modelo generativo, pero en implementaciones contemporáneas se emplean modelos de mayor escala como GPT-4, LLaMA o Mistral, entre otros. El modelo generativo puede operar en dos modos principales: en el modo *RAG-Sequence*, el mismo conjunto de documentos recuperados se utiliza para generar la respuesta completa; en el modo *RAG-Token*, el proceso de recuperación puede actualizarse token a token, lo que permite una mayor granularidad en el condicionamiento de la generación.

Más allá de estos tres procesos nucleares, un sistema RAG completo incluye componentes adicionales que son indispensables para su funcionamiento. En la fase de indexación, los documentos de la base de conocimiento se fragmentan, se representan computacionalmente y se almacenan en estructuras de datos que permiten una búsqueda eficiente. Esta fase es previa a la operación del sistema y puede involucrar tanto índices invertidos como índices vectoriales, dependiendo del tipo de recuperación que se desee realizar. El diseño de estos componentes tiene un impacto directo sobre la calidad de la evidencia que el sistema es capaz de recuperar, y por tanto sobre la calidad de las respuestas generadas [43].

4.3 Conocimiento paramétrico y conocimiento no paramétrico

Una distinción conceptual fundamental para comprender el funcionamiento de los sistemas RAG es la que existe entre el conocimiento paramétrico y el conocimiento no paramétrico en los modelos de lenguaje.

El **conocimiento paramétrico** es aquel que un modelo de lenguaje almacena implícitamente en sus pesos durante el proceso de preentrenamiento. Cuando un modelo como GPT-4 o LLaMA es entrenado sobre miles de millones de tokens de texto, aprende distribuciones de probabilidad sobre secuencias de palabras que codifican, de manera distribuida, una enorme cantidad de información sobre el mundo: hechos históricos, relaciones entre entidades, patrones lingüísticos, convenciones discursivas y conocimiento de dominio. Este conocimiento es paramétrico porque reside en los parámetros del modelo (sus pesos) y no en ninguna estructura de datos externa accesible en tiempo de inferencia. Se demostró de manera sistemática que los modelos de lenguaje preentrenados pueden funcionar como bases de conocimiento implícitas, siendo capaces de recuperar hechos a partir de consultas formuladas en lenguaje natural, sin acceso a ninguna fuente externa [44].

Sin embargo, el conocimiento paramétrico tiene limitaciones estructurales importantes. En primer lugar, es estático: una vez que el modelo ha sido entrenado, su conocimiento no puede actualizarse sin un nuevo proceso de entrenamiento o ajuste fino, lo que lo hace inadecuado para dominios donde la información cambia con frecuencia. En segundo lugar, es opaco: no es posible identificar con precisión de qué fuentes proviene un hecho particular que el modelo genera, lo que dificulta su verificación. En tercer lugar, tiende a ser impreciso en dominios de larga cola, es decir, en aquellos temas que estuvieron escasamente representados en los datos de entrenamiento [40].

El **conocimiento no paramétrico**, por contraste, es aquel que se almacena explícitamente en estructuras de datos externas al modelo (bases de datos, índices documentales, grafos de conocimiento) y que puede consultarse en tiempo de inferencia. Este tipo de conocimiento es dinámico, actualizable y trazable: es posible saber exactamente qué documento respaldó una afirmación particular. La arquitectura RAG es, en esencia, una forma de combinar el conocimiento paramétrico del modelo generativo (su capacidad para producir texto fluido y coherente) con el conocimiento no paramétrico almacenado en una base documental externa [2].

4.4 El fenómeno de las alucinaciones en modelos de lenguaje

Una de las limitaciones más documentadas de los LLM es su tendencia a generar contenido que resulta lingüísticamente fluido y sintácticamente correcto, pero que es factualmente incorrecto o completamente inventado. Este fenómeno, conocido como alucinación, ha recibido una atención creciente debido a la incorporación de los LLM en aplicaciones que requieren un alto grado de precisión y confiabilidad [45].

El término alucinación fue introducido en el ámbito del procesamiento de lenguaje natural por Maynez et al [46], quienes lo emplearon para describir la generación de información que no puede inferirse ni verificarse a partir de la fuente de entrada en tareas de resumen abstractivo. Posteriormente, el concepto se amplió para abarcar cualquier contenido generado por un modelo de lenguaje que carezca de respaldo en evidencia verificable, independientemente de la tarea realizada.

Desde una perspectiva técnica, una alucinación puede definirse como la generación de texto que es internamente coherente, pero externamente infiel, es decir, que no corresponde con hechos verificables, contradice la información disponible o incorpora afirmaciones sin sustento documental [45]. Esta definición diferencia las alucinaciones de otros tipos de errores, como los problemas gramaticales o de coherencia discursiva, al situarlas específicamente en la relación entre la respuesta generada y la información que debería respaldarla.

La incidencia de este fenómeno es significativa. Ji et al [45], señalan que las alucinaciones afectan, en distinta medida, a prácticamente todos los modelos de lenguaje actuales y que su frecuencia aumenta en tareas que requieren conocimiento especializado, razonamiento de múltiples pasos o información poco frecuente. Asimismo, Zhang et al[47] proponen una taxonomía que distingue dos categorías principales: las **alucinaciones factuales**, en las que el modelo afirma hechos incorrectos acerca del mundo, y las **alucinaciones de fidelidad**, en las que la respuesta resulta inconsistente con el contexto o con las fuentes proporcionadas.

Las causas de las alucinaciones pueden agruparse en tres categorías. La primera está relacionada con los datos de entrenamiento. Si el corpus contiene información incorrecta, sesgada o contradictoria, el modelo puede aprender estos patrones y reproducirlos durante la generación. Además, la representación desigual de los distintos dominios provoca que algunos temas sean modelados con mayor precisión que otros. La segunda categoría corresponde a la arquitectura y al proceso de entrenamiento. Los LLM están optimizados para predecir la siguiente palabra con mayor probabilidad, pero su función objetivo no incorpora restricciones explícitas sobre la veracidad del contenido generado. Como consecuencia, el modelo puede producir respuestas estadísticamente plausibles, aunque sean incorrectas desde el punto de vista factual. Finalmente, la tercera categoría se relaciona con el proceso de inferencia. Cuando una consulta requiere información que no está representada adecuadamente en los parámetros del modelo, este tiende a completar la respuesta con contenido plausible, en lugar de reconocer la ausencia de conocimiento.

Las consecuencias de las alucinaciones son especialmente relevantes en aplicaciones donde la precisión factual es indispensable. En el ámbito médico, una respuesta incorrecta sobre la dosis de un medicamento o las contraindicaciones de un tratamiento puede afectar directamente la seguridad del paciente [48]. En el ámbito legal, la generación de referencias

jurídicas inexistentes compromete la validez de los argumentos presentados. De igual forma, en aplicaciones educativas y culturales, la producción de información histórica incorrecta puede distorsionar la interpretación del patrimonio y la memoria colectiva.

4.5 Estrategias para reducir alucinaciones

La literatura ha propuesto un amplio abanico de estrategias para mitigar el fenómeno de las alucinaciones, que pueden clasificarse según el momento del ciclo de vida del modelo en el que intervienen: estrategias de preentrenamiento, de ajuste fino (*fine-tuning*), de inferencia y de postprocesamiento [45].

Las estrategias de preentrenamiento buscan mejorar la calidad y la representatividad de los datos con los que se entrena el modelo. Esto incluye la depuración del corpus de entrenamiento para eliminar información errónea o sesgada, la aplicación de técnicas de balanceo para garantizar una cobertura más uniforme de los dominios de conocimiento, y la incorporación de señales de veracidad durante el proceso de aprendizaje. Sin embargo, estas estrategias son costosas y no eliminan completamente el problema, puesto que ningún corpus puede ser exhaustivo ni libre de errores.

Las estrategias de ajuste fino buscan especializar el modelo para que sea más fiel a las fuentes de información en tareas específicas. El ajuste fino con retroalimentación humana (*RLHF*, *Reinforcement Learning from Human Feedback*) ha demostrado ser efectivo para reducir ciertos tipos de alucinaciones, al incorporar preferencias humanas sobre la calidad y la veracidad de las respuestas generadas [49].

Las estrategias de inferencia constituyen el conjunto más amplio y diverso. Dentro de este grupo, la más relevante para el paradigma RAG es el acceso a conocimiento externo: al proporcionar al modelo fragmentos de texto recuperados de fuentes confiables, se reduce su necesidad de recurrir a conocimiento paramétrico potencialmente impreciso. Shuster et al. [41] demostraron que el acceso a información recuperada reduce significativamente las alucinaciones en sistemas de diálogo, y Mallen et al. [40] mostraron que esta reducción es especialmente pronunciada para entidades de larga cola, precisamente donde el conocimiento paramétrico es más débil. Otras estrategias de inferencia incluyen el *chain-of-thought prompting* [50], que induce al modelo a producir razonamientos intermedios antes de llegar a una conclusión, y las técnicas de calibración de confianza, que permiten al modelo expresar incertidumbre en lugar de generar respuestas falsamente seguras.

Las estrategias de postprocesamiento intervienen después de que el modelo ha generado una respuesta, para verificar su veracidad y corregirla si es necesario. Esto incluye sistemas de detección de alucinaciones que comparan el texto generado contra fuentes externas, así como mecanismos de verificación de hechos (*fact-checking*) que evalúan cada afirmación del texto generado de manera independiente [51]. Aunque estas estrategias son prometedoras, su

implementación práctica es costosa en términos computacionales y no siempre es posible aplicarlas en tiempo real.

En conjunto, la evidencia disponible sugiere que ninguna estrategia aislada es suficiente para eliminar completamente las alucinaciones, y que los sistemas más robustos combinan múltiples enfoques.

4.6 Recuperación de información dentro de RAG

En el contexto de los sistemas RAG, la recuperación de información desempeña un papel central: es el componente que determina qué evidencia estará disponible para el modelo generativo al momento de producir una respuesta. La calidad de esta evidencia (su relevancia, precisión y cobertura) condiciona directamente la calidad de la respuesta final. Un recuperador que selecciona fragmentos irrelevantes o ruidosos proporcionará al modelo generativo una base débil sobre la cual construir su respuesta, lo que aumenta el riesgo de alucinaciones o de respuestas imprecisas. Por el contrario, un recuperador que selecciona fragmentos altamente relevantes y concisos proporciona al modelo la evidencia necesaria para generar respuestas fundamentadas y verificables [43][52].

Los métodos de recuperación pueden clasificarse en dos grandes familias: los métodos de recuperación dispersa (*sparse retrieval*), basados en representaciones léxicas de documentos y consultas, y los métodos de recuperación densa (*dense retrieval*), basados en representaciones vectoriales continuas. Cada familia tiene ventajas y limitaciones complementarias, y la combinación de ambas (conocida como recuperación híbrida) ha emergido como la estrategia más robusta en la literatura reciente. Estas tres familias se describen en detalle en las secciones 4.7, 4.8 y 4.9 de este capítulo.

4.7 Recuperación dispersa

Los métodos de recuperación dispersa se basan en representaciones léxicas de documentos y consultas, en las que cada dimensión del espacio de representación corresponde a un término del vocabulario. La representación de un documento es «dispersa» porque la mayoría de sus componentes son cero: un documento típico utiliza solo una pequeña fracción del vocabulario total de la colección.

El modelo más influyente en la historia de la recuperación de información es el Modelo de Espacio Vectorial (*Vector Space Model*, VSM) propuesto por Salton et al. [7], que representa documentos y consultas como vectores en un espacio de alta dimensión y calcula su similitud mediante el producto escalar o el coseno del ángulo entre ellos. La ponderación de los términos en el VSM clásico se realiza mediante el esquema TF-IDF (*Term Frequency-Inverse Document Frequency*), que asigna mayor peso a los términos que son frecuentes en un documento, pero raros en la colección, lo que los hace más discriminativos [53].

Sin embargo, el modelo de recuperación léxica que ha demostrado mayor efectividad empírica y que se ha convertido en el estándar de facto en la industria y la investigación es BM25 (*Best Match 25*), propuesto por Robertson et al.[54] en el marco del modelo probabilístico de recuperación. BM25 es una función de puntuación que estima la relevancia de un documento d con respecto a una consulta q mediante la siguiente fórmula:

$$\text{BM25}(d, q) = \sum_{t \in q} \text{IDF}(t) \cdot \frac{\text{TF}(t, d) \cdot (k_1 + 1)}{\text{TF}(t, d) + k_1 \cdot \left(1 - b + b \cdot \frac{|d|}{\text{avgdl}}\right)} \quad (2)$$

donde $\text{TF}(t, d)$ es la frecuencia del término t en el documento d , $|d|$ es la longitud del documento, avgdl es la longitud promedio de los documentos de la colección, y k_1 y b son parámetros de ajuste. El término $\text{IDF}(t)$ es la frecuencia inversa de documento del término, que penaliza los términos que aparecen en muchos documentos de la colección.

La fortaleza de BM25 radica en su robustez y eficiencia: ha demostrado ser difícil de superar de manera consistente en *benchmarks* de recuperación estándar, incluso frente a métodos mucho más complejos. Su implementación es eficiente gracias a los índices invertidos, que permiten calcular las puntuaciones BM25 sin examinar todos los documentos de la colección. Además, BM25 es interpretable: es posible entender por qué un documento recibió una puntuación alta o baja, lo que facilita la depuración y el ajuste del sistema.

La limitación principal de BM25 y de los métodos léxicos en general es el problema del vocabulario (*vocabulary mismatch*): si la consulta y el documento relevante no comparten términos en común, el recuperador léxico no podrá encontrar ese documento, independientemente de su relevancia semántica. Este problema es especialmente agudo en dominios donde los usuarios utilizan terminología diferente a la de los documentos, por ejemplo, cuando un usuario formula una consulta en lenguaje coloquial y los documentos están escritos en lenguaje técnico, o cuando existen múltiples sinónimos para el mismo concepto [55].

4.8 Recuperación densa

La recuperación densa surgió como respuesta directa a las limitaciones de los métodos léxicos, al proponer que documentos y consultas sean representados como vectores densos en un espacio de alta dimensión, y que la relevancia se calcule como la similitud geométrica entre estos vectores. A diferencia de las representaciones dispersas, donde cada dimensión corresponde a un término del vocabulario, las representaciones densas no tienen una interpretación dimensional directa, sino que codifican el significado de manera distribuida a lo largo de todas sus dimensiones.

El trabajo fundacional de la recuperación densa moderna es el modelo DPR (*Dense Passage Retrieval*), propuesto por Karpukhin et al. [43]. DPR utiliza dos codificadores independientes

basados en BERT: uno para codificar la consulta y otro para codificar los pasajes de texto. Ambos codificadores son entrenados de manera conjunta mediante aprendizaje contrastivo, utilizando pares de consultas y pasajes relevantes como ejemplos positivos y pasajes no relevantes como ejemplos negativos. El objetivo del entrenamiento es que la representación vectorial de una consulta sea similar a la representación de los pasajes relevantes, y disimilar a la de los pasajes no relevantes.

DPR demostró mejoras significativas sobre BM25 en el *benchmark* de preguntas y respuestas de dominio abierto *Natural Questions* Kwiatkowski et al., [56], estableciendo la recuperación densa como una alternativa competitiva a los métodos léxicos. Sin embargo, DPR requiere datos de entrenamiento anotados con pares de consultas y pasajes relevantes, lo que limita su aplicabilidad a dominios donde dichos datos no están disponibles.

Para abordar esta limitación, Izacard et al.[52] propusieron Contriever, un modelo de recuperación densa entrenado de manera no supervisada mediante aprendizaje contrastivo sobre pares de fragmentos de texto generados automáticamente. Contriever demostró que es posible obtener representaciones vectoriales de alta calidad para recuperación sin necesidad de datos anotados, lo que lo hace especialmente valioso para dominios especializados donde los datos de entrenamiento son escasos.

La similitud entre vectores en la recuperación densa se calcula típicamente mediante el coseno del ángulo entre los vectores de consulta y documento, o mediante el producto escalar entre ellos. Ambas métricas son equivalentes cuando los vectores están normalizados a norma unitaria.

Una ventaja importante de la recuperación densa sobre la recuperación léxica es su capacidad para capturar similitud semántica: dos fragmentos pueden ser recuperados como relevantes, aunque no compartan ningún término en común, siempre que sus representaciones vectoriales sean similares. Esto permite al recuperador denso encontrar documentos relevantes formulados con vocabulario diferente al de la consulta, superando el problema del vocabulario que afecta a los métodos léxicos.

Sin embargo, la recuperación densa también tiene limitaciones. En primer lugar, los modelos de *embedding* son computacionalmente costosos de entrenar y de aplicar, especialmente cuando el corpus es grande. En segundo lugar, los modelos de *embedding* pueden tener dificultades para capturar la relevancia de términos muy específicos o de larga cola, precisamente donde los métodos léxicos son más fuertes [57]. En tercer lugar, los modelos de *embedding* son opacos: no es fácil entender por qué un fragmento recibió una puntuación alta o baja, lo que dificulta la depuración del sistema.

4.9 Recuperación híbrida

La recuperación híbrida combina las señales de los métodos de recuperación dispersa y densa con el objetivo de aprovechar sus fortalezas complementarias y mitigar sus debilidades individuales. La motivación fundamental es que los métodos léxicos y los métodos semánticos capturan tipos de relevancia diferentes y complementarios: los primeros son más precisos para términos específicos y de larga cola, mientras que los segundos son más efectivos para capturar similitud semántica y manejar la variabilidad léxica [58].

La combinación de señales en la recuperación híbrida puede realizarse mediante diferentes estrategias. La más directa es la fusión de puntuaciones (*score fusion*), que consiste en combinar linealmente las puntuaciones asignadas por el recuperador léxico y el recuperador denso a cada fragmento, ponderando cada señal según su importancia relativa. Esta estrategia requiere que las puntuaciones de ambos recuperadores estén en la misma escala, lo que generalmente implica un paso de normalización previo.

En la formulación ponderada adoptada en esta investigación, α se restringe al intervalo $0 \leq \alpha \leq 1$. Este parámetro representa el peso asignado a la señal léxica: $\alpha = 1$ corresponde a una recuperación puramente léxica y $\alpha = 0$ a una recuperación puramente semántica. Los valores intermedios combinan ambas señales y asignan un peso de $1 - \alpha$ a la componente semántica.

Una estrategia alternativa es la fusión por rangos (*rank fusion*), que combina las listas de resultados ordenadas producidas por cada recuperador, sin necesidad de que las puntuaciones estén en la misma escala. El método más utilizado en esta categoría es *Reciprocal Rank Fusion* (RRF), propuesto por Cormack et al. [59], que asigna a cada documento una puntuación basada en su posición en cada lista de resultados mediante la fórmula:

$$\text{RRF}(d) = \sum_{r \in R} \frac{1}{k + \text{rank}_r(d)} \quad (3)$$

donde R es el conjunto de recuperadores, $\text{rank}_r(d)$ es la posición del documento d en la lista del recuperador r , y k es una constante de suavizado (típicamente $k = 60$). RRF tiene la ventaja de ser robusto frente a diferencias de escala entre las puntuaciones de distintos recuperadores y de no requerir ajuste de pesos, lo que lo hace fácil de implementar y aplicar.

Otra estrategia de fusión es la denominada *Relative Score Fusion* (RSF), que normaliza las puntuaciones de cada recuperador al rango $[0,1]$ antes de combinarlas, permitiendo una comparación directa entre las señales de distintos recuperadores. A diferencia de RRF, RSF preserva la información de magnitud de las puntuaciones, lo que puede ser ventajoso cuando las diferencias de puntuación entre documentos son informativas.

La evidencia empírica sobre la efectividad de la recuperación híbrida es consistente en la literatura. Se demostró que la combinación de BM25 con un recuperador denso supera a

ambos métodos individuales en múltiples *benchmarks* de recuperación[60]. Luan et al. [58] mostraron resultados similares en el contexto de la recuperación de pasajes para sistemas de preguntas y respuestas. Más recientemente, el *benchmark* BEIR, que evalúa la capacidad de generalización de los métodos de recuperación en múltiples dominios y tipos de consulta, ha confirmado que los métodos híbridos son más robustos que los métodos individuales en escenarios de dominio cruzado [57].

4.10 Flujo completo de un sistema RAG

La integración de todos los componentes descritos en las secciones anteriores da lugar al flujo operativo completo de un sistema RAG. Comprender este flujo de manera integral es indispensable para apreciar cómo cada componente contribuye al desempeño global del sistema y cómo las decisiones de diseño en cada etapa se propagan a lo largo del pipeline.

El flujo comienza con la formulación de la consulta por parte del usuario. En sistemas de diálogo, la consulta puede ser una pregunta directa, una solicitud de información o una continuación de una conversación previa. En algunos sistemas, esta consulta es preprocesada antes de ser enviada al recuperador: se puede aplicar normalización léxica, lematización, corrección ortográfica o expansión de consultas para mejorar la compatibilidad entre la consulta y los documentos indexados [6][61].

La consulta preprocesada se envía al componente de recuperación, que puede ser un recuperador léxico, un recuperador denso o un sistema híbrido, como se ha descrito en las secciones anteriores. El recuperador consulta el índice y devuelve una lista de fragmentos ordenados por su puntuación de relevancia estimada. En sistemas con recuperación híbrida, las puntuaciones de los distintos recuperadores se combinan mediante una función de fusión antes de producir la lista final.

Los fragmentos recuperados pasan opcionalmente por un componente de reordenamiento, que refina el orden de los candidatos utilizando un modelo más expresivo que el recuperador inicial. Este componente puede descartar fragmentos que, aunque recibieron una puntuación alta del recuperador, no son suficientemente relevantes para la consulta específica, o puede incorporar criterios adicionales de relevancia como la diversidad o la cobertura temática.

A continuación, el sistema construye el contexto de entrada para el modelo generativo, integrando los fragmentos reordenados, la consulta del usuario, las instrucciones de sistema y cualquier otro elemento contextual relevante. Este contexto se ensambla en un *prompt* estructurado que respeta las convenciones del modelo generativo utilizado.

El modelo generativo recibe el *prompt* y produce una respuesta en lenguaje natural. En esta etapa, el modelo combina su conocimiento paramétrico con la información explícita proporcionada en el contexto. La calidad de la respuesta generada depende tanto de la capacidad del modelo como de la calidad del contexto: un contexto rico en información

relevante y libre de ruido permite al modelo generar respuestas más precisas, completas y verificables.

4.11 Métricas para evaluación RAG

BLEU

BLEU (*Bilingual Evaluation Understudy*) es una métrica automática de solapamiento de n-gramas ampliamente utilizada en evaluaciones de sistemas RAG y QA [62]. Mide la precisión del solapamiento de n-gramas entre la respuesta candidata generada y una o más respuestas de referencia, favoreciendo coincidencias exactas de secuencias de palabras.

Ventajas: BLEU es computacionalmente eficiente, reproducible y permite comparaciones rápidas entre múltiples sistemas sin requerir anotación humana adicional, lo que facilita la evaluación iterativa durante el desarrollo de sistemas RAG [62].

Desventajas: BLEU es sensible a la redacción exacta y penaliza paráfrasis válidas que expresan el mismo contenido factual con diferentes palabras, lo que limita su capacidad para evaluar la corrección semántica y factual de respuestas en dominios especializados donde múltiples formulaciones pueden ser igualmente correctas [62].

ROUGE

ROUGE (*Recall-Oriented Understudy for Gisting Evaluation*) comprende una familia de métricas (ROUGE-1, ROUGE-2, ROUGE-L) orientadas a medir el *recall* de solapamiento entre texto generado y referencias. ROUGE-1 y ROUGE-2 capturan solapamiento de unigramas y bigramas respectivamente, mientras que ROUGE-L mide la subsecuencia común más larga, favoreciendo la similitud superficial de contenido.

Ventajas: Las variantes de ROUGE son especialmente útiles para evaluar la cobertura de contenido en resúmenes y respuestas generadas, y su orientación al *recall* complementa métricas orientadas a precisión como BLEU, permitiendo una evaluación más balanceada de la completitud de las respuestas.

Desventajas: Al igual que BLEU, ROUGE favorece la similitud superficial de palabras y no puede garantizar la corrección factual ni la fidelidad al contexto recuperado en sistemas RAG, lo que puede resultar en puntuaciones altas para respuestas que parafrasean correctamente, pero introducen información no fundamentada.

METEOR

METEOR (*Metric for Evaluation of Translation with Explicit ORdering*) es una métrica reportada en múltiples pipelines de evaluación RAG que considera *stemming* y sinónimos, ofreciendo mayor tolerancia a paráfrasis que BLEU. METEOR intenta capturar similitud

semántica más allá del solapamiento exacto de n-gramas mediante el uso de recursos lingüísticos.

Ventajas: METEOR es más robusto ante variaciones de redacción y paráfrasis que BLEU y ROUGE, lo que lo hace más apropiado para evaluar respuestas en dominios donde la expresión del conocimiento puede variar considerablemente manteniendo la corrección factual.

Desventajas: A pesar de su mayor flexibilidad semántica, METEOR no puede garantizar la fidelidad factual al contexto recuperado en RAG, y su dependencia de recursos lingüísticos (listas de sinónimos, *stemmers*) puede limitar su efectividad en dominios especializados con terminología técnica o cultural específica no cubierta por diccionarios generales.

BERTScore

BERTScore es una métrica basada en *embeddings* contextuales que utiliza modelos BERT para capturar similitud semántica entre textos generados y referencias. En lugar de comparar tokens superficiales, *BERTScore* compara representaciones vectoriales contextuales, permitiendo reconocer equivalencias semánticas más allá del solapamiento léxico.

Ventajas: *BERTScore* es significativamente más robusto ante paráfrasis que las métricas basadas en n-gramas, capturando similitud semántica profunda mediante representaciones contextuales, lo que lo hace más apropiado para evaluar la corrección conceptual de respuestas en dominios especializados.

Desventajas: Una alta similitud semántica según *BERTScore* no garantiza que los hechos generados estén fundamentados en los documentos recuperados, ya que el modelo puede generar contenido semánticamente coherente pero factualmente incorrecto o no respaldado por el contexto RAG.

Faithfulness (Fidelidad)

Faithfulness o fidelidad es un eje de evaluación explícito en marcos como ARES (*Automated RAG Evaluation System*) y trabajos relacionados, que mide si las afirmaciones generadas por el modelo están respaldadas por el contexto recuperado en lugar de ser inventadas o post-racionalizadas[63]. La literatura muestra que la fidelidad es distinta de la mera corrección de citas y que la falta de fidelidad (post-racionalización) es común en sistemas RAG.

Ventajas: *Faithfulness* es crucial en RAG porque evalúa directamente la propiedad fundamental del sistema: que las respuestas estén fundamentadas en evidencia recuperada. Esta métrica permite detectar alucinaciones y contenido inventado incluso cuando la respuesta es superficialmente coherente y relevante.

Desventajas: La evaluación de fidelidad típicamente requiere anotación humana o modelos juez entrenados, lo que incrementa el costo y la complejidad de la evaluación. Además, la

definición operacional de «respaldo suficiente» puede variar según el dominio y el nivel de inferencia aceptable, complicando la estandarización de la métrica.

Answer Relevance (Relevancia de respuesta)

Answer Relevance es uno de los ejes primarios en ARES y otras evaluaciones RAG, midiendo si la respuesta generada aborda apropiadamente la consulta del usuario. Se operacionaliza mediante jueces humanos o modelos aprendidos en los sistemas de evaluación citados.

Ventajas: La relevancia de respuesta evalúa la utilidad práctica del sistema desde la perspectiva del usuario, asegurando que las respuestas no solo sean correctas y fundamentadas, sino también pertinentes a la pregunta formulada, lo que es esencial para la experiencia de usuario en aplicaciones museográficas.

Desventajas: Una respuesta puede ser relevante pero infiel (relevante pero alucinada), por lo que la relevancia debe combinarse con métricas de fidelidad y contexto para garantizar calidad integral. Además, la evaluación de relevancia puede ser subjetiva y dependiente del contexto cultural y las expectativas del usuario.

Context Precision y Context Recall

Context Precision y *Context Recall* son un par de métricas utilizadas en flujos de trabajo estilo RAGAS e implementaciones como ragR para medir qué proporción del contexto recuperado es relevante (precisión) y qué proporción del contexto relevante de verdad fundamental fue recuperado (*recall*)[64]. ragR reporta puntuación basada en LLM para métricas RAGAS incluyendo precisión y *recall* de contexto.

Ventajas: Alta precisión de contexto ayuda a reducir el riesgo de alucinación al minimizar información irrelevante que el modelo podría malinterpretar, mientras que alto *recall* de contexto asegura que la evidencia necesaria esté disponible para el generador, permitiendo respuestas completas y bien fundamentadas.

Desventajas: La evaluación de estas métricas requiere conocimiento de la verdad fundamental (*ground truth*) sobre qué contexto es relevante, lo que puede no estar disponible en dominios especializados sin anotación experta. Además, la implementación basada en LLM como juez introduce dependencia de la calidad del modelo evaluador.

Hallucination Rate (Tasa de alucinación)

Hallucination Rate mide la fracción de contenido generado que no está respaldado por la evidencia recuperada, y es utilizado como objetivo de evaluación en múltiples trabajos y como motivación central para detectores especializados como Luna[65]. *Ragtruth* proporciona un corpus dedicado de alucinaciones con anotaciones de categorías finas de alucinación.

Ventajas: La tasa de alucinación evalúa directamente uno de los problemas más críticos en sistemas RAG para dominios culturales e históricos, donde la precisión factual es esencial. La detección puede realizarse con clasificadores especializados o anotación humana, permitiendo identificar y mitigar contenido no fundamentado.

Desventajas: La detección automática de alucinaciones requiere modelos especializados o reglas complejas, y la definición de qué constituye una alucinación puede ser ambigua en casos de inferencia razonable versus invención. La anotación humana de alucinaciones es costosa y requiere experiencia en el dominio para distinguir errores factuales de interpretaciones válidas.

Recall@k

Recall@k es una de las métricas más utilizadas para evaluar sistemas de recuperación de información y motores de búsqueda dentro de arquitecturas RAG. Su objetivo es medir la proporción de documentos relevantes que fueron recuperados entre los primeros K resultados devueltos por el sistema. Formalmente, *Recall@k* representa la fracción de documentos relevantes presentes en el *Top-K* respecto al total de documentos relevantes existentes para una consulta determinada.

Ventajas: *Recall@k* permite evaluar la capacidad del sistema para recuperar evidencia útil antes de la etapa de generación. Un valor elevado indica que el recuperador logra localizar la mayor parte de la información necesaria para responder adecuadamente una consulta, reduciendo la probabilidad de que el modelo generativo produzca respuestas incompletas debido a la ausencia de contexto relevante. Esta métrica resulta especialmente importante en sistemas RAG, donde la calidad de la generación depende directamente de la cobertura del proceso de recuperación.

Desventajas: *Recall@k* no considera el orden en que aparecen los documentos recuperados. Un sistema puede obtener un valor alto de Recall recuperando todos los documentos relevantes, aunque estos aparezcan en posiciones muy bajas dentro del *Top-K*. Asimismo, tampoco penaliza la recuperación de documentos irrelevantes, por lo que suele complementarse con métricas que evalúan la calidad del ordenamiento, como MRR o nDCG.

Mean Reciprocal Rank (MRR@K)

Mean Reciprocal Rank (MRR) es una métrica ampliamente utilizada para evaluar sistemas de búsqueda, recuperación de información y sistemas de preguntas y respuestas. Su propósito consiste en medir la posición del primer documento relevante recuperado por el sistema mediante el recíproco de su posición dentro del ranking. Posteriormente, este valor se promedia sobre todas las consultas evaluadas.

Ventajas: MRR recompensa a los sistemas que presentan el primer documento relevante en las primeras posiciones del ranking, característica particularmente importante en sistemas RAG, ya que los modelos generativos suelen utilizar principalmente los primeros documentos recuperados para construir sus respuestas. Una puntuación elevada de MRR

indica que el usuario encuentra rápidamente la información relevante, mejorando tanto la eficiencia del sistema como la calidad potencial de la generación.

Desventajas: MRR únicamente considera la posición del primer documento relevante e ignora el resto de los documentos recuperados. Como consecuencia, dos sistemas pueden obtener el mismo valor de MRR aun cuando uno de ellos recupere mayor cantidad de evidencia útil después del primer documento. Debido a esta limitación, MRR suele emplearse junto con *Recall@k* o nDCG para obtener una evaluación más completa del proceso de recuperación.

Normalized Discounted Cumulative Gain (nDCG@k)

Normalized Discounted Cumulative Gain (nDCG) es una métrica diseñada para evaluar la calidad del ordenamiento producido por un sistema de recuperación considerando tanto la relevancia de los documentos como la posición que ocupan dentro del ranking. A diferencia de métricas binarias, nDCG permite asignar diferentes niveles de relevancia a cada documento recuperado, otorgando mayor importancia a aquellos ubicados en las primeras posiciones.

Ventajas: nDCG proporciona una evaluación más completa del ranking al considerar simultáneamente la relevancia y el orden de los resultados recuperados. Esta característica resulta especialmente útil cuando existen múltiples documentos parcialmente relevantes o con distintos grados de utilidad para responder una consulta. En sistemas RAG, un valor elevado de nDCG indica que el contexto más relevante se encuentra disponible en las primeras posiciones, favoreciendo una generación más precisa y fundamentada.

Desventajas: El cálculo de nDCG requiere disponer de niveles de relevancia previamente definidos para cada documento, lo que generalmente implica anotación manual o conocimiento experto. Además, la interpretación de la métrica puede resultar menos intuitiva que otras medidas más simples como *Recall* o *Hit Rate*, especialmente cuando se trabaja con diferentes escalas de relevancia.

Hit Rate (Hit@k)

Hit Rate, también conocida como *Hit@k*, es una métrica binaria utilizada para determinar si al menos un documento relevante aparece entre los primeros K resultados recuperados por el sistema. Para cada consulta, la métrica asigna un valor de uno cuando existe al menos un elemento relevante dentro del *Top-K*, y cero en caso contrario. El resultado final corresponde al promedio obtenido sobre todas las consultas evaluadas.

Ventajas: *Hit@k* es una métrica sencilla de interpretar y proporciona una estimación directa de la probabilidad de éxito del sistema durante la recuperación inicial. En aplicaciones RAG resulta particularmente útil para verificar si el generador dispone de al menos una evidencia relevante que le permita construir una respuesta fundamentada. Asimismo, su simplicidad facilita la comparación entre diferentes estrategias de recuperación.

Desventajas: *Hit@k* no distingue entre recuperar un único documento relevante o recuperar múltiples evidencias útiles, ni considera la posición específica que ocupa dicho documento dentro del ranking. Por ello, dos sistemas con comportamientos muy diferentes pueden obtener exactamente el mismo valor de *Hit Rate*. Debido a esta limitación, normalmente se utiliza como una métrica complementaria junto con *Recall*, MRR y nDCG para obtener una evaluación más completa del desempeño del recuperador.

Capítulo V. Método de solución

En este capítulo se describe el método de solución seguido para desarrollar la propuesta de investigación. El objetivo del método es establecer una secuencia organizada de actividades que permita seleccionar los componentes más adecuados para un sistema RAG, integrarlos dentro de una arquitectura funcional y validar experimentalmente su desempeño.

El método se estructuró en tres etapas consecutivas, mostradas en la Figura 5.1. La primera etapa se enfocó en la investigación y selección de estrategias de recuperación, modelos generativos y métricas de evaluación. La segunda consistió en el diseño e integración del sistema RAG utilizando los componentes seleccionados. Finalmente, la tercera etapa estuvo dedicada a la validación experimental del sistema mediante diferentes experimentos orientados a evaluar tanto el módulo de recuperación como la calidad de las respuestas generadas.

Cada una de estas etapas contribuye al cumplimiento de los objetivos específicos de la investigación y proporciona la evidencia necesaria para evaluar la viabilidad de la propuesta.

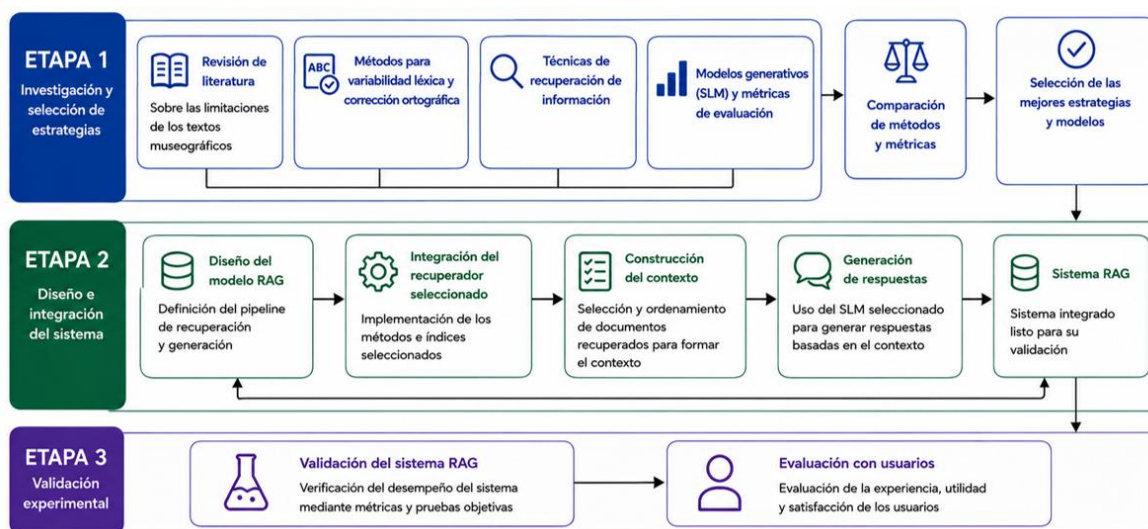


Figura 5.1. Etapas del método de solución

5.1 Etapa 1. Investigación y selección de estrategias

La primera etapa tuvo como propósito identificar las técnicas, modelos y métricas más apropiadas para construir el sistema RAG. Para ello se realizó una revisión de literatura sobre los textos museográficos y sobre cada uno de los componentes del RAG, ya mencionados en el capítulo III y IV.

Durante esta etapa se analizaron cuatro aspectos principales.

5.1.1 Revisión contextual de los problemas de comunicación en los textos museográficos

A diferencia del estado del arte presentado en el Capítulo III, orientado al análisis de sistemas RAG, técnicas de procesamiento de lenguaje natural y estrategias de mitigación de alucinaciones, esta actividad tuvo un propósito contextual. Su objetivo fue identificar los principales problemas de comunicación, comprensión, legibilidad y accesibilidad presentes en los textos utilizados en entornos museográficos mexicanos. Los resultados de esta actividad se presentan en el Capítulo VI y se utilizaron para caracterizar las necesidades del dominio de aplicación y fundamentar el desarrollo del sistema propuesto.

La búsqueda bibliográfica se llevó a cabo inicialmente en Google Scholar mediante una estrategia estructurada basada en las siguientes cadenas de búsqueda:

- (legibilidad OR readability) AND (museos OR museums) AND México AND (comprensión OR comprehension)
- (infografías OR infographics) AND (museos OR museums) AND México AND (comunicación OR communication)
- (mejorar OR improve) AND (comunicación OR communication) AND (museos OR museums) AND México AND (textos OR texts)

Posteriormente, con el fin de ampliar la cobertura de la revisión, se realizó una búsqueda dirigida en revistas especializadas relacionadas con patrimonio cultural, museología y comunicación, entre ellas la *Revista Latino-Americana de Turismología*, la *Revista del Museo de Antropología* y *Dilemas Contemporáneos: Educación, Política y Valores*.

La búsqueda inicial recuperó aproximadamente 843 documentos. A continuación, se aplicó un proceso de selección basado en la revisión del título y el resumen de cada publicación, utilizando criterios de inclusión y exclusión previamente definidos. Este procedimiento permitió reducir el conjunto inicial a un grupo de trabajos potencialmente relevantes. Finalmente, se seleccionaron 60 documentos para su lectura y análisis completo, los cuales constituyeron la base para el estudio del problema de investigación y la identificación de las estrategias utilizadas en trabajos previos.

Los criterios de inclusión considerados fueron los siguientes:

- Publicaciones en español o inglés.
- Estudios relacionados con museos, patrimonio cultural o divulgación científica.
- Trabajos enfocados en legibilidad, comprensión o comunicación de información.
- Investigaciones que mencionaran explícitamente el contexto mexicano en el título o resumen.

Por otra parte, se excluyeron aquellos trabajos que:

- No presentaban una relación directa con la comprensión o comunicación textual.

- Corresponderían a contextos geográficos que no aportaban evidencia relevante para el objetivo de la investigación.

5.1.2 Métodos para variabilidad léxica y corrección ortográfica

Durante la etapa de investigación se identificó que uno de los problemas más frecuentes en los sistemas de recuperación de información basados en lenguaje natural corresponde a la variabilidad léxica presente en las consultas de los usuarios. Esta variabilidad incluye errores ortográficos, diferencias en la escritura de palabras, uso de sinónimos, abreviaturas y distintas formas de expresar una misma necesidad de información. En sistemas RAG, estas variaciones pueden afectar el proceso de recuperación, ya que la calidad de las respuestas generadas depende directamente de la pertinencia de los documentos recuperados.

Con el propósito de identificar estrategias que permitieran reducir el impacto de este problema, se realizó una revisión de los principales enfoques utilizados en la literatura. Entre ellos se encuentran los métodos basados en distancia de edición, como *Levenshtein* y *Damerau-Levenshtein*, los cuales permiten cuantificar la similitud entre cadenas de texto mediante operaciones de inserción, eliminación, sustitución y transposición de caracteres. Asimismo, se analizaron técnicas de corrección ortográfica (*spell checking*), orientadas a normalizar las consultas antes del proceso de recuperación.

Además de los enfoques léxicos, se revisaron métodos basados en representaciones semánticas mediante modelos de *embeddings*. A diferencia de las técnicas tradicionales, estos modelos representan palabras y oraciones en un espacio vectorial, permitiendo recuperar información a partir de la similitud de significado y no únicamente de la coincidencia exacta entre términos. También se identificaron propuestas híbridas que combinan evidencia léxica y semántica con el objetivo de aprovechar las fortalezas de ambos enfoques.

El análisis realizado mostró que cada estrategia presenta ventajas y limitaciones. Los métodos basados en distancia de edición resultan adecuados para corregir errores tipográficos y consultas con pequeñas variaciones ortográficas; sin embargo, presentan dificultades cuando las diferencias entre consulta y documento corresponden a variaciones semánticas. Por otra parte, los métodos basados en *embeddings* permiten recuperar documentos conceptualmente relacionados aun cuando no exista coincidencia léxica, aunque su desempeño depende de la calidad del modelo de representación utilizado. Finalmente, las estrategias híbridas buscan equilibrar ambos tipos de evidencia mediante la combinación de información léxica y semántica.

Debido a que la literatura revisada no establece un consenso respecto al comportamiento de estas estrategias en consultas relacionadas con información histórica y cultural, se decidió incorporarlas dentro de la fase experimental de la presente investigación. Esta decisión permitió evaluar, bajo un mismo conjunto de condiciones experimentales, la robustez de los diferentes enfoques frente a variaciones léxicas y errores ortográficos, con el propósito de seleccionar la estrategia más adecuada para integrarse al sistema RAG propuesto.

5.1.3 Técnicas de recuperación de información

Debido a que el componente de recuperación determina la evidencia que posteriormente utilizará el modelo generativo, su selección constituye una de las decisiones más importantes dentro del diseño del sistema RAG.

Como parte de este análisis se consideraron tres familias de métodos ampliamente utilizadas en la literatura: recuperación léxica, recuperación semántica y recuperación híbrida. La descripción teórica y el funcionamiento de cada una de estas estrategias se presentan en el Capítulo IV; por ello, en esta sección únicamente se describe el proceso seguido para su selección dentro de la propuesta.

Para cada familia de métodos se identificaron implementaciones representativas, modelos de *embeddings* y mecanismos de combinación entre evidencia léxica y semántica. Asimismo, se revisaron diferentes criterios utilizados en trabajos previos para evaluar el desempeño de los recuperadores, considerando tanto la capacidad para localizar documentos relevantes como la robustez frente a diferentes tipos de consultas.

Tomando en cuenta que el desempeño de estas estrategias depende del dominio de aplicación, del corpus documental y del modelo de representación utilizado, no fue posible seleccionar un método únicamente con base en la literatura. Por esta razón, se decidió realizar una evaluación experimental comparando las distintas estrategias de recuperación bajo un mismo conjunto de condiciones, con el fin de identificar la alternativa más adecuada para integrarse al sistema RAG propuesto.

5.1.4 Modelos generativos (SLM) y métricas de evaluación

Se analizaron diferentes SLMs y métricas de evaluación con el propósito de seleccionar los componentes más adecuados para el sistema RAG propuesto. Debido a las restricciones de recursos computacionales y a la necesidad de contar con un modelo que pudiera ejecutarse localmente, se consideraron arquitecturas representativas de SLM reportadas en la literatura.

En la literatura todos los modelos SLM presentan problemas en el seguimiento de instrucciones en el *prompt*, una mala generación de texto, ya que presentan generando tokens en otro idioma o sin sentido alguno colocando símbolos; por ello, se decidió realizar una evaluación experimental de los modelos y métricas seleccionados. Los resultados de esta evaluación sirvieron como base para la selección del modelo generativo y del esquema de evaluación utilizados en la propuesta.

5.2 Etapa 2. Diseño e integración del sistema

Una vez seleccionadas las estrategias de recuperación, el modelo generativo y las métricas de evaluación durante la etapa anterior, se procedió al diseño e integración del sistema RAG propuesto. El propósito de esta etapa fue construir una arquitectura que permitiera recuperar

información relevante a partir de consultas en lenguaje natural, generar un contexto con la evidencia recuperada y producir respuestas fundamentadas mediante un modelo de lenguaje pequeño.

El desarrollo se organizó en cinco actividades principales: el diseño de la arquitectura del sistema, la integración del recuperador seleccionado, la construcción del contexto de entrada, la generación de respuestas y la integración del sistema completo. Estas actividades permitieron incorporar de manera secuencial los componentes seleccionados durante la etapa de investigación, obteniendo un sistema funcional preparado para su validación experimental.

5.2.1 Diseño del sistema RAG

Como primera actividad se definió la arquitectura general del sistema siguiendo el paradigma *Retrieval-Augmented Generation* (RAG). El diseño se estructuró en módulos independientes con el propósito de facilitar la integración, sustitución y evaluación de cada uno de sus componentes como se muestra en la Figura 5.2.

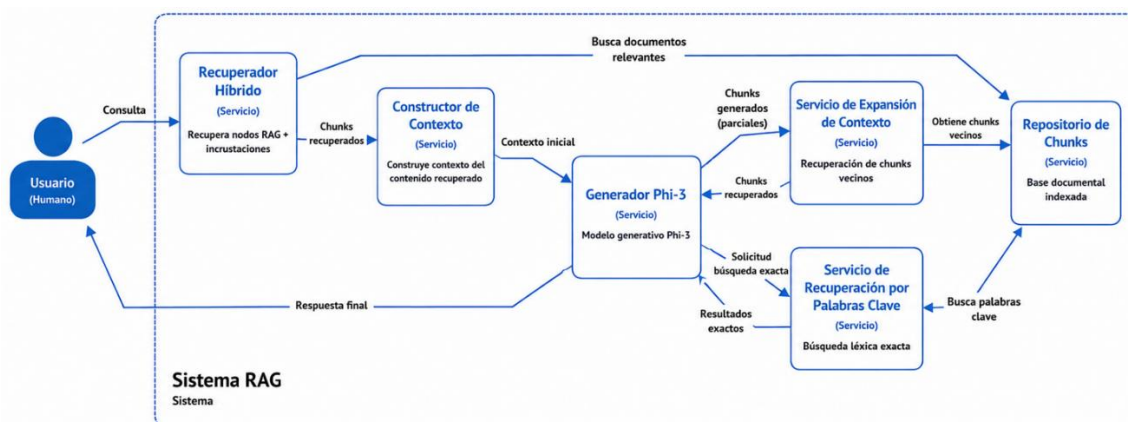


Figura 5.2. Arquitectura RAG propuesta

La arquitectura considera un flujo de procesamiento que inicia con la recepción de una consulta en lenguaje natural. Posteriormente, el módulo de recuperación identifica los documentos más relevantes del corpus, los cuales son utilizados para construir el contexto que servirá como entrada del modelo generativo. Finalmente, el modelo produce una respuesta fundamentada en la evidencia recuperada.

La adopción de una arquitectura modular permitió desacoplar el proceso de recuperación del proceso de generación, facilitando la evaluación independiente de ambos componentes y permitiendo sustituir modelos sin modificar el resto del sistema.

5.2.2 Integración del recuperador seleccionado

Una vez definida la arquitectura, se incorporó el módulo de recuperación de información utilizando la estrategia seleccionada durante la etapa de investigación. Para ello se implementaron los índices y modelos necesarios para procesar las consultas y obtener un conjunto ordenado de documentos relevantes.

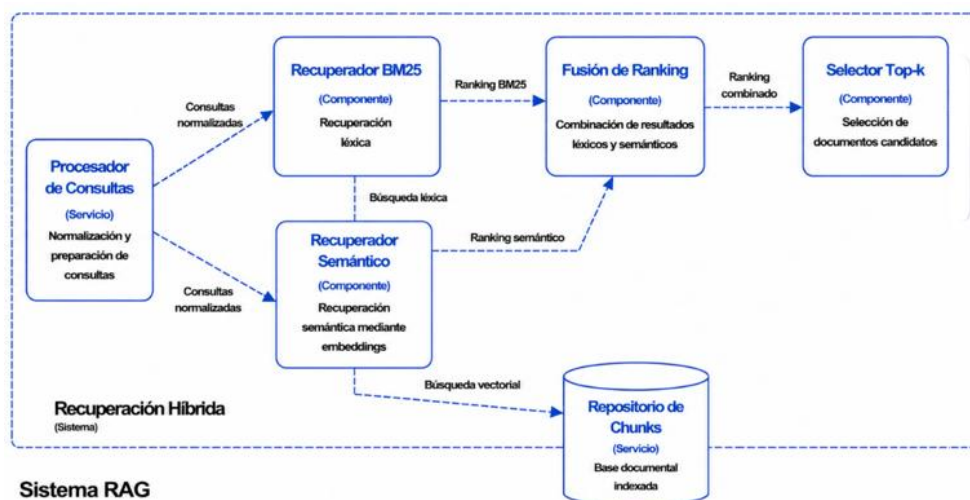


Figura 5.3. Módulo de recuperación propuesto

La integración contempló tanto la representación léxica como la representación semántica de los documentos representada en la Figura 5.3, permitiendo utilizar diferentes estrategias de recuperación dentro de una misma arquitectura. De esta forma, el sistema quedó preparado para incorporar los distintos métodos considerados en la fase experimental sin requerir modificaciones estructurales en el pipeline.

El diseño del recuperador también permitió controlar parámetros como el número de documentos recuperados y el mecanismo de combinación entre evidencia léxica y semántica, facilitando posteriormente la comparación experimental de diferentes configuraciones.

5.2.3 Construcción del contexto

Después del proceso de recuperación, los documentos obtenidos fueron utilizados para construir el contexto que sería proporcionado al modelo generativo. Esta actividad consistió en seleccionar los documentos recuperados, ordenarlos de acuerdo con su relevancia y concatenar su contenido en una única entrada para el modelo de lenguaje.

Con el propósito de mantener la información más pertinente dentro de la ventana de contexto del modelo, se empleó una estrategia de fragmentación de documentos (*chunking*) y recuperación *Top-k*, permitiendo controlar la cantidad de información utilizada durante la generación de respuestas.

La construcción del contexto constituye una etapa fundamental dentro de la arquitectura RAG, ya que determina la evidencia disponible para el modelo generativo durante el proceso de inferencia.

5.2.4 Generación de respuestas

Una vez construido el contexto, éste fue enviado al modelo generativo seleccionado durante la etapa de investigación. El modelo recibió como entrada la consulta del usuario junto con la información recuperada, generando una respuesta fundamentada en el contenido proporcionado por el recuperador.

Para mantener la consistencia del sistema, la generación se realizó mediante una plantilla de instrucciones (*prompt*) diseñada para orientar al modelo a responder utilizando exclusivamente la información disponible en el contexto recuperado. Esta estrategia buscó reducir la generación de información no respaldada por el corpus documental y favorecer respuestas coherentes con el contenido recuperado.

La separación entre recuperación y generación permitió modificar cualquiera de los dos componentes sin afectar el funcionamiento general del sistema.

5.2.5 Sistema RAG integrado

Como resultado de esta etapa se obtuvo un sistema RAG completamente integrado, compuesto por un módulo de recuperación de información, un mecanismo de construcción de contexto y un modelo generativo encargado de elaborar las respuestas finales.

El funcionamiento del sistema inicia con la recepción de una consulta, continúa con la recuperación y selección de documentos relevantes, posteriormente construye el contexto que será utilizado por el modelo generativo y finaliza con la generación de una respuesta basada en la evidencia recuperada. Esta integración permitió disponer de una plataforma funcional sobre la cual se realizaron los experimentos descritos en la etapa de validación.

5.2.6 Configuración del sistema

La configuración final integró los componentes seleccionados en las etapas anteriores y fijó los parámetros utilizados en los experimentos posteriores. La Tabla 5.1 resume el modelo generativo, el modelo de *embeddings*, el recuperador, la profundidad de recuperación y los parámetros de segmentación que se mantuvieron constantes durante la evaluación.

<i>Parámetro</i>	<i>Valor</i>
<i>Modelo generativo</i>	Phi-3-mini-128k-instruct
<i>Modelo de embeddings</i>	intfloat/multilingual-e5-base

<i>Recuperación</i>	BM25 + <i>embeddings</i> (configuración experimental)
<i>Top-k</i>	10
<i>Chunk size</i>	220
<i>Chunk overlap</i>	40
<i>Idioma</i>	Español

Tabla 5.1. Configuración de parámetros del sistema RAG

5.3 Etapa 3. Validación experimental

La tercera etapa tuvo como propósito validar el desempeño del sistema RAG desarrollado durante la etapa de integración. Para ello se diseñó un conjunto de experimentos orientados a evaluar de manera independiente los principales componentes del sistema, así como su funcionamiento integral una vez incorporados dentro de la arquitectura propuesta.

La validación se estructuró mediante cinco experimentos complementarios. El primero analizó la robustez del módulo de recuperación ante errores ortográficos; el segundo permitió seleccionar el modelo generativo más adecuado; el tercero comparó distintas estrategias de recuperación; el cuarto evaluó el sistema RAG completo mediante métricas automáticas y múltiples jueces; y el quinto estudió su interacción con usuarios. Esta organización permitió analizar los componentes de forma independiente antes de valorar el funcionamiento global de la propuesta.

5.3.1 Diseño experimental

Con el propósito de garantizar una evaluación objetiva del sistema, se diseñó un conjunto de experimentos controlados que permitieran medir el desempeño de los diferentes componentes bajo condiciones reproducibles. Cada experimento fue desarrollado sobre conjuntos de datos específicos, utilizando métricas acordes con el objetivo particular de la evaluación.

La validación experimental contempló el uso de corpus especializados para recuperación de información, conjuntos de preguntas de opción múltiple para la selección del modelo generativo y colecciones de preguntas y respuestas para la evaluación del sistema RAG. En todos los casos se mantuvieron constantes las condiciones experimentales correspondientes a cada evaluación, con el fin de asegurar la comparabilidad entre los métodos analizados.

La incertidumbre se reportó de acuerdo con la unidad de repetición disponible. Cuando existieron varias ejecuciones o distribuciones por consulta, se incluyeron medidas de dispersión o intervalos de confianza. Las tablas derivadas de una sola ejecución se presentan como estimaciones puntuales y se identifican como tales, ya que una desviación

estándar entre repeticiones no puede calcularse sin réplicas independientes. Este criterio evita atribuir variabilidad experimental a resultados obtenidos en una única corrida.

Los resultados obtenidos durante esta etapa permitieron seleccionar las configuraciones finales empleadas en el sistema y validar el cumplimiento de los objetivos planteados en la investigación.

5.3.2 Experimento 1. Robustez ante errores ortográficos

El objetivo de este experimento es evaluar la robustez de distintos métodos de recuperación de información cuando las consultas contienen errores ortográficos. Para ello, se comparan tres enfoques de recuperación: un método léxico basado en la distancia de Damerau-Levenshtein (DL), un método semántico basado en *embeddings* y un método híbrido que combina ambos enfoques. El propósito es determinar si los modelos de *embeddings* son suficientemente robustos para recuperar información relevante aun cuando las consultas presentan ruido ortográfico.

La puntuación del método híbrido se calcula mediante la siguiente combinación lineal:

$$d_{hyb}(q, d) = \alpha d_{DL} + (1 - \alpha) d_{emb}(q, d) \quad (4)$$

donde α es un parámetro de ponderación que controla la contribución de cada método, q representa la consulta y d , el documento candidato. En este experimento se evaluaron los valores $\alpha \in \{0.3, 0.5, 0.7\}$.

Como conjunto de datos se utilizó el corpus DanielRjdz/BarcnasMexico, disponible en Hugging Face. El corpus contiene cuatro atributos: categoría, subcategoría, pregunta y respuesta, con un total de 670 mil registros distribuidos en las categorías de historia de México, cultura y tradiciones, gastronomía mexicana, educación, ciencia y arte, y cultura popular contemporánea.

Para la evaluación se seleccionó una muestra de 300 pares pregunta-documento. A partir de cada consulta original se generó una versión modificada introduciendo dos o cuatro errores ortográficos de manera aleatoria. Los errores considerados corresponden a las operaciones básicas de la distancia de Damerau-Levenshtein: transposición, eliminación, sustitución y adición de caracteres. La generación del ruido se ejecutó una sola vez con la semilla 42, por lo que las métricas de este experimento representan estimaciones puntuales para esa realización y no permiten calcular variabilidad entre repeticiones.

El desempeño de los métodos de recuperación se evaluó mediante las métricas MRR, Recall@k y nDCG@k, ampliamente utilizadas para medir la efectividad de los sistemas de recuperación de información.

Los modelos de embeddings evaluados se presentan en la Tabla 5.2.

bge-small-en-v1.5	e5-mistral-7b-instruct
-------------------	------------------------

bge-base-en-v1.5	e5-small
bge-large-en-v1.5	e5-base
gte-small	e5-large
gte-base	all-MiniLM-L6-v2
multilingual-e5-base	all-mpnet-base-v2
paraphrase-multilingual-MiniLM-L12-v2	bert-base-multilingual-cased
<i>Qwen3-Embedding-8B</i>	bert-base-spanish-wwm-cased
paraphrase-spanish-distilroberta	

Tabla 5.2. Modelos de embeddings usados en experimento 1

5.3.3 Experimento 2. Evaluación de modelos generativos

El objetivo de este experimento fue seleccionar el modelo generativo que posteriormente sería integrado al sistema RAG. Para ello, se compararon los 10 modelos de lenguaje pequeños (SLM) presentados en la Tabla 5.3, con el propósito de identificar aquel que ofreciera el mejor desempeño en tareas de conocimiento general.

01-ai/Yi-6B-Chat	microsoft/Phi-3-mini-128k-instruct
Qwen/Qwen2-7B-Instruct	microsoft/Phi-3-mini-4k-instruct
Qwen/Qwen2.5-3B-Instruct	microsoft/Phi-4-mini-instruct
meta-llama/Llama-3.1-8B-Instruct	mistralai/Mistral-7B-Instruct-v0.3
tiiuae/Falcon3-7B-Instruct	openchat/openchat_3.5

Tabla 5.3. Modelos de lenguaje pequeños comparados para seleccionar el generador del sistema RAG (Experimento 2)

La evaluación se realizó utilizando el conjunto de datos openai/MMMLU [4] en su configuración para español latinoamericano (ES_LA), uno de los *benchmarks* más utilizados para medir el conocimiento y las capacidades de razonamiento de los modelos de lenguaje. El conjunto de datos contiene aproximadamente 14 000 preguntas de opción múltiple distribuidas en 57 categorías, entre las que se incluyen leyes, historia, matemáticas, química, física y otras áreas del conocimiento.

Con el fin de garantizar una comparación equitativa entre los modelos, se realizó un muestreo estratificado por categoría. Se seleccionaron 10 preguntas por cada una de las 57 categorías, obteniendo un total de 570 preguntas por configuración experimental. Cada modelo fue evaluado bajo las tres configuraciones de prompting presentadas en la Tabla 5.4, las cuales difieren en las instrucciones proporcionadas al modelo y en el formato esperado de la respuesta.

CONFIGURACIÓN	FORMATO ESPERADO	CRITERIO DE EVALUACIÓN
CONFIGURACIÓN 1	Una única letra (A, B, C o D)	Se considera la primera opción válida encontrada en la respuesta.
CONFIGURACIÓN 2	Respuesta estructurada	Se requiere el formato "Respuesta: <A/B/C/D>",

CONFIGURACIÓN 3	Respuesta cerrada o abstención	permitiendo una justificación opcional. Se acepta como respuesta una opción (A, B, C o D) o la expresión "No sé".
------------------------	--------------------------------	--

Tabla 5.4. Configuraciones del experimento 2

La evaluación comparativa de los SLM se ejecutó en Google Colab con una GPU NVIDIA L4. La selección final también consideró la posibilidad de ejecutar el modelo en el equipo local disponible, equipado con una NVIDIA GeForce GTX 1650 Ti de 4096 MiB de memoria gráfica. La Tabla 5.5 distingue el entorno utilizado para la evaluación del recurso local que condicionó la elección del modelo.

Aspecto	Especificación registrada
<i>Evaluación de los SLM</i>	Google Colab con GPU NVIDIA L4 y 24 GB de memoria gráfica nominal.
<i>CPU y RAM del entorno</i>	Asignación dinámica de Google Colab; la configuración exacta no se conservó en el registro experimental.
<i>Tiempos de inferencia</i>	No se registraron de forma sistemática; no se presentan comparaciones de latencia entre modelos.
<i>Verificación de despliegue local</i>	NVIDIA GeForce GTX 1650 Ti con 4096 MiB de memoria gráfica. La prueba confirmó la ejecución funcional del modelo seleccionado; no se conservó un registro detallado de cuantización, precisión numérica o distribución de capas entre CPU y GPU.
<i>Criterio práctico de selección</i>	El criterio práctico consistió en comprobar que Phi-3-mini pudiera cargarse y ejecutar inferencia sin errores en el equipo local disponible. Esta verificación fue operativa y no constituye una medición reproducible del consumo de memoria, debido a que no se registraron la cuantización, la precisión numérica ni la posible descarga de componentes a la RAM del sistema.

Tabla 5.5. Especificaciones de los entornos utilizados

Para evaluar el desempeño de los modelos se empleó la Métrica de Contención de Respuesta Correcta (MCRC). En este trabajo, la MCRC mide la proporción de preguntas en las que la opción correcta está contenida en la salida textual del asistente, aun cuando la respuesta incluya texto adicional no solicitado.

Sea $N_{m,c}$ el número de preguntas evaluadas para el modelo m bajo la configuración c , e I_i un indicador binario que toma el valor de 1 cuando la opción correcta está contenida en la salida textual del asistente y 0 en caso contrario. La MCRC para cada configuración se calcula mediante:

$$\text{MCRC}_{m,c} = \frac{1}{N_{m,c}} \sum_{i=1}^{N_{m,c}} I_i \quad (5)$$

Dado que en cada configuración se evaluaron **570 preguntas** ($N_{m,c} = 570$), el número de respuestas correctas obtenidas por cada modelo puede expresarse como:

$$\text{Aciertos}_{m,c} = 570 \cdot \text{MCRC}_{m,c} \quad (6)$$

Finalmente, la MCRC global de cada modelo se obtuvo considerando las tres configuraciones evaluadas, para un total de **1 710 preguntas**:

$$\text{MCRC}_m = \frac{1}{1710} \sum_{c=1}^3 \sum_{i=1}^{570} I_i \quad (7)$$

5.3.4 Experimento 3. Evaluación de estrategias de recuperación

El objetivo de este experimento fue evaluar la estrategia de recuperación híbrida propuesta, basada en una combinación lineal de evidencia léxica y evidencia semántica, y compararla con métodos de recuperación ampliamente utilizados en la literatura. En particular, se busca determinar si la combinación ponderada mejora tanto la posición del documento relevante en el ranking como la capacidad del sistema para recuperarlo dentro de los primeros resultados.

Sea q una consulta y d un documento candidato. La puntuación de la estrategia híbrida se define mediante la siguiente combinación lineal:

$$s_\alpha(q, d) = \alpha s_{lex}(q, d) + (1 - \alpha) s_{sem}(q, d) \quad (8)$$

$$\alpha \in \{0.3, 0.5, 0.7\}$$

donde s_{lex} representa la similitud léxica normalizada y s_{sem} la similitud semántica normalizada.

Los valores de α fueron seleccionados para evaluar distintos niveles de contribución entre ambos componentes sin considerar los casos extremos. Un valor de $\alpha = 0.3$ asigna mayor peso a la evidencia semántica, mientras que $\alpha = 0.7$ favorece la evidencia léxica. Por su parte, $\alpha = 0.5$ representa una combinación equilibrada entre ambos enfoques. Los valores extremos ($\alpha = 0$ y $\alpha = 1$) corresponden a recuperación puramente semántica y puramente léxica, respectivamente, por lo que ya se encuentran representados mediante los métodos base utilizados como referencia (*Embedding* y BM25). De esta manera, los valores seleccionados permiten analizar si una combinación intermedia de ambas evidencias ofrece ventajas sobre cada enfoque por separado.

La evaluación se realizó utilizando el conjunto de datos MessIRve [5], un *benchmark* para tareas de recuperación de información en español. El conjunto experimental estuvo conformado por:

- 5000 pares consulta-documento relevante, considerando un único documento relevante por consulta.
- 97 configuraciones experimentales.

El número total de configuraciones evaluadas se obtuvo mediante la siguiente expresión:

$$N_{conf} = (16 \text{ modelos} \times 6 \text{ métodos}) + 1 \text{ (BM25)} = 97 \quad (9)$$

Para cada uno de los 16 modelos de *embeddings* se evaluaron seis estrategias de recuperación:

- *Embedding*.
- *Lineal*_{0.3}
- *Lineal*_{0.5}
- *Lineal*_{0.7}
- RSF
- RRF

Adicionalmente, se evaluó BM25 como referencia de recuperación léxica.

Los 16 modelos semánticos fueron evaluados utilizando el mismo conjunto de consultas y documentos, manteniendo constantes el corpus, la definición de relevancia, la profundidad del ranking y las métricas de evaluación, con el fin de garantizar una comparación justa entre arquitecturas de *embeddings* heterogéneas.

Índice	Modelo
1	<i>Qwen/Qwen3-Embedding-8B</i>
2	intfloat/multilingual-e5-large
3	BAAI/bge-m3
4	intfloat/multilingual-e5-base
5	intfloat/multilingual-e5-small
6	intfloat/e5-mistral-7b-instruct
7	sentence-transformers/paraphrase-multilingual-mpnet-base-v2
8	sentence-transformers/use-cmlm-multilingual
9	somosnlp-hackathon-2022/paraphrase-spanish-distilroberta
10	hiiamsid/sentence_similarity_spanish_es
11	sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2
12	sentence-transformers/LaBSE
13	sentence-transformers/distiluse-base-multilingual-cased-v2
14	sentence-transformers/all-MiniLM-L6-v2
15	dccuchile/bert-base-spanish-wwm-cased
16	bert-base-multilingual-cased

El conjunto de modelos de este experimento se definió mediante un criterio distinto al empleado en el Experimento 1. En aquel caso se incluyeron 17 modelos de diferentes familias y tamaños para caracterizar la robustez de las representaciones semánticas ante errores ortográficos y analizar la dependencia respecto de la arquitectura. En el Experimento 3 se seleccionaron 16 arquitecturas heterogéneas documentadas en la literatura especializada y en *benchmarks* públicos de *embeddings*, incluyendo modelos multilingües y modelos entrenados específicamente en español. Estos últimos se incorporaron para explorar si la especialización lingüística ofrecía una ventaja en la representación semántica de las consultas; sin embargo, los resultados no mostraron una superioridad sistemática frente a los modelos multilingües. Por tanto, la diferencia entre ambos conjuntos corresponde a una decisión metodológica asociada con el objetivo de cada experimento y no al descarte accidental de un modelo.

Para la evaluación complementaria de la ponderación ajustable se utilizó MessIRve en su configuración full/test. Se construyeron tres muestras disjuntas de 5000 pares consulta-documento mediante las semillas 2024, 2025 y 2026. Para cada modelo y semilla se evaluó una rejilla de α entre 0.00 y 1.00, con incrementos de 0.05 y normalización min-max por consulta. La selección de α se realizó mediante validación cruzada anidada de cinco *folds*: el peso se eligió únicamente con los *folds* de entrenamiento y se evaluó en el *fold* externo. La incertidumbre de la diferencia pareada de MRR frente a las referencias se estimó con 10 000 réplicas *bootstrap*; adicionalmente, se aplicó la prueba de Wilcoxon con tratamiento Pratt de los ceros y corrección de Holm para las comparaciones confirmatorias. Las consultas se identificaron mediante un índice único dentro de cada muestra para evitar duplicaciones durante la reorganización de los resultados.

Este diseño experimental permite evaluar el efecto de las distintas estrategias de recuperación sobre la calidad del ranking y comparar su desempeño en un conjunto diverso de modelos de representación semántica, proporcionando además evidencia sobre la estabilidad de la combinación lineal propuesta frente a arquitecturas heterogéneas.

5.3.5 Experimento 4. Evaluación integral del sistema RAG

El objetivo de este experimento fue evaluar el desempeño del sistema RAG completo, considerando tanto la calidad de la recuperación como la generación de respuestas. A diferencia de los experimentos anteriores, en este caso se evaluó el comportamiento del sistema de extremo a extremo, incluyendo la recuperación de documentos, la construcción del contexto y la generación de la respuesta.

Se utilizaron dos conjuntos de datos independientes.

Para la evaluación comparativa del pipeline RAG se empleó el corpus Danielbrdz/BarcenaxMexico [6]. Los registros relacionados con la historia de México se identificaron mediante un filtro de términos aplicado a los campos textuales; posteriormente se seleccionaron, con semilla 42, hasta 3000 registros y se eliminaron los pares pregunta-respuesta duplicados, con lo que se obtuvieron 587 pares únicos. Mediante una partición reproducible de 80 % para entrenamiento y 20 % para evaluación se conformó el subconjunto

de 118 consultas. En concordancia con la configuración general del sistema, ambas implementaciones se evaluaron con una profundidad de recuperación $k = 10$; para la comparación directa se mantuvo $\alpha = 0.5$, donde α representa el peso asignado a BM25.

Adicionalmente, para evaluar el desempeño del sistema de recuperación se utilizó un conjunto independiente de 3487 consultas, empleado para analizar la calidad de la recuperación y la consistencia de la evaluación automática mediante múltiples jueces LLM.

Se compararon dos implementaciones del pipeline RAG:

- Custom, correspondiente a la implementación desarrollada en este trabajo.
- LangChain, utilizando la implementación proporcionada por dicha biblioteca.

Asimismo, se evaluó el desempeño utilizando dos modelos generativos:

- Phi-3 Mini, empleado como modelo principal del sistema.
- GPT-4o Mini, utilizado como referencia para comparar el comportamiento del pipeline con un modelo propietario de mayor capacidad.

Las métricas consideradas fueron:

- Recuperación: Hit@k, MRR@K y nDCG@K.
- Calidad de las respuestas: ROUGE-1 y ROUGE-L.
- Evaluación automática: métricas agregadas obtenidas mediante múltiples jueces LLM.

La evaluación automática se realizó mediante un *prompt* diseñado para tareas de LLM-as-a-Judge, el cual especifica los criterios de evaluación, las reglas de decisión y el formato de salida requerido. El *prompt* completo se presenta en el Anexo A.

Este diseño experimental permitió comparar el desempeño de ambas implementaciones del pipeline, analizar la influencia del modelo generativo y evaluar la calidad de las respuestas producidas por el sistema RAG.

5.3.6 Experimento 5. Evaluación con usuarios

El objetivo de este experimento fue evaluar el impacto del sistema RAG sobre la interacción de los usuarios durante una tarea de recuperación de información histórica. En particular, se analizó si el uso de consultas en lenguaje natural influye en la carga cognitiva, el tiempo de respuesta y el desempeño en la recuperación de información.

Se realizó un estudio comparativo entre dos modalidades de acceso a la información:

- **Museo virtual**, en el que los participantes únicamente podían consultar la exhibición y el panel informativo correspondiente.

- **Sistema RAG**, que además permitía realizar consultas en lenguaje natural sobre el contenido de las exhibiciones.

Las exhibiciones correspondieron a pinturas del Museo Nacional de Bellas Artes de Argentina relacionadas con la conquista de México. El corpus utilizado por el sistema RAG se construyó a partir de la información del museo, comentarios de historiadores y contenido complementario de Wikipedia.

La muestra estuvo conformada por 90 participantes procedentes de grupos escolares facilitados por profesores del Instituto Tecnológico de Zacatepec (ITZ). Dentro de cada grupo disponible, las condiciones se asignaron de forma alternada, procurando distribuir aproximadamente la mitad de los estudiantes al sistema RAG y la otra mitad al museo virtual. No se realizó un balance posterior ni se eliminaron registros para forzar tamaños iguales: se conservaron los 90 registros completos obtenidos, lo que produjo 41 participantes en la condición RAG y 49 en la condición de museo virtual. Los participantes tenían entre 18 y 28 años y provenían de distintas licenciaturas; quienes utilizaron el sistema RAG realizaron en conjunto 220 consultas.

Por la disponibilidad de los grupos y el procedimiento de asignación alternada, el estudio corresponde a un diseño cuasiexperimental y no a un ensayo con asignación aleatoria individual. Esta condición debe considerarse al interpretar las comparaciones entre grupos.

Para la evaluación se emplearon los siguientes instrumentos:

- Plataforma web con las dos modalidades de acceso.
- Cuestionario de comprensión integrado por diez preguntas (cinco literales y cinco inferenciales).
- Cuestionario de carga cognitiva de *Paas*.

El experimento evaluó tres variables principales: carga cognitiva, tiempo de respuesta y desempeño, medido mediante la proporción de respuestas correctas (MCRC). Para ello se plantearon dos grupos de hipótesis.

El primer grupo corresponde a la comparación entre las dos modalidades de acceso (museo virtual y sistema RAG), contrastando la igualdad de medias para la carga cognitiva, el tiempo de respuesta y el desempeño.

El segundo grupo analiza la relación entre las variables mediante pruebas de correlación, evaluando la asociación entre carga cognitiva y tiempo de respuesta, carga cognitiva y desempeño, así como entre desempeño y tiempo de respuesta.

En todos los casos se formularon hipótesis nulas de igualdad de medias o ausencia de correlación, junto con sus correspondientes hipótesis alternativas de diferencia entre medias o existencia de correlación.

Debido a que se realizaron siete contrastes de hipótesis sobre la misma muestra —cuatro comparaciones entre grupos y tres análisis de correlación—, se aplicó el procedimiento secuencial de Holm para controlar el incremento del riesgo de error tipo I asociado con las comparaciones múltiples. Se conservaron los valores p originales y se reportaron adicionalmente los valores p ajustados. Las decisiones estadísticas se establecieron con un nivel de significancia $\alpha = 0.05$ aplicado a los valores ajustados.

Capítulo VI. Resultados y discusión

En este capítulo se presentan los resultados obtenidos durante la validación experimental del sistema propuesto. Los resultados se organizan de acuerdo con los objetivos específicos planteados en el Capítulo I. En primer lugar, se muestran los resultados relacionados con la evaluación del módulo de recuperación de información. Posteriormente se presentan los resultados de la evaluación del modelo generativo y, finalmente, se analiza el desempeño del sistema RAG completo mediante evaluaciones automáticas.

6.1 Resultados de la revisión contextual sobre comunicación museográfica

El análisis de la literatura permitió identificar diversos factores que dificultan la comunicación efectiva de la información en los museos mexicanos. Estos problemas involucran tanto aspectos lingüísticos como elementos relacionados con el diseño y la accesibilidad de los materiales museográficos.

Uno de los problemas más frecuentes es el uso de lenguaje altamente especializado. Diversos estudios coinciden en que las cédulas y paneles informativos suelen ser elaborados por especialistas, lo que da lugar a textos con terminología técnica y estructuras sintácticas complejas que dificultan su comprensión por parte del público general.

Asimismo, se identificaron deficiencias en la organización del contenido. En muchos casos, las cartelas presentan información extensa y con escasa jerarquización, lo que dificulta localizar las ideas principales y favorece una lectura superficial.

Desde el punto de vista del diseño, la literatura también reporta problemas relacionados con la legibilidad visual, entre ellos el uso de tamaños de letra reducidos, bajo contraste entre texto y fondo y una selección tipográfica poco adecuada, factores que incrementan el esfuerzo requerido para la lectura.

En relación con los recursos gráficos, diversos trabajos señalan que las infografías y otros elementos visuales no siempre cumplen una función complementaria al texto. La ausencia de etiquetas claras o de explicaciones suficientes limita su capacidad para apoyar la interpretación de la información presentada.

Finalmente, se identificaron limitaciones en materia de accesibilidad. La falta de formatos alternativos, recursos táctiles y otras adaptaciones reduce las posibilidades de acceso a la información para personas con discapacidad visual, restringiendo el alcance inclusivo de las exposiciones museográficas.

Los resultados de esta revisión permitieron establecer las principales necesidades del dominio de aplicación y evidenciaron que existe una limitación importante asociada a la naturaleza estática de la información museográfica, la cual restringe la profundización y exploración de los temas por parte de los visitantes. A partir de esta limitación, se plantea la necesidad de abordar el problema mediante el uso de herramientas que permitan ofrecer información dinámica, lo que sirvió como fundamento para la selección de las estrategias de recuperación de información, los modelos de lenguaje y los criterios de evaluación considerados en las etapas posteriores de la investigación.

6.2 Experimento 1. Robustez ante errores ortográficos

La Tabla 6.1 muestra el valor óptimo del parámetro α obtenido para cada familia de modelos. Se observa que los modelos E5 e Instruct seleccionaron predominantemente $\alpha=0.3$, mientras que los modelos BGE y MiniLM obtuvieron mejores resultados con $\alpha=0.5$. Por su parte, los modelos BERT base presentaron el mayor MRR con $\alpha=0.7$.

Familia de modelos	α óptimo predominante
E5 / Instruct	0.3
BGE	0.5
BERT base	0.7
Multilingual MiniLM	0.5

Tabla 6.1. Valor óptimo de alfa por familia que indica el mejor ajuste para la métrica MRR

La Figura 6.1 compara la distribución del MRR con dos y cuatro errores ortográficos. Con dos errores, los métodos basados en *embeddings* y el híbrido presentan medianas próximas a 0.45 y rangos aproximados de 0.42–0.49 y 0.41–0.48, respectivamente. Con cuatro errores, las medianas permanecen cercanas, pero el límite inferior del método basado únicamente en *embeddings* desciende hasta aproximadamente 0.38, mientras que el híbrido se mantiene alrededor de 0.41; los límites superiores continúan próximos a 0.49. Este patrón sugiere que el aumento del ruido afecta principalmente los casos de peor desempeño del recuperador semántico, mientras la combinación híbrida reduce parcialmente esa caída. DL aparece como un único valor por nivel de ruido porque no depende del modelo de *embeddings*. Estas diferencias son descriptivas y no sustituyen una prueba inferencial.

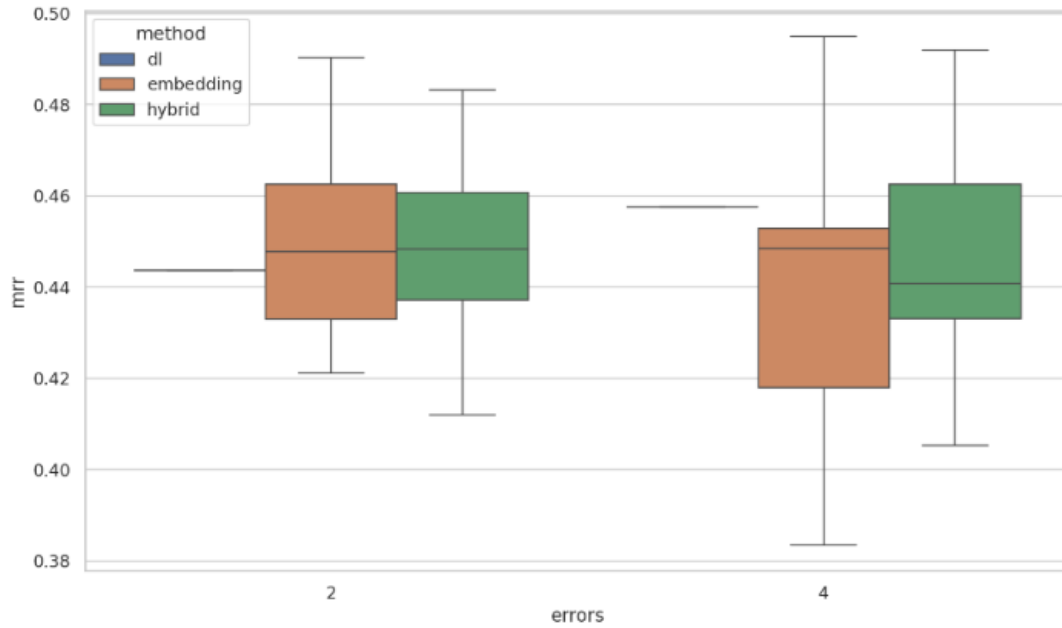


Figura 6.1. Distribución de MRR por método y nivel de ruido ortográfico

Las Tablas 6.2 y 6.3 presentan los resultados obtenidos con dos y cuatro errores ortográficos, respectivamente. En ambos escenarios, las diferencias entre métodos fueron reducidas. Tanto DL como el método híbrido alcanzaron *Recall@10* igual a 1.0, mientras que el método basado únicamente en *embeddings* obtuvo valores ligeramente inferiores.

MÉTODO	RECALL@1	RECALL@5	RECALL@10	MRR	NDCG@10
DL	0.2000	0.8333	1.0000	0.4437	0.5778
EMBEDDING	0.2078	0.8402	0.9980	0.4490	0.5813
HÍBRIDO	0.2029	0.8373	1.0000	0.4497	0.5825

Tabla 6.2. Resultados promedio para consultas con 2 errores ortográficos.

MÉTODO	RECALL@1	RECALL@5	RECALL@10	MRR	NDCG@10
DL	0.2167	0.8500	1.0000	0.4576	0.5885
EMBEDDING	0.1892	0.8333	0.9912	0.4513	0.5707
HÍBRIDO	0.2010	0.8490	1.0000	0.4426	0.5795

Tabla 6.3. Resultados promedio para consultas con 4 errores ortográficos.

Finalmente, la Tabla 6.4 muestra los modelos con mejor y peor desempeño. El mayor MRR promedio correspondió a *e5-mistral-7b-instruct* (0.486), seguido de *Qwen3-Embedding-8B*

(0.478) y *bge-large-en-v1.5* (0.469). Los menores valores fueron obtenidos por *bert-base-multilingual-cased*, *bert-base-spanish-wwm-cased* y *paraphrase-spanish-distilroberta*.

CATEGORÍA	MODELO	MRR PROMEDIO
MEJORES	e5-mistral-7b-instruct	0.486
	<i>Qwen3-Embedding-8B</i>	0.478
	bge-large-en-v1.5	0.469
PEORES	bert-base-multilingual-cased	0.381
	bert-base-spanish-wwm-cased	0.374
	paraphrase-spanish-distilroberta	0.368

Tabla 6.4. Modelos con mejor y peor desempeño global medido mediante MRR promedio en alfa

6.3 Experimento 2. Evaluación de modelos generativos

La Tabla 6.5 presenta el desempeño global de las tres configuraciones de *prompting* evaluadas mediante la métrica MCRC.

La configuración 2 obtuvo el mejor desempeño, alcanzando 3880 aciertos de un total de 5700 preguntas, equivalente a un MCRC de 68.07 %. En comparación, la configuración 3 obtuvo un MCRC de 49.56 %, mientras que la configuración 1 registró el menor desempeño con 44.61 %.

CONFIGURACIÓN	ACIERTOS	TOTAL	MCRC
CONF 1	2543	5700	44.61 %
CONF 2	3880	5700	68.07 %
CONF 3	2825	5700	49.56 %

Tabla 6.5. Desempeño global por configuración medido con MCRC.

La Tabla 6.6 presenta el número de respuestas correctas obtenidas por cada modelo en las tres configuraciones evaluadas, así como el MCRC global.

Los mayores valores de MCRC correspondieron a *Falcon3-7B-Instruct* (62.22 %), *Qwen2-7B-Instruct* (61.87 %) y *Phi-4-mini-instruct* (61.64 %). En contraste, los menores desempeños fueron obtenidos por *Qwen2.5-3B-Instruct* (39.24 %) y *Yi-6B-Chat* (42.34 %).

Se observa además que la configuración 2 produjo el mayor número de respuestas correctas para todos los modelos evaluados, independientemente de la arquitectura utilizada.

MODELO	CONF 1	CONF 2	CONF 3	TOTAL ACIERTOS	TOTAL PREGUNTAS	MCRC GLOBAL
01-AI/YI-6B-CHAT	159	330	235	724	1710	42.34 %
QWEN/QWEN2-7B-INSTRUCT	324	403	331	1058	1710	61.87 %
QWEN/QWEN2.5-3B-INSTRUCT	17	368	286	671	1710	39.24 %
META-LLAMA/LLAMA-3.1-8B-INSTRUCT	258	390	275	923	1710	53.98 %
MICROSOFT/PHI-3-MINI-128K-INSTRUCT	332	390	266	988	1710	57.78 %
MICROSOFT/PHI-3-MINI-4K-INSTRUCT	301	409	298	1008	1710	58.95 %
MICROSOFT/PHI-4-MINI-INSTRUCT	274	495	285	1054	1710	61.64 %
MISTRALAI/MISTRAL-7B-INSTRUCT-V0.3	268	317	262	847	1710	49.53 %
OPENCHAT/OPENCHAT_3.5	268	382	261	911	1710	53.27 %
THUAE/FALCON3-7B-INSTRUCT	342	396	326	1064	1710	62.22 %

Tabla 6.6. Total de aciertos y MCRC global de los modelos por configuración de prompting

6.4 Experimento 3. Evaluación de estrategias de recuperación

La Tabla 6.7 presenta el desempeño promedio de los métodos evaluados considerando los 16 modelos de *embeddings*. La coincidencia entre Lineal ($\alpha = 0.50$) y RSF no corresponde a una duplicación de filas: bajo la misma normalización, ambos calculan $0.50 \cdot \text{sBM25} + 0.50 \cdot \text{ssemántico}$. En consecuencia, los valores idénticos de MRR (0.6601 ± 0.0871), *Recall@1* (0.5605), *Recall@5* (0.7825) y *Recall@10* (0.8360) constituyen un control de implementación esperado, no evidencia de superioridad de una estrategia sobre la otra.

MÉTODO	MRR ($M \pm \Sigma$)	RECALL@1 (μ)	RECALL@5 (μ)	RECALL@10 (μ)
LINEAL (ALFA = 0.50)	0.6601 $\pm$ 0.0871	0.5605	0.7825	0.8360
RSF (CONTRO L EQUIVAL ENTE)	0.6601 $\pm$ 0.0871	0.5605	0.7825	0.8360
LINEAL (ALFA = 0.70)	0.6327 $\pm$ 0.0345	0.5433	0.7361	0.7787

LINEAL (ALFA = 0.30)	0.6263 $\pm$ 0.1729	0.5383	0.7289	0.7779
RRF	0.5796 $\pm$ 0.1143	0.4799	0.6845	0.7655
BM25	0.5594	0.4700	0.6628	0.7136
EMBEDDING	0.5196 $\pm$ 0.2422	0.4391	0.6089	0.6624

Tabla 6.7. Desempeño promedio por método en el contraste principal

La equivalencia se verificó también a nivel de consulta: RSF y Lineal ($\alpha = 0.50$) produjeron cero discrepancias de ranking en las 48 combinaciones de modelo y semilla. Para evaluar la aportación real del método, α se seleccionó dentro de cada partición de entrenamiento y se aplicó después al *fold* externo, evitando escoger y evaluar el peso con las mismas observaciones.

La Figura 6.2 muestra el contraste descriptivo del experimento inicial. El empate entre RSF y Lineal ($\alpha = 0.50$) se mantiene visible como control, mientras que la evaluación fuera de muestra presentada a continuación permite determinar si otros valores de α aportan una mejora reproducible.

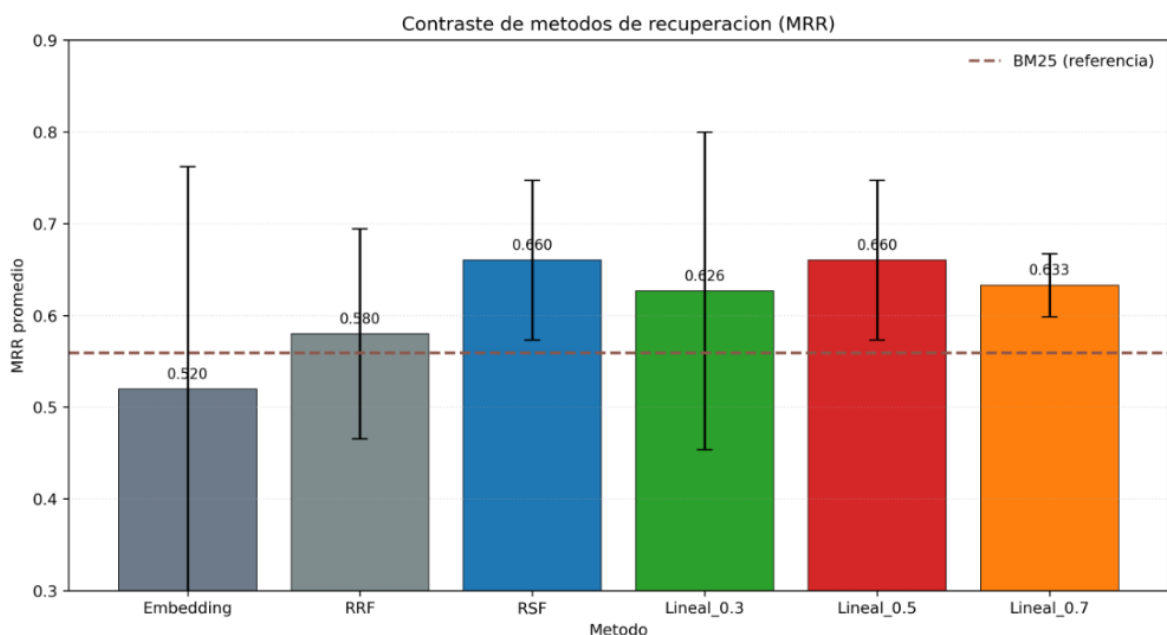


Figura 6.2. MRR promedio por método con dispersión; la línea punteada indica la referencia BM25

Verificación de equivalencia y evaluación de la ponderación ajustable

La igualdad observada responde a la identidad $RSF = \text{Lineal}$ ($\alpha = 0.50$), no a un error del pipeline. La hipótesis relevante es si permitir $\alpha \neq 0.50$ mejora la recuperación en datos no utilizados para elegir el peso. El experimento complementario evaluó 16 modelos sobre tres muestras de 5000 consultas y seleccionó α mediante validación cruzada anidada de cinco folds. La estabilidad del peso seleccionado y la variación del MRR respecto de RSF se resumen en la Tabla 6.8.

MODELO	A SELECCIONADO	ΔMRR PROMEDIO	SEMILLAS ROBUSTAS
MULTILINGUAL-E5-LARGE	0.05–0.10	+0.1582	3/3
MULTILINGUAL-E5-BASE	0.05–0.10	+0.1444	3/3
BGE-M3	0.05–0.10	+0.1414	3/3
QWEN3-EMBEDDING-8B	0.00–0.10	+0.1371	3/3
MULTILINGUAL-E5-SMALL	0.10–0.15	+0.1369	3/3
E5-MISTRAL-7B	0.30–0.35	+0.0277	3/3
MULTILINGUAL-MPNET	0.30–0.35	+0.0252	3/3
USE-CMLM	0.35	+0.0240	3/3
SPANISH DISTILROBERTA	0.35–0.40	+0.0176	3/3
SENTENCE SIMILARITY ES	0.40–0.45	+0.0055	3/3
BERT MULTILINGUAL	0.55–0.65	+0.0058	2/3
MINILM MULTILINGUAL	0.40–0.45	+0.0049	2/3
LABSE	0.35–0.45	+0.0042	2/3
ALL-MINILM-L6-V2	0.40–0.45	+0.0019	1/3
DISTILUSE MULTILINGUAL	0.45–0.50	+0.0008	0/3
BERT ESPAÑOL	0.40–0.50	–0.0006	0/3

Tabla 6.8. Estabilidad del peso seleccionado y cambio de MRR respecto a RSF

Efecto de seleccionar α mediante validación cruzada anidada

Punto: Δ MRR promedio en tres semillas; segmento: envolvente de los IC95% por semilla

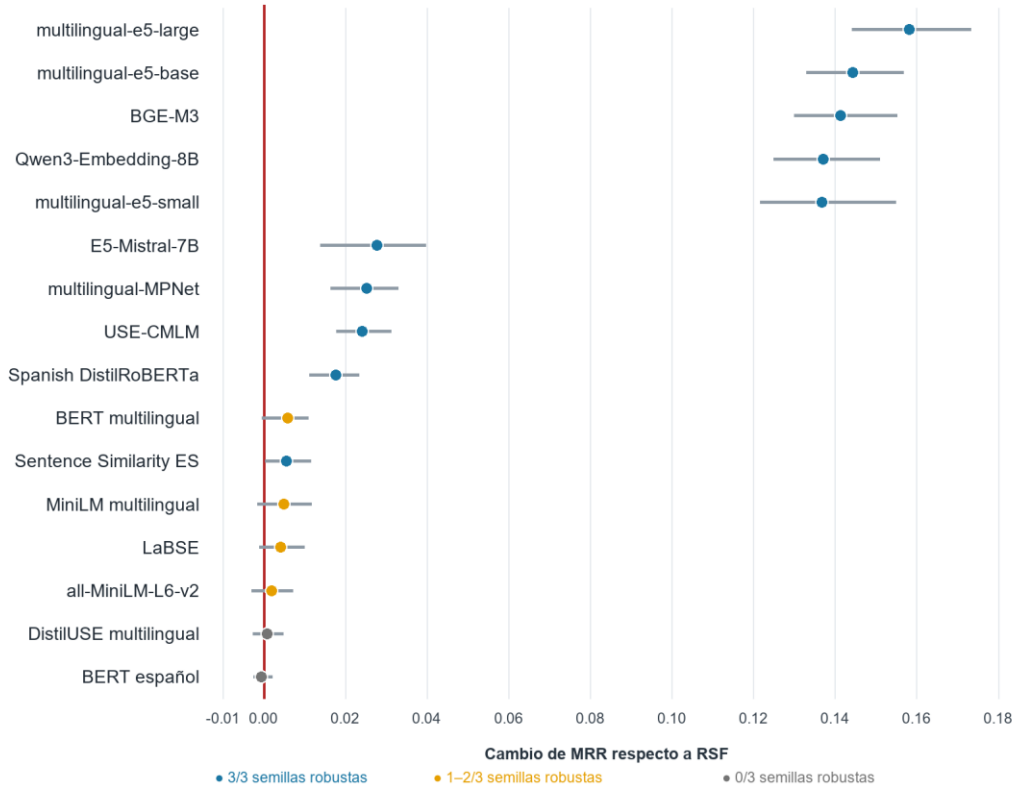


Figura 6.3. Cambio de MRR de la ponderación ajustada respecto a RSF por modelo

La Figura 6.3 muestra que el efecto no es uniforme. Los modelos E5, BGE-M3 y Qwen3 se benefician de asignar un peso reducido a BM25 (α entre 0.05 y 0.15), mientras que otros modelos seleccionan valores próximos a 0.50 y presentan mejoras pequeñas o no robustas. En conjunto, 43 de 48 escenarios mostraron una variación positiva de MRR y 37 mostraron una mejora robusta. Por tanto, los datos respaldan la adaptabilidad del peso de fusión, pero no una superioridad universal frente a RSF.

La Tabla 6.9 presenta la ganancia relativa de cada método respecto a BM25.

La mayor mejora correspondió a *Lineal*_{0.5}, con incrementos de 0.1007 en MRR, 0.0905 en Recall@1 y 0.1224 en Recall@10. Las variantes *Lineal*_{0.3} y *Lineal*_{0.7} también mostraron mejoras positivas respecto a la referencia. En contraste, el método basado únicamente en *embeddings* presentó una disminución en las tres métricas evaluadas.

MÉTODO	Δ MRR BM25	VS Δ RECALL@1 BM25	VS Δ RECALL@10 BM25
LINEAL_0.5	+0.1007	+0.0905	+0.1224

LINEAL_0.7	+0.0733	+0.0733	+0.0651
LINEAL_0.3	+0.0669	+0.0683	+0.0643
RRF	+0.0202	+0.0099	+0.0519
EMBEDDING	-0.0398	-0.0309	-0.0512

Tabla 6.9. Ganancia promedio respecto a la referencia léxica pura (BM25)

La Tabla 6.10 presenta los cinco modelos que obtuvieron el mayor MRR durante la evaluación.

El mejor resultado correspondió a Qwen3-Embedding-8B (0.8699), seguido por multilingual-e5-large (0.8596), BGE-M3 (0.8454), multilingual-e5-base (0.8275) y multilingual-e5-small (0.8110).

MODELO	MEJOR MRR OBSERVADO
QWEN/QWEN3-EMBEDDING-8B	0.8699
INTFLOAT/MULTILINGUAL-E5-LARGE	0.8596
BAAI/BGE-M3	0.8454
INTFLOAT/MULTILINGUAL-E5-BASE	0.8275
INTFLOAT/MULTILINGUAL-E5-SMALL	0.8110

Tabla 6.10. Top 5 modelos según su mejor MRR.

La Figura 6.4 muestra la matriz de calor correspondiente al MRR obtenido por cada modelo y método de recuperación.

Se observa una diferencia importante entre modelos de *embeddings*. Los mayores valores de MRR corresponden a Qwen3-Embedding-8B, multilingual-e5-large y BGE-M3, mientras que bert-base-multilingual-cased y bert-base-spanish-wwm-cased presentan los menores desempeños. Asimismo, la matriz permite apreciar que el efecto del método de recuperación depende del modelo utilizado, ya que la configuración con mejor desempeño varía entre arquitecturas.

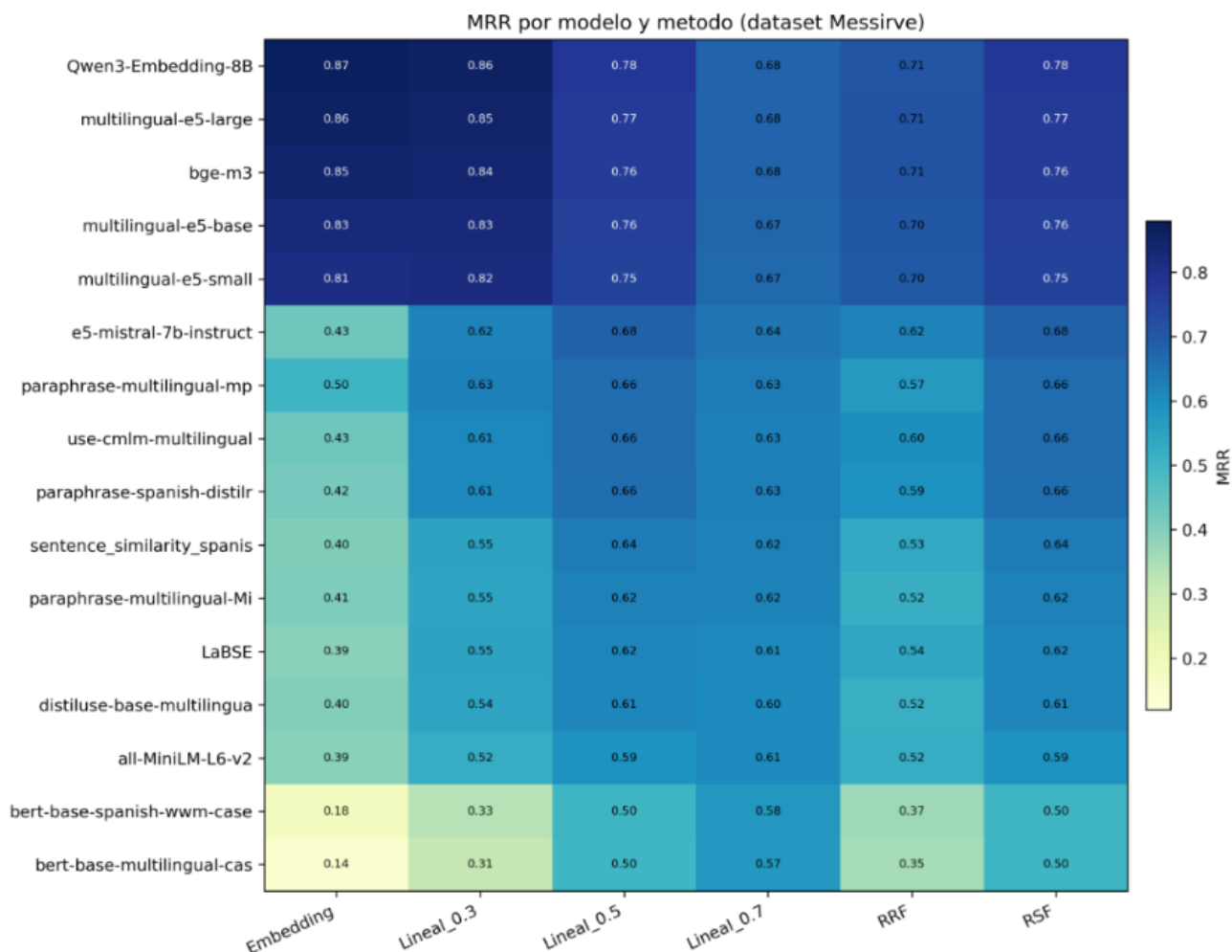


Figura 6.4. Matriz MRR por modelo y método para identificar estabilidad y casos extremos

6.5 Experimento 4. Evaluación integral del sistema RAG

La Tabla 6.11 presenta las métricas de recuperación obtenidas al reproducir la evaluación del subconjunto de 118 consultas con $k = 10$ en ambas implementaciones.

Con $\alpha = 0.5$, la implementación custom obtuvo un $\text{Hit}@10$ de 0.6525, un $\text{MRR}@10$ de 0.3901 y un $\text{nDCG}@10$ de 0.4532. Por su parte, la implementación basada en LangChain alcanzó valores de 0.5254, 0.2179 y 0.2882, respectivamente.

Asimismo, la implementación custom presentó mayores valores de $\text{ContextPrecision}@10$ (0.0653) y $\text{ContextRecall}@10$ (0.6525) que LangChain (0.0525 y 0.5254). Dado que cada consulta tiene un único documento relevante, $\text{ContextRecall}@10$ coincide con $\text{Hit}@10$ y $\text{ContextPrecision}@10$ equivale a $\text{Hit}@10/10$; estas igualdades corresponden a la definición de las métricas y no a una duplicación de resultados.

PIPELINE	HIT@10	MRR@10	NDCG@10	CONTEXTPRECISION@10	CONTEXTRECALL@10
CUSTOM	0.6525	0.3901	0.4532	0.0653	0.6525
LANGCHAIN	0.5254	0.2179	0.2882	0.0525	0.5254

Tabla 6.11. Resultados de recuperación a $k = 10$ en el subconjunto de evaluación ($n = 118$)

La Tabla 6.12 muestra los resultados de generación medidos mediante las métricas ROUGE-1 y ROUGE-L.

La combinación *custom* + GPT-4o-mini obtuvo los mayores valores de ROUGE-L (0.2008) y ROUGE-1 (0.3279). En segundo lugar, se ubicó *custom* + Phi-3-mini, mientras que ambas configuraciones implementadas con *LangChain* presentaron valores inferiores.

PIPELINE	LLM	ROUGE-1	ROUGE-L
CUSTOM	gpt-4o-mini	0.3279	0.2008
CUSTOM	phi3-mini	0.2806	0.1713
LANGCHAIN	phi3-mini	0.2586	0.1519
LANGCHAIN	gpt-4o-mini	0.2578	0.1537

Tabla 6.12. Resultados de generación (ROUGE) en el subconjunto de evaluación ($n = 118$)

La Tabla 6.13 presenta el desempeño del sistema de recuperación sobre el conjunto completo de 3487 consultas.

El sistema obtuvo un Hit@10 y Recall@10 de 0.9338, un MRR@10 de 0.5527 y un nDCG@10 de 0.6441.

N	HIT@10	MRR@10	NDCG@10	RECALL@10
3487	0.9338	0.5527	0.6441	0.9338

Tabla 6.13. Evaluación del sistema de recuperación sobre el conjunto completo de consultas.

Se evaluó el conjunto de 3487 consultas utilizando seis modelos de lenguaje como evaluadores automáticos. Para cada registro se analizó la presencia del documento de referencia entre los documentos recuperados (GT@10), la corrección de la respuesta, su comprensibilidad, la presencia de *refusals*, la corrección de estos y la existencia de alucinaciones.

La Tabla 6.14 resume los resultados obtenidos por cada evaluador. En todos los casos, la métrica GT@10 presentó el mismo valor (3256 registros), mientras que las restantes variables mostraron diferencias entre evaluadores.

EVALUADOR	N	GT@10	CORRECTAS	COMPENSIBLES	REFUSALS	REFUSAL CORR.	ALUCINACIONES
GPT-5.3	3487	3256	3418	3481	42	31	6
GEMINI	3487	3256	3200	3095	76	20	211
MISTRAL	3487	3256	3203	3424	63	10	221
KIMI	3487	3256	3195	3487	71	10	221
DEEPSEEK	3487	3256	3203	3487	63	10	221
GPT-5.4	3487	3256	3192	3487	74	10	221

Tabla 6.14. Resultados agregados reportados por los evaluadores LLM.

Posteriormente se calculó el coeficiente de kappa de Fleiss utilizando los evaluadores seleccionados para el análisis de acuerdo. Los resultados se presentan en la Tabla 6.15.

VARIABLE	KAPPA DE FLEISS	NIVEL DE ACUERDO (INTERPRETACIÓN)
GT_IN_TOP10	0.788	Acuerdo sustancial
REFUSAL	0.762	Acuerdo sustancial
ANSWER_LABEL	0.649	Acuerdo moderado / sustancial
REFUSAL_TYPE	0.438	Acuerdo moderado
DEMOGRAPHIC_OK	0.004	Sin acuerdo (puro azar)
HALLUCINATION	-0.107	Acuerdo inferior al azar (desacuerdo)
CONTEXT_SUPPORTED	-0.128	Acuerdo inferior al azar (desacuerdo)

Tabla 6.15. Acuerdo interevaluador medido mediante kappa de Fleiss en las variables de evaluación

Las mayores concordancias se obtuvieron en GT_in_top10 ($\kappa = 0.788$), variable que indica si el documento de referencia fue recuperado dentro de los diez primeros resultados, y en *refusal* ($\kappa = 0.762$), que evalúa si el modelo identifica correctamente que no existe información suficiente para responder. Asimismo, *answer_label* ($\kappa = 0.649$), correspondiente

a la clasificación global de la respuesta como correcta o incorrecta, presentó un nivel de acuerdo sustancial.

Por otra parte, *refusal_type* ($\kappa = 0.438$), que distingue entre rechazos correctos e incorrectos, mostró un acuerdo moderado. Las variables *demographic_ok* ($\kappa = 0.004$), utilizada para evaluar si la respuesta es comprensible para el perfil demográfico proporcionado, *hallucination* ($\kappa = -0.107$), que identifica información no respaldada por el contexto recuperado, y *context_supported* ($\kappa = -0.128$), que determina si la respuesta está sustentada por los documentos recuperados, presentaron los menores niveles de acuerdo entre los evaluadores. (Para más información de las variables consultar el Anexo A)

Los resultados relacionados con la detección de alucinaciones deben interpretarse con cautela. GPT-5.3 identificó 6 casos, Gemini 211 y Mistral, Kimi, DeepSeek y GPT-5.4 identificaron 221 casos cada uno. Esta dispersión, junto con el coeficiente de kappa de Fleiss negativo obtenido para la variable *hallucination* ($\kappa = -0.107$), indica que los evaluadores no aplicaron de manera consistente el criterio de información no respaldada por el contexto. En consecuencia, estos conteos se conservan como resultados individuales de los jueces, pero no se agregan ni se interpretan como una estimación confiable de la tasa real de alucinación del sistema.

Por esta razón, los resultados no permiten sostener que el sistema presente un nivel bajo de alucinaciones. Tampoco se adoptó una votación mayoritaria, debido a que el acuerdo inferior al azar no ofrece evidencia suficiente para considerar que la decisión más frecuente constituya una aproximación válida a una anotación de referencia. La construcción de un conjunto de evaluación anotado manualmente y la calibración de los jueces respecto de dicho conjunto se plantean como trabajo futuro.

La variable de comprensibilidad presentó una limitación semejante. Kimi, DeepSeek y GPT-5.4 clasificaron como comprensibles las 3487 respuestas, mientras que los demás evaluadores mostraron cierta variación. Debido a que estos tres jueces asignaron la misma clasificación positiva a todos los casos, sus resultados no permitieron distinguir entre respuestas con diferentes grados de claridad. Esta falta de variación, junto con el valor de *demographic_ok* ($\kappa = 0.004$), impide interpretar los porcentajes elevados como evidencia concluyente de que todas las respuestas sean comprensibles.

Finalmente, los evaluadores utilizados corresponden a servicios de modelos de lenguaje accesibles mediante infraestructura en la nube. Sus resultados pueden verse afectados por actualizaciones de los modelos, cambios de versión, modificaciones de los proveedores y variaciones en la infraestructura de inferencia. Por ello, los valores obtenidos describen la ejecución realizada durante el experimento, pero no garantizan una reproducción numérica exacta mediante ejecuciones posteriores. Esta limitación refuerza la necesidad de conservar las salidas originales y complementar las evaluaciones automáticas con anotaciones humanas.

6.6 Experimento 5. Evaluación con usuarios

La Tabla 6.16 presenta la distribución de los participantes que integraron las dos condiciones experimentales. En total participaron 90 usuarios, distribuidos entre el grupo que utilizó el sistema RAG y el grupo que utilizó el museo virtual. Asimismo, se muestran las características generales de sexo y edad media de ambos grupos.

GRUPO	PARTICIPANTES	MASCULINOS	FEMENINOS	EDAD MEDIA
RAG	41	26	15	20.98
MUSEO_VIRTUAL	49	27	22	20.88

Tabla 6.16. Distribución de participantes por grupo experimental

La Tabla 6.17 presenta los resultados de los cuatro contrastes entre grupos y sus valores p ajustados mediante el procedimiento de Holm. Después de controlar las siete comparaciones, únicamente el tiempo de exploración mantuvo una diferencia estadísticamente significativa entre ambos grupos (p ajustado = 0.0217). Las variables restantes no mostraron evidencia suficiente para rechazar la hipótesis nula.

Variable	Prueba	p original	p ajustado (Holm)	Decisión
Tiempo de exploración	Mann-Whitney	0.0031	0.0217	Rechazar H_0
Carga mental	Mann-Whitney	0.4846	1.0000	No rechazar H_0
MCRC	Mann-Whitney	0.1113	0.5988	No rechazar H_0
Eficiencia cognitiva	t independiente	0.2126	0.8504	No rechazar H_0

Tabla 6.17. Pruebas estadísticas y valores p ajustados mediante Holm

La Tabla 6.18 muestra los resultados del análisis de correlación mediante el coeficiente de Spearman y los valores p ajustados mediante Holm. Ninguna de las tres correlaciones evaluadas presentó significancia estadística después de la corrección por comparaciones múltiples.

Variables	Rho	p original	p ajustado (Holm)
Carga mental vs tiempo	0.0437	0.6828	1.0000
Carga mental vs MCRC	0.1746	0.0998	0.5988
MCRC vs tiempo	-0.0743	0.4865	1.0000

Tabla 6.18. Correlaciones y valores p ajustados mediante Holm

Se analizaron las consultas realizadas por los participantes durante el uso del sistema RAG. La Tabla 6.19 presenta la distribución de las consultas por categoría semántica, mientras que la Tabla 6.20 resume los lemas con mayor frecuencia de aparición.

Categoría	Cantidad	Porcentaje
Buscando o indagando más información	118	53.6%
Entender el texto	78	35.5%
Interactuar con el bot	15	6.8%
Irrelevante o fuera del tópico	6	2.7%
Condensar la información	3	1.4%
Total	220	100%

Tabla 6.19. Distribución de categorías de consultas realizadas al sistema RAG

Lema	Frecuencia
cortés	22
imagen	25
tabla	11
ser (fue, era, son, eran, está)	>30
llegada	7

Tabla 6.20. Frecuencia de lemas identificados en las consultas

Posteriormente, la Tabla 6.21 muestra los tópicos identificados mediante LDA y la Tabla 6.22 presenta los resultados obtenidos con los métodos de agrupamiento semántico.

Categoría	Tópico 1	Tópico 2	Tópico 3
Términos asociados	• quién • imagen • usaba • servía • pintura	• tabla • Cortés • era • hubo	• imagen • explica • significa • Cortés
Descripción del tópico	Agrupar consultas orientadas a identificar, describir e interpretar elementos visuales presentes en las obras del museo. También incluye preguntas sobre la función o utilidad de los objetos representados en las piezas.	Incluye consultas relacionadas con acontecimientos históricos, personajes y procesos descritos en las cédulas informativas de la exposición, especialmente aquellos vinculados con la llegada de Hernán Cortés y la conquista de México.	Reúne preguntas dirigidas a aclarar el significado de imágenes, textos o acontecimientos históricos, solicitando explicaciones adicionales para facilitar la comprensión del contenido museográfico.

Tabla 6.21. Tópicos identificados mediante modelado LDA

MÉTODO	MEJOR CONFIGURACIÓN	SILHOUETTE SCORE
K-MEANS + TF-IDF	k = 8	0.094
NMF + TF-IDF	8 temas	0.424

Tabla 6.22. Comparación de métodos para clustering semántico

La Tabla 6.21 resume tres tópicos representativos obtenidos mediante LDA para facilitar su interpretación cualitativa. Este resumen no corresponde al número de grupos evaluado en la comparación posterior. La Tabla 6.22 presenta un análisis exploratorio independiente: K-Means se evaluó con $k = 8$ y NMF con ocho temas sobre representaciones TF-IDF. Por ello, los valores de *silhouette* se utilizan únicamente para describir la separación interna obtenida por cada configuración y no para establecer una equivalencia directa entre los tópicos de LDA, los clústeres de K-Means y los componentes de NMF.

El valor de *silhouette* obtenido por K-Means (0.094) es cercano a cero, lo que indica que las consultas asignadas a distintos grupos presentan un grado importante de solapamiento y que los límites entre los clústeres no están claramente definidos. Por tanto, este resultado no aporta evidencia de una estructura de agrupamiento sólida y los clústeres de K-Means solo deben considerarse una descripción exploratoria del corpus, no categorías estables o generalizables. NMF obtuvo un valor mayor (0.424), compatible con una separación temática más clara en estos datos; sin embargo, este resultado tampoco constituye una validación externa de los temas identificados.

6.7 Discusión de resultados

Los resultados obtenidos permiten analizar el desempeño del sistema propuesto desde tres perspectivas complementarias: la recuperación de información, la generación de respuestas y la interacción de los usuarios con el sistema RAG. En conjunto, los experimentos muestran que la calidad de la recuperación constituye el principal factor que determina el comportamiento del sistema completo, mientras que el modelo generativo y la interfaz actúan sobre la información previamente recuperada.

Recuperación de información

Los experimentos 1 y 3 muestran que la incorporación de evidencia léxica y semántica produce un comportamiento más robusto que el empleo aislado de cualquiera de estos enfoques. En el primer experimento, el método híbrido mantuvo un desempeño estable aun cuando las consultas incorporaban errores ortográficos, mientras que la degradación observada en los *embeddings* fue reducida. Esto indica que los modelos modernos de *embeddings* poseen cierta tolerancia al ruido léxico y que la combinación con información léxica contribuye a mantener la recuperación del documento relevante.

En el tercer experimento se observó un comportamiento similar. Las variantes de combinación lineal superaron sistemáticamente a BM25 y al método basado únicamente en *embeddings* en las métricas MRR, Recall@1 y Recall@10. Particularmente, la configuración con $\alpha = 0.5$ obtuvo el mejor desempeño promedio, lo que sugiere que un equilibrio entre evidencia léxica y semántica resulta más adecuado que privilegiar exclusivamente uno de ambos componentes.

Estos resultados coinciden con la literatura reciente sobre recuperación híbrida en sistemas RAG, donde diversos autores señalan que la combinación de recuperación dispersa y recuperación densa mejora la capacidad para recuperar documentos relevantes frente al uso aislado de cualquiera de los dos enfoques. Sin embargo, a diferencia de trabajos que emplean estrategias de fusión por rangos como RRF o RSF, la propuesta presentada incorpora una combinación lineal ponderada cuya influencia puede ajustarse mediante el parámetro α , permitiendo analizar experimentalmente el efecto de distintos balances entre ambas evidencias.

Selección del modelo generativo

El segundo experimento mostró que el desempeño de un modelo generativo depende no solamente de su arquitectura, sino también del formato de interacción utilizado. La configuración basada en respuestas estructuradas obtuvo un incremento considerable respecto a las demás configuraciones para todos los modelos evaluados, lo que evidencia que la forma en que se formula la instrucción tiene un impacto directo sobre la capacidad del modelo para producir respuestas consistentes.

Asimismo, aunque Falcon3-7B-Instruct y Qwen2-7B-Instruct obtuvieron los mayores valores de MCRC, las diferencias entre los modelos más competitivos fueron relativamente reducidas. Por esta razón, la selección de Phi-3-mini como modelo principal resulta consistente con el objetivo de desarrollar un sistema que pueda ejecutarse en hardware con recursos limitados, manteniendo un desempeño competitivo sin depender de modelos propietarios de gran tamaño.

Evaluación integral del sistema RAG

Los resultados del cuarto experimento muestran que las mejoras obtenidas en el módulo de recuperación se reflejan directamente en el comportamiento del sistema RAG completo. La implementación desarrollada obtuvo mejores resultados que la implementación basada en LangChain tanto en recuperación como en generación, indicando que las decisiones adoptadas durante el diseño del recuperador tuvieron un efecto positivo sobre la calidad del contexto suministrado al modelo generativo.

Por otra parte, la evaluación mediante múltiples modelos de lenguaje mostró un alto grado de acuerdo para variables relativamente objetivas, como la presencia del documento de referencia entre los documentos recuperados o la identificación de respuestas sin evidencia suficiente. En contraste, las variables relacionadas con alucinaciones, respaldo contextual y comprensibilidad presentaron una concordancia insuficiente. En consecuencia, los resultados de estas últimas variables se interpretan como evidencia de las limitaciones de los jueces automáticos y no como una estimación concluyente de la incidencia de alucinaciones ni de la comprensibilidad general del sistema.

Este resultado coincide con diversos trabajos recientes que señalan que la evaluación mediante LLM-as-a-Judge resulta confiable para tareas relativamente objetivas, mientras que

la valoración de aspectos semánticos complejos continúa presentando una elevada variabilidad entre evaluadores.

Evaluación con usuarios

Los resultados del experimento con usuarios muestran que la incorporación del sistema RAG modificó principalmente el comportamiento de exploración de los participantes. El incremento significativo en el tiempo de interacción sugiere que los usuarios utilizaron el sistema para profundizar en la información disponible más allá de la presentada originalmente por el museo virtual.

Sin embargo, no se observaron diferencias significativas en carga mental, eficiencia cognitiva ni desempeño en el cuestionario. Estos resultados indican que el uso del sistema RAG no incrementó el esfuerzo percibido por los participantes, pero tampoco produjo mejoras inmediatas en la comprensión medida mediante las preguntas del experimento.

El análisis de las consultas permite explicar parcialmente este comportamiento. Más de la mitad de las preguntas realizadas estuvieron orientadas a ampliar información, mientras que otra proporción importante buscó aclarar conceptos específicos presentes en las cédulas museográficas. Esto sugiere que los usuarios utilizaron el sistema principalmente como una herramienta de exploración y profundización, más que como un mecanismo para responder únicamente el cuestionario.

Implicaciones y limitaciones

En conjunto, los resultados muestran que la principal aportación de esta investigación radica en la estrategia de recuperación híbrida propuesta y en su integración dentro de un sistema RAG para información histórica y cultural. La evidencia experimental indica que mejorar la recuperación tiene un efecto directo sobre el desempeño global del sistema, reforzando la importancia de este componente dentro de las arquitecturas RAG.

No obstante, el trabajo presenta algunas limitaciones. Los experimentos se realizaron sobre colecciones específicas del dominio museográfico e histórico, por lo que los resultados podrían variar en otros dominios o con corpus de distinta naturaleza. Asimismo, la evaluación con usuarios se realizó con estudiantes universitarios, por lo que sería conveniente ampliar la muestra hacia visitantes reales de museos con perfiles demográficos más diversos.

Capítulo VII. Conclusiones

7.1 Conclusiones

El objetivo general de esta investigación fue evaluar la viabilidad de utilizar un modelo de lenguaje pequeño (SLM) integrado en un sistema de Retrieval-Augmented Generation (RAG) para generar respuestas contextualizadas sobre información museográfica preservando la fidelidad con respecto a los documentos de referencia.

Los resultados obtenidos permiten concluir que la propuesta es viable para el dominio considerado. La incorporación de un componente de recuperación híbrida permitió recuperar de manera consistente los documentos relevantes, proporcionando al modelo generativo el contexto necesario para producir respuestas fundamentadas en la información disponible.

Respecto al primer objetivo específico, la evaluación de diferentes estrategias de recuperación mostró que la combinación de evidencia léxica y evidencia semántica mediante una combinación lineal obtuvo un desempeño superior al de los métodos basados únicamente en BM25 o en *embeddings*. En particular, la configuración con un balance equivalente entre ambos componentes presentó el mejor comportamiento promedio, demostrando que la integración de ambas fuentes de evidencia favorece la recuperación de documentos relevantes.

En relación con el segundo objetivo específico, la evaluación de modelos generativos permitió seleccionar un modelo adecuado para su integración en el sistema RAG. Asimismo, se observó que el formato del *prompt* influye de manera importante en el desempeño del modelo, evidenciando que la calidad de las respuestas depende no solo del modelo utilizado, sino también de la forma en que se estructura la interacción.

Respecto al tercer objetivo específico, la evaluación del sistema RAG mostró que la implementación propuesta supera a una implementación equivalente basada en LangChain tanto en las métricas de recuperación como en la calidad de las respuestas generadas. Estos resultados indican que las decisiones adoptadas durante el diseño del recuperador tienen un efecto directo sobre el desempeño global del sistema.

Finalmente, la evaluación con usuarios mostró que el sistema RAG favorece la exploración de la información museográfica mediante consultas en lenguaje natural. Aunque no se observaron diferencias significativas en el desempeño de las tareas de comprensión, los participantes realizaron consultas orientadas principalmente a profundizar y ampliar la información presentada por el museo, lo que evidencia el potencial del sistema como herramienta de apoyo para la consulta de contenidos históricos y culturales.

Por tanto, esta investigación aporta el diseño y la evaluación experimental de una estrategia híbrida de recuperación basada en una combinación lineal ponderada entre evidencia léxica y evidencia semántica, así como su integración y validación dentro de un sistema RAG para

el dominio museográfico. Los resultados obtenidos muestran que la calidad de la recuperación constituye un componente determinante para el desempeño de sistemas RAG especializados en información histórica y cultural.

7.2 Productos y aportaciones

Durante el desarrollo de la presente investigación se obtuvieron diversos productos académicos y de divulgación científica que contribuyeron a la difusión del conocimiento en las áreas de inteligencia artificial, procesamiento del lenguaje natural y recuperación de información. Estos productos se describen a continuación.

7.2.1 Publicación en congreso nacional

Se presentó el artículo titulado: «Postprocesamiento Geométrico de Espacios de *Embeddings* para Clasificación de Textos con Voronoi»

en el XVIII Congreso Mexicano de Inteligencia Artificial (COMIA 2026), realizado del 25 al 29 de mayo de 2026 en Cuernavaca, Morelos, México.

Evidencia: Anexo B.

7.2.2 Conferencia de divulgación científica

Se impartió la conferencia: «La magia detrás de los LLM»

como ponente en las Jornadas Académicas de Innovación, Tecnología, Liderazgo y Sostenibilidad 2025, organizadas por el Instituto Tecnológico de Zacatepec.

Esta actividad permitió difundir conceptos fundamentales sobre modelos de lenguaje de gran escala y sus aplicaciones actuales.

Evidencia: Anexo C.

7.2.3 Taller especializado

Se impartió el taller: «Sistemas de Lenguaje Natural en Inteligencia Artificial»

durante el Primer Campamento de Verano de Inteligencia Artificial del Centro Nacional de Investigación y Desarrollo Tecnológico (CENIDET).

El taller abordó fundamentos y aplicaciones prácticas del procesamiento del lenguaje natural mediante modelos modernos de inteligencia artificial.

Evidencia: Anexo D.

7.2.4 Participación en proyectos de innovación tecnológica

Se presentó el proyecto SanEEG en la categoría Servicios para la Salud Humana dentro del certamen InnovaTec 2025, etapa local.

Esta participación fortaleció la experiencia en el desarrollo y evaluación de proyectos tecnológicos multidisciplinares.

Evidencia: Anexo E.

7.3 Trabajo futuro

Como líneas de trabajo futuro se plantean las siguientes:

- Evaluar la estrategia de recuperación propuesta en otros dominios de conocimiento, como archivos históricos, bibliotecas digitales o repositorios documentales, con el fin de analizar su capacidad de generalización.
- Incorporar métodos de reranking basados en modelos *cross-encoder* o técnicas de recuperación multietapa que permitan mejorar el ordenamiento de los documentos recuperados.
- Analizar el comportamiento de la estrategia propuesta utilizando modelos de *embeddings* de nueva generación y modelos generativos con razonamiento explícito, con el propósito de determinar si los beneficios observados se mantienen con arquitecturas más recientes.
- Ampliar la evaluación con usuarios incorporando visitantes reales de museos y una muestra demográfica más diversa, permitiendo estudiar el impacto del sistema en distintos perfiles de usuarios.
- Integrar información multimodal, como imágenes de las obras, mapas, líneas del tiempo o material audiovisual, para enriquecer el contexto proporcionado al sistema RAG y ampliar las posibilidades de interacción con el patrimonio cultural.
- Evaluar el sistema en escenarios de actualización continua del conocimiento, donde nuevas piezas o documentos museográficos puedan incorporarse al repositorio sin necesidad de volver a entrenar el modelo generativo.
- Construir un conjunto de referencia mediante la anotación manual de una submuestra de respuestas, con el fin de calibrar los jueces automáticos y evaluar con mayor precisión la presencia de alucinaciones, el respaldo contextual y la comprensibilidad. Asimismo, investigar estrategias automáticas para la detección y mitigación de alucinaciones.

Referencias

- [1] H. Valverde, «Los dispositivos digitales como herramientas de interpretación en entornos museales,» *Critical Hermeneutics*, vol. 4, núm. 1, pp. 119–136, 2020, doi: 10.13125/CH/4325.
- [2] P. Lewis, «Retrieval-augmented generation for knowledge-intensive NLP tasks,» *NIPS'20: Proceedings of the 34th International Conference on Neural Information Processing Systems*, pp. 9459-9474, 2020.
- [3] M. Klesel y H. F. Wittmann, «Retrieval-Augmented Generation (RAG),» *Business & Information Systems Engineering*, vol. 67, pp. 551–561, 2025, doi: 10.1007/s12599-025-00945-3.
- [4] A. Vaswani et al., «Attention is All you Need,» *Advances in Neural Information Processing Systems*, 2017.
- [5] D. Naik, «Large Data Begets Large Data: Studying Large Language Models (LLMs) and Its History, Types, Working, Benefits and Limitations,» *The International Conference on Computing, Communication, Cybersecurity & AI*, 2024.
- [6] C. D. Manning, *Introduction to information retrieval.*, Cambridge University Press, 2008.
- [7] G. Salton, «A vector space model for automatic indexing.,» *Communications of the ACM*, vol. 18, pp. 613-620, 1975.
- [8] S. Chang, «Engineering material failure analysis report generation based on QWen and Llama2,» *Results in Engineering*, 2025.
- [9] T. Mikolov, «Distributed representations of words and phrases and their compositionality.,» *Advances in Neural Information Processing Systems*, vol. 26, p. 3111–3119, 2013.
- [10] J. Devlin, «BERT: Pre-training of deep bidirectional transformers for language understanding.,» *Proceedings of NAACL-HLT*, 2019.
- [11] P. Turney, «From frequency to meaning: Vector space models of semantics.,» *Journal of Artificial Intelligence Research*, vol. 37, 2010.
- [12] M. Peters, «Deep contextualized word representations.,» *Proceedings of NAACL-HLT*, 2018.
- [13] Fabio Clarizia, Massimo De Santo, Rosario Gaeta y Rocco Loffredo, «Enhancement Large Language Models Domain Through Ontology-Based Retrieval-Augmented Generation,» *International Journal on Semantic Web and Information Systems*, vol. 21, núm. 1, pp. 1–29, 2025. [En línea]. Disponible en: <https://doi.org/10.4018/IJSWIS.392507>.
- [14] Jinhu Qi, Shuai Yan, Yibo Zhang, Wentao Zhang, Rong Jin, Yuwei Hu y Ke Wang, «RAG-Optimized Tibetan Tourism LLMs: Enhancing Accuracy and Personalization,» *arXiv:2408.12003* (aceptado en AIPR 2024), 2024. [En línea]. Disponible en: <https://doi.org/10.48550/arXiv.2408.12003>.
- [15] Jing Wan, Xinrong Li, Hao Zhang, Ao Zou y Rumei Wang, «An LLM-Based QA System for Chinese Painting and Calligraphy with Knowledge Graphs and External

- Documents,» npj Heritage Science, 2025. [En línea]. Disponible en: <https://doi.org/10.1038/s40494-025-02219-3>.
- [16] Yuqian Lin, Wenhui Zhang y Jiayi Lin, «Question Answering System Based on Retrieval-Augmented Generation,» Proc. ICCDC 2025, 2026. [En línea]. Disponible en: <https://doi.org/10.1145/3783669.3783745>.
- [17] Anmol Guragain, Jaime Bellver-Soler, Samuel Ramos-Varela, Long Lin, David Aragón Díaz y Luis Fernando D'Haro, «A Personalized, Multimodal AI Assistant for Enhancing Museum Visitor Experience,» Proc. Workshop on Awareness in Learning Agents (ALA), CEUR-WS, vol. 4136, 2025. [En línea]. Disponible en: <https://ceur-ws.org/Vol-4136/ala3.pdf>.
- [18] Yifan Hu, «MuseRAG++: A Deep Retrieval-Augmented Generation Framework for Semantic Interaction and Multi-Modal Reasoning in Virtual Museums,» Scientific Reports, 2026. [En línea]. Disponible en: <https://doi.org/10.1038/s41598-026-55700-9>.
- [19] Ritu Ram Ojha, «Vision-Augmented RAG System for Interactive Local Heritage Exploration,» TechRxiv (preprint), 2025. [En línea]. Disponible en: <https://doi.org/10.36227/techrxiv.175691179.92356541/v1>.
- [20] Hung-Shin Lee, Chen-Chi Chang, Ching-Yuan Chen y Yun-Hsiang Hsu, «Evaluating Cultural Knowledge Processing in Large Language Models: A Cognitive Benchmarking Framework Integrating Retrieval-Augmented Generation,» arXiv:2511.01649, 2025. [En línea]. Disponible en: <https://doi.org/10.48550/arXiv.2511.01649>.
- [21] Filip J. Kucia, Bartosz Grabek, Szymon D. Trochimiak y Anna Wróblewska, «How to Make Museums More Interactive? Case Study of Artistic Chatbot,» Proc. ACM CIKM 2025, pp. 6654–6658, 2025. [En línea]. Disponible en: <https://doi.org/10.48550/arXiv.2509.00572>.
- [22] Haofeng Wang, Yilin Guo, Zehao Li, Tong Yue, Yizong Wang, Enci Zhang, Rongqun Lin, Feng Gao, Shiqi Wang y Siwei Ma, «RiverEcho: Real-Time Interactive Digital System for Ancient Yellow River Culture,» IEEE ICME Workshop 2025, 2025. [En línea]. Disponible en: <https://doi.org/10.48550/arXiv.2506.21865>.
- [23] Yurong Zeng, Chengxi Yan, Jiayi Li, Jingxuan Yao y Zhihan Sun, «Is AIGC Technology Useful? A Comparative Analysis of Search Technologies for Online Museums,» Proc. Association for Information Science and Technology, vol. 62, pp. 1155–1160, 2025. [En línea]. Disponible en: <https://doi.org/10.1002/pra2.1362>.
- [24] Jing Zhou, Li Si y Wenjun Hou, «Humanities-in-the-Loop: Using Close Reading as a Method for Retrieval-Augmented Generation (RAG),» Proc. Association for Information Science and Technology, vol. 62, núm. 1, pp. 1185–1189, 2025. [En línea]. Disponible en: <https://doi.org/10.1002/pra2.1355>.
- [25] Iordanis Sapidis, Valantis Zervos, Michalis Mountantonakis y Yannis Tzitzikas, «Interactive and Provenance-Aware Search and QA over Documents Using LLMs, RAG and Knowledge Graph Verbalization,» New Trends in Theory and Practice of Digital Libraries, pp. 248–257, 2025. [En línea]. Disponible en: https://doi.org/10.1007/978-3-032-06136-2_24.

- [26] Paul Kelly, Jonathan Schild y Amir Jafari, «FolkRAG: A Retrieval-Augmented Generation System for Cultural Heritage Materials,» *Neural Computing and Applications*, vol. 37, núm. 24, pp. 20281–20297, 2025. [En línea]. Disponible en: <https://doi.org/10.1007/s00521-025-11455-4>.
- [27] Taemoun Jung e Inwhae Joe, «An Intelligent Docent System with a Small Large Language Model (sLLM) Based on Retrieval-Augmented Generation (RAG),» *Applied Sciences*, vol. 15, núm. 17, art. 9398, 2025. [En línea]. Disponible en: <https://doi.org/10.3390/app15179398>.
- [28] Jun Xu, Hao Zhang, Haijing Zhang, Jiawei Lu y Gang Xiao, «ChatTf: A Knowledge Graph-Enhanced Intelligent Q&A System for Mitigating Factuality Hallucinations in Traditional Folklore,» *IEEE Access*, vol. 12, pp. 162638–162650, 2024. [En línea]. Disponible en: <https://doi.org/10.1109/ACCESS.2024.3485877>.
- [29] Fabio Clarizia, Massimo De Santo, Rosario Gaeta, Roberta Gimigliano y Rocco Loffredo, «Reducing Hallucinations in LLMs: A Semantic Approach for Cultural Heritage,» *Proc. i-Cities 2025*, 2025. [En línea]. Disponible en: <https://icities25.unicas.it/papers/43.pdf>.
- [30] Sihan Song, Chuncheng Yang, Li Xu, Haibin Shang, Zhuo Li y Yinghui Chang, «TravelRAG: A Tourist Attraction Retrieval Framework Based on Multi-Layer Knowledge Graph,» *ISPRS International Journal of Geo-Information*, vol. 13, núm. 11, art. 414, 2024. [En línea]. Disponible en: <https://doi.org/10.3390/ijgi13110414>.
- [31] Mohamed Ali, Mohamed Mabrouk y Zaki Taha, «Hallucination Mitigation Techniques in Large Language Models,» *International Journal of Intelligent Computing and Information Sciences*, vol. 24, núm. 4, pp. 73–81, 2024. [En línea]. Disponible en: <https://doi.org/10.21608/ijicis.2024.336135.1365>.
- [32] Yuxia Wang, Minghan Wang, Muhammad Arslan Manzoor, Fei Liu, Georgi Georgiev, Rocktim Jyoti Das y Preslav Nakov, «Factuality of Large Language Models: A Survey,» *arXiv:2402.02420*, 2024. [En línea]. Disponible en: <https://doi.org/10.48550/arXiv.2402.02420>.
- [33] Duy Khoa Pham y Bao Quoc Vo, «Towards Reliable Medical Question Answering: Techniques and Challenges in Mitigating Hallucinations in Language Models,» *arXiv:2408.13808*, 2024. [En línea]. Disponible en: <https://doi.org/10.48550/arXiv.2408.13808>.
- [34] Minda Hu, Bowei He, Yufei Wang, Liangyou Li, Chen Ma e Irwin King, «Mitigating Large Language Model Hallucination with Faithful Finetuning,» *arXiv:2406.11267*, 2024. [En línea]. Disponible en: <https://doi.org/10.48550/arXiv.2406.11267>.
- [35] Peiran Wang, Yang Liu, Yunfei Lu, Jue Hong y Ye Wu, «What Are Models Thinking About? Understanding Large Language Model Hallucinations ‘Psychology’ Through Model Inner State Analysis,» *arXiv:2502.13490*, 2025. [En línea]. Disponible en: <https://doi.org/10.48550/arXiv.2502.13490>.
- [36] Cunxiang Wang, Xiaoze Liu, Yuanhao Yue, Xiangru Tang, Tianhang Zhang et al., «Survey on Factuality in Large Language Models: Knowledge, Retrieval and Domain-Specificity,» *arXiv:2310.07521*, 2023. [En línea]. Disponible en: <https://doi.org/10.48550/arXiv.2310.07521>.

- [37] Jason Kirchenbauer y Caleb Barns, «Hallucination Reduction in Large Language Models with Retrieval-Augmented Generation Using Wikipedia Knowledge,» OSF Preprints, 2024. [En línea]. Disponible en: <https://doi.org/10.31219/osf.io/pv7r5>.
- [38] Xinxin Liu, «A Survey of Hallucination Problems Based on Large Language Models,» Applied and Computational Engineering, vol. 97, pp. 24–30, 2024. [En línea]. Disponible en: <https://doi.org/10.54254/2755-2721/97/2024.17851>.
- [39] Y. Gao, «Retrieval-augmented generation for large language models: A survey,» 2023.
- [40] A. Mallen, «When not to trust language models: Investigating effectiveness of parametric and non-parametric memories,» Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, 2023.
- [41] K. Shuster, «Retrieval augmentation reduces hallucination in conversation,» Findings of the Association for Computational Linguistics: EMNLP, 2021.
- [42] W. Fan, «A survey on RAG meeting LLMs: Towards retrieval-augmented large language models,» Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2024.
- [43] V. Karpukhin, «Dense passage retrieval for open-domain question answering,» Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020.
- [44] F. Petroni, «Language models as knowledge bases?,» Conference on Empirical Methods in Natural Language Processing, 2019.
- [45] Z. Ji, «Survey of hallucination in natural language generation,» ACM Computing Surveys, vol. 55, 2023.
- [46] J. Maynez, «On faithfulness and factuality in abstractive summarization,» Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020.
- [47] W. Zhang, «Hallucination Mitigation for Retrieval-Augmented Large Language Models: A Review,» Mathematics, vol. 13, n° 5, 2025.
- [48] J. Omiye, «Large language models propagate race-based medicine,» Digital Medicine, vol. 6, 2023.
- [49] L. Ouyang, «Training language models to follow instructions with human feedback,» Advances in Neural Information Processing Systems, vol. 35, 2022.
- [50] J. Wei, «Chain-of-thought prompting elicits reasoning in large language models,» Advances in Neural Information Processing Systems, vol. 35, 2022.
- [51] Z. Guo, M. Schlichtkrull y A. Vlachos, «A survey on automated fact-checking,» Transactions of the Association for Computational Linguistics, vol. 10, pp. 178–206, 2022.
- [52] G. Izacard, «Few-shot learning with retrieval augmented language models,» Journal of Machine Learning Research, vol. 23, 2022.
- [53] K. Sparck Jones, «A statistical interpretation of term specificity and its application in retrieval,» Journal of Documentation, vol. 28, 1972.
- [54] S. Robertson, «The probabilistic relevance framework: BM25 and beyond,» Foundations and Trends in Information Retrieval, vol. 3, 2009.

- [55] G. Furnas, «The vocabulary problem in human-system communication,» *Communications of the ACM*, vol. 30, 1987.
- [56] T. Kwiatkowski, «Natural questions: A benchmark for question answering research,» *Transactions of the Association for Computational Linguistics*, vol. 7, 2019.
- [57] N. Thakur, «BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models,» *Advances in Neural Information Processing Systems*, vol. 34, 2021.
- [58] Y. Luan, «Sparse, dense, and attentional representations for text retrieval,» *Transactions of the Association for Computational Linguistics*, vol. 9, 2021.
- [59] G. Cormack, «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*, 2009.
- [60] X. Ma, «PROP: Pre-training with representative words prediction for ad-hoc retrieval,» *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, 2021.
- [61] D. Jurafsky, *Speech and language processing*, 2024.
- [62] H. Yu et al., «Evaluation of Retrieval-Augmented Generation: A Survey,» *arXiv:2405.07437*, 2024, doi: 10.48550/arXiv.2405.07437.
- [63] J. Saad-Falcon et al., «ARES: An Automated Evaluation Framework for Retrieval-Augmented Generation Systems,» in *Proc. NAACL*, 2024.
- [64] M. A. Rehman, «ragR: Retrieval-Augmented Generation and RAG Assessment in R,» *arXiv:2604.23515*, 2026.
- [65] M. Belyi, «Luna: A Lightweight Evaluation Model to Catch Language Model Hallucinations with High Accuracy and Low Cost,» *Proceedings of the 31st International Conference on Computational Linguistics: Industry Track*, 2025.

Anexos

Anexo A. *Prompt* utilizado en la evaluación automática

El siguiente *prompt* fue utilizado para evaluar automáticamente las respuestas generadas por el sistema RAG. Cada evaluador recibió la misma instrucción y el mismo conjunto de datos, con el objetivo de garantizar que las diferencias observadas correspondieran al comportamiento del modelo evaluador y no a variaciones en la instrucción.

You are evaluating the output of a Retrieval-Augmented Generation (RAG) system.

You will receive multiple rows from a dataset.

Each row contains:

- row_id
- question
- generated_answer
- ctx1-ctx10 (retrieved contexts)
- demographic profile (age, gender, education_level)
- reference_document (ground truth)

Your task is to evaluate EACH ROW independently.

EVALUATION CRITERIA

1. Ground Truth in Top10 (GT_in_top10)

Check if the reference_document appears in ctx1-ctx10.

Output:

1 = yes

0 = no

2. Answer correctness (answer_label)

Evaluate ONLY using the contexts:

1 = Correct -> grounded, correct, complete, understandable, or partially grounded

0 = Incorrect -> hallucination, unsupported, incorrect refusal, or incomplete

3. Context grounding (context_supported)

Is the answer supported by the retrieved contexts?

1 = yes

0 = no

4. Demographic comprehension (demographic_ok)

Would a person with the given age and education_level understand the answer?

1 = yes

0 = no

5. Refusal detection (refusal)

Does the answer explicitly state lack of evidence?

1 = yes

0 = no

6. Refusal type (refusal_type)

If refusal = 1:

3 = correct_refusal -> contexts truly lack information

2 = incorrect_refusal -> contexts contain the information

If refusal = 0:

1 = none

7. Hallucination (hallucination)

Answer contains information NOT present in contexts.

1 = yes

0 = no

IMPORTANT RULES

- Use ONLY the provided contexts

- Do NOT use external knowledge
- Evaluate STRICTLY per row
- Be consistent across all rows
- Do NOT skip any row

OUTPUT FORMAT (STRICT JSON)

Return a JSON array:

```
[  
{  
  "row_id": <id>,  
  "gt_in_top10": 0 or 1,  
  "answer_label": 0 or 1,  
  "context_supported": 0 or 1,  
  "demographic_ok": 0 or 1,  
  "refusal": 0 or 1,  
  "refusal_type": 1 | 2 | 3,  
  "hallucination": 0 or 1  
}  
]
```

DO NOT return explanations.

DO NOT return aggregated metrics.

DO NOT choose only one number and set it for all answers; think about each row.

ONLY return the JSON array.

Anexo B. Constancia de presentación del artículo Postprocesamiento Geométrico de Espacios de *Embeddings* para Clasificación de Textos con Voronoi en COMIA 2026.



Figura B.1. Anexo B. Constancia de presentación de artículo

Anexo C. Reconocimiento como ponente de la conferencia *La magia detrás de los LLM*.



Figura C.1. Anexo C. Reconocimiento como ponente

Anexo D. Reconocimiento por impartición del taller Sistemas de Lenguaje Natural en Inteligencia Artificial.



Figura D.1. Anexo D. Reconocimiento por impartir un taller

Anexo E. Constancia de participación en el certamen InnovaTec 2025 con el proyecto SanEEG.



Figura E.1. Anexo E. Constancia de participación en el certamen InnovaTec 2025