# Generación Aumentada por Recuperación sobre Películas

Autor: Daniel Polanco Pérez

Tutor: Dr.C. Yudivián Almeida Cruz

Tesis en opción al grado de Licenciado en Ciencia de la Computación


Línea de Investigación: Generación Aumentada por Recuperación
Grupo de Investigación: Inteligencia Artificial

Universidad de La Habana
Facultad de Matemática y Computación
La Habana, Cuba
Junio, 2025

# Resumen

Esta investigación desarrolla un sistema inteligente capaz de analizar películas y responder preguntas complejas sobre su contenido, integrando todos los elementos que las componen: diálogos, sonidos ambientales, música y secuencias visuales. El sistema transforma esta información audiovisual en descripciones textuales detalladas que capturan no solo lo que se dice, sino también lo que ocurre en pantalla y fuera de ella. Estas descripciones se organizan de manera que permiten encontrar rápidamente información relevante cuando el usuario realiza una consulta, ya sea sobre personajes, escenas específicas o elementos narrativos sutiles.

El principal aporte de este trabajo es mostrar cómo la inteligencia artificial puede comprender y procesar el lenguaje cinematográfico en toda su complejidad, superando limitaciones de sistemas anteriores que solo consideraban los diálogos o metadatos básicos. Los resultados muestran que este enfoque permite responder preguntas más sofisticadas sobre las películas. Este avance acerca la posibilidad de interactuar con el contenido audiovisual de forma tan natural como lo hacemos hoy con la información textual.

**Palabras clave:** Generación Aumentada por Recuperación (RAG), multimodalidad, procesamiento audiovisual, modelos de lenguaje, inteligencia artificial, películas.

# Abstract

This research develops an intelligent system capable of analyzing films and answering complex questions about their content, integrating all the elements that compose them: dialogue, ambient sounds, music, and visual sequences. The system transforms this audiovisual information into detailed textual descriptions that capture not only what is said, but also what happens on and off screen. These descriptions are organized in such a way that they allow users to quickly find relevant information when they make a query, whether about characters, specific scenes, or subtle narrative elements.

The main contribution of this work is to show how artificial intelligence can understand and process cinematic language in all its complexity, overcoming the limitations of previous systems that only considered dialogue or basic metadata. The results show that this approach allows for more sophisticated questions about films to be answered. This advance brings us closer to the possibility of interacting with audiovisual content as naturally as we do today with textual information.

**Keywords:** Retrieval-Augmented Generation (RAG), multimodality, audiovisual processing, language models, artificial intelligence, movies.

# Agradecimientos

A mi hermano por ser mi guía desde mi primer día de vida hasta al escribir estas palabras, por siempre estar a mi lado y hacer más fácil el camino.

A mis padres por sus consejos certeros, por apoyarme siempre en los momentos más difíciles y por encarnar el molde de la persona que deseo ser.

A mi abuela Onelia y mi abuela Leo por su amor incondicional y por acompañarme cada día espiritualmente.

A mi abuelo Nelson por su esfuerzo y dedicación a la familia y a mi abuelo Elpidio por su humor heredado.

A mis tíos Yamil, Carmen y Toni y a mis primos hermanos Ana y Samuel por su preocupación constante y por tantos momentos de alegría.

A mi familia toda, por esos domingos en el piso 4, por las visitas a Uña y por siempre tenerme presente.

A mi novia Nathaly por siempre creer en mí, por hacerme soñar en grande, por ser mi confidente y mi complemento.

A la familia de mi pareja, en especial a los abuelos Ario, Bibiana y Mercedes; a las tías Leticia y Anelis y a mis suegros Mercy y Abel, por adoptarme como uno más en su familia, por el cariño que siempre me han dado y por consentirme tanto.

A Ernesto, mi hermano de otra madre, por estar en cada etapa de mi vida y por tantas historias y anécdotas juntos.

A Olivia, Miguel, Massiel, Adrián y en especial a Nanda por ser mis compañeros de batalla todos estos años y por tantas risas y aventuras.

A Venezuela por los tantos amigos que a la distancia siguen presentes y por marcar de vinotinto mi alma para siempre.

A todos los profesores que me formaron en toda mi etapa estudiantil y en especial al profe Somoza, a las profes Celia e Idania y a mi tutor Yudivián, por el apoyo y dedicación brindados en cada momento.

A cada persona que un día estuvo a mi lado y dejó una huella en mí.

# Opinión del Tutor

**Título de la tesis:** Generación Aumentada por Recuperación sobre Películas
**Estudiante:** Daniel Polanco Pérez
**Tutor:** Dr. Yudivián Almeida Cruz

El estudiante Daniel Polanco Pérez desarrolló satisfactoriamente el trabajo de diploma titulado "Generación Aumentada por Recuperación sobre Película". En este trabajo el estudiante propuso el diseño de un sistema que permite analizar películas y responder preguntas complejas sobre su contenido, integrando todos los elementos que las componen: diálogos, sonidos ambientales, música y secuencias visuales.

En el sistema que Daniel propone los distintos elementos del video se transforman en descripciones textuales que capturan no solo lo que se dice, sino también lo que ocurre en pantalla. Estas descripciones se organizan de manera que se puede recuperar información relevante cuando el usuario realiza una consulta, ya sea sobre personajes, escenas específicas u otros elementos narrativos. Para ello se basa en la potencia de los *embeddings*, los grandes modelos de lenguaje (LLM) y la generación aumentada por recuperación (RAG).

Para poder afrontar el trabajo, el estudiante tuvo que revisar literatura científica relacionada con la temática así como soluciones existentes y bibliotecas de software que pueden ser apropiadas para su utilización. Todo ello con sentido crítico, determinando las mejores aproximaciones y también las dificultades que presentan.

Todo el trabajo fue realizado por el estudiante con una buena dosis de constancia, capacidad de trabajo individual y habilidades, tanto de gestión,

como de desarrollo y de investigación.

Por estas razones pedimos que le sea otorgada al estudiante Daniel Polanco Pérez una calificación que le permita obtener el título de Licenciado en Ciencia de la Computación y así, por derecho propio, se pueda integrar al gremio de los profesionales de la computación.

Dr. Yudivián Almeida Cruz

# Índice

# Introducción

Desde el surgimiento del cine a finales del siglo XIX hasta la consolidación de las plataformas de *streaming* en el siglo XXI, el contenido audiovisual se ha convertido en una de las formas de comunicación y expresión cultural más influyentes. Los videos, en sus múltiples formatos —desde películas, series y documentales hasta videoclips, tutoriales, videos de videojuegos y contenidos de redes sociales—, han evolucionado de ser simples narrativas lineales a complejas obras artísticas multimodales que integran narrativa, imagen, sonido, música, efectos visuales y tecnología digital.

Hoy día, el consumo de contenido audiovisual es parte fundamental de la vida cotidiana de millones de personas. Plataformas como Netflix, Disney+, Amazon Prime y YouTube han contribuido a una verdadera explosión de contenidos, haciendo que la cantidad de películas, series y videos disponibles sea abrumadora. Este fenómeno ha generado no sólo un nuevo paradigma de distribución, sino también una necesidad creciente de sistemas inteligentes que ayuden a explorar, organizar y comprender este contenido [5].

En paralelo, la forma en que las personas acceden a la información también ha cambiado radicalmente. La búsqueda tradicional basada en palabras clave ha sido progresivamente reemplazada por interacciones conversacionales con asistentes virtuales y *chatbots* [39]. De los que se espera que entiendan contextos complejos y proporcionen respuestas elaboradas, incluso sobre temas subjetivos o interpretativos.

Esto ha generado una demanda creciente por sistemas de inteligencia artificial capaces de comprender contenido audiovisual y responder preguntas específicas sobre él. Por ejemplo: "¿Qué personajes aparecen en la escena del crimen?" o "¿Cuándo se menciona por primera vez el conflicto familiar del protagonista?". Este tipo de consultas requiere una comprensión profunda

y contextual del material audiovisual, algo que excede las capacidades de los buscadores tradicionales o de los sistemas actuales basados en metadatos simples [11].

El desarrollo de modelos de lenguaje de gran escala (LLMs) como GPT-3, GPT-4 (OpenAI), LLaMA (Meta), Claude (Anthropic) o Gemini (Google DeepMind), ha sido una de las revoluciones más notorias en la inteligencia artificial contemporánea. Estos modelos han demostrado una capacidad sin precedentes para generar lenguaje natural, responder preguntas, redactar textos creativos, resumir documentos, y entablar diálogos fluidos. Sin embargo, su conocimiento está limitado por su entrenamiento previo [10], y no siempre puede responder de manera precisa sobre contenidos que no han sido incluidos explícitamente en sus datos de entrenamiento [30].

Frente a este límite, surge la arquitectura *Retrieval-Augmented Generation* (RAG) como un método que combina modelos preentrenados con mecanismos externos de búsqueda para mejorar la generación de texto [33]. Esto funciona como un enfoque que permite a los LLMs complementar sus respuestas con información extraída dinámicamente de fuentes externas. Esta arquitectura marca un punto de inflexión en la relación entre búsqueda y generación, ofreciendo una solución al dilema de "memoria estática" de los modelos. Sin embargo, la implementación de RAG ha sido en su mayoría textual, dejando de lado dominios ricamente semánticos como el audiovisual.

Los estudios más representativos sobre RAG han abordado diversos enfoques para mejorar su arquitectura y eficiencia. Lewis et al. [33] introdujeron la arquitectura original, utilizando *Dense Passage Retrieval* (DPR) como mecanismo de recuperación. Por su parte, Izacard y Grave [22] exploraron cómo los modelos generativos pueden beneficiarse de pasajes recuperados en entornos de preguntas abiertas. Un avance significativo fue propuesto por Izacard et al. [23], quienes presentaron un modelo RAG preentrenado capaz de aprender con muy pocos ejemplos. Más recientemente, Shi et al. [49] desarrollaron una arquitectura modular que optimiza el proceso de recuperación y mejora el rendimiento mediante una gestión estratégica del conocimiento. Adicionalmente, An et al. [3] introdujeron un enfoque innovador que mejora la relación entre calidad y eficiencia computacional en RAG mediante la reutilización del caché del reranking. Estos trabajos se han centrado principalmente en aplicaciones como pregunta-respuesta, generación de resúmenes y diálogo conversacional basado en documentos.

Posteriores mejoras, como REALM [20] o RETRO [6], han perfeccionado la eficiencia y precisión de los sistemas RAG, pero sin salir del dominio estrictamente textual. Esta falta de atención al contenido multimodal revela una brecha entre el potencial técnico de los modelos y las necesidades reales de los usuarios que interactúan con medios complejos como las películas.

A su vez, la investigación sobre la comprensión automática del contenido audiovisual ha estado avanzando por caminos parcialmente independientes. Por un lado, se han desarrollado modelos de visión por computadora para la descripción de escenas visuales (*Dense Captioning* [42], *Visual Grounding* [60], *Image-to-Text* [18]). Por otro lado, se han entrenado modelos para la transcripción de audio a texto (ASR - *Automatic Speech Recognition* [1]), y para la detección de sonidos no verbales (*sound event detection* [47]). Sin embargo, estos modelos no han sido aún integrados de manera sistemática dentro de un sistema RAG que permita su explotación estructurada para la consulta semántica de películas.

En investigaciones recientes, se ha observado un avance significativo en el desarrollo de modelos multimodales, capaces de procesar e interpretar imágenes, video y sonido. Uno de los primeros referentes en esta línea es Flamingo [2], el cual combina modelos de lenguaje preentrenados con componentes visuales, permitiendo abordar tareas de visión y lenguaje de manera conjunta. Por su parte, GPT-4V, lanzado por OpenAI en 2023, extiende las capacidades del modelo GPT-4 al incorporar procesamiento visual, lo que le permite analizar elementos gráficos diversos y generar respuestas que integran información textual y visual de manera coherente. Finalmente, Gemini, desarrollado por Google DeepMind entre 2023 y 2024, representa una arquitectura multimodal nativa, diseñada para integrar y razonar simultáneamente sobre texto, imágenes y audio.

Sin embargo, el foco de estas investigaciones ha estado más en el desarrollo de capacidades perceptivas que en la estructuración del conocimiento recuperable. Es decir, aunque estos modelos "entienden" imágenes o escenas, no han sido optimizados para permitir que el usuario acceda o consulte ese conocimiento mediante sistemas de recuperación orientados a preguntas específicas, y mucho menos bajo esquemas RAG.

De forma complementaria, herramientas como Whisper [43] para transcripción de audio, CLIP [44] para asociación imagen-texto, o los modelos como BLIP-2 [34], han demostrado que los cimientos tecnológicos para trans-

formar contenido audiovisual en texto ya existen. No obstante, hasta la fecha, la investigación en RAG se ha centrado principalmente en dominios textuales, y no se conoce un sistema ampliamente adoptado que integre de manera sistemática y escalable todos los componentes de una película (diálogo, imágenes, música, sonidos, narrativa) en un marco unificado de RAG [25].

A pesar del progreso alcanzado en el desarrollo de modelos de lenguaje y tecnologías multimodales, persiste una disociación estructural entre la riqueza semántica de los contenidos audiovisuales y la capacidad de los sistemas de inteligencia artificial para acceder a ellos de forma consultable y precisa. Las películas, por su naturaleza narrativa, visual y sonora, constituyen un tipo de dato complejo en el que la información relevante no se encuentra únicamente en los diálogos, sino también en la composición visual, los gestos, los sonidos ambientales, la música, los silencios y su progresión temporal.

Actualmente, las consultas realizadas a modelos de lenguaje sobre películas dependen en gran medida de bases de datos manuales, descripciones incompletas o información proveniente de fuentes no estructuradas. Esto limita la profundidad y exactitud de las respuestas, especialmente cuando se requieren detalles de escenas específicas, relaciones implícitas entre personajes o análisis narrativos que no están textualmente disponibles. En otras palabras, el conocimiento que reside en el lenguaje audiovisual no es directamente accesible mediante los mecanismos actuales de recuperación semántica [25].

Para abordar estas limitaciones, se propone la integración de diversas herramientas que permitan mejorar significativamente la capacidad de los sistemas de inteligencia artificial para comprender y responder preguntas semánticas complejas sobre contenido audiovisual. Para la consulta del contenido, se presenta la implementación de un sistema RAG que componga las ideas expuestas como son los modelos de visión por computadora, el reconocimiento automático del habla y la detección de eventos sonoros. Con este fin, y considerando las restricciones técnicas en cuanto a potencia computacional y disponibilidad de tiempo, el enfoque adoptado se basa en el desarrollo de un caso de estudio. Esta metodología brinda la oportunidad de estudiar a profundidad una parte de cierto problema con un tiempo que generalmente es limitado.

Por tanto, el objetivo de esta investigación es diseñar e implementar un software que permita realizar consultas en lenguaje natural sobre una base

de datos de películas y obtener resultados relacionados con sus apartados audiovisuales. Los objetivos específicos planteados para dar cumplimiento al objetivo general son:

1. Realizar el estudio del marco teórico de la recuperación de información multimedia, específicamente en películas, indagando en sistemas RAG.
2. Modelar un sistema para transformar toda la información audiovisual en texto y adaptarla a un sistema RAG.
3. Comprobar el funcionamiento correcto del software previamente modelado para permitir escribir consultas y recibir respuestas con una interfaz funcional.
4. Analizar los resultados del programa y contrastarlos con modelaciones similares en la literatura.

Además de esta introducción el documento de tesis consta de 3 capítulos, conclusiones, recomendaciones y referencias bibliográficas.

El capítulo 1, Estado del arte, define el marco teórico que sustenta esta investigación, se abordan los principios fundamentales de los sistemas de Generación Aumentada por Recuperación (RAG), así como las principales tecnologías para el procesamiento y extracción de información audiovisual. Además, se analizan los avances recientes en modelos de lenguaje multimodal y su aplicabilidad en tareas de consulta sobre contenido no textual.

El capítulo 2, Diseño de la propuesta, presenta el corazón técnico de esta investigación: un marco innovador que transforma películas en datos estructurados para sistemas RAG. Aquí se articulan metodologías clave para procesar diálogos, sonidos y elementos visuales, convertirlos en representaciones semánticas y diseñar un motor de búsqueda inteligente. El sistema no solo recupera información relevante, sino que genera respuestas verificables, abordando desde el procesamiento inicial hasta los desafíos de calidad en la generación final. Esta arquitectura integrada sienta las bases para un nuevo enfoque en el análisis de contenido audiovisual mediante IA.

Finalmente, en el capítulo 3, Detalles de Implementación y Experimentos, describe la implementación técnica de la solución propuesta en el capítulo anterior y se realiza un análisis del sistema desarrollado mediante escenarios de

prueba diseñados para medir la calidad y precisión de las respuestas generadas. Los resultados obtenidos se analizan y se contrastan con otros enfoques documentados en la literatura.

La tesis finaliza exponiendo las Conclusiones generales de la investigación, se destacan sus aportes científicos y técnicos, y se formulan las Recomendaciones que orienten futuras líneas de trabajo en el área. El documento finaliza con la Bibliografía.

# Capítulo 1

# Estado del Arte

En este capítulo se presentan los fundamentos técnicos y conceptuales necesarios para contextualizar el trabajo propuesto. Se inicia con un recorrido por los grandes modelos de lenguaje (LLM) y su progresiva transformación hacia modelos multimodales, subrayando cómo los avances recientes en inteligencia artificial han posibilitado que estos sistemas no solo procesen texto, sino también imágenes, audio y vídeo. Dicha evolución hacia la multimodalidad ha sido impulsada por la incorporación de arquitecturas como los Transformers y el desarrollo de técnicas avanzadas de aprendizaje profundo multimodal, que capacitan a los modelos para identificar y manejar relaciones complejas entre distintos tipos de datos, permitiendo abordar tareas novedosas sin necesidad de ajustes específicos para cada modalidad.

A continuación, se expone en detalle el paradigma "Generación Aumentada por Recuperación" (RAG), explicando su funcionamiento basado en la combinación de recuperación de información relevante y generación de respuestas contextualizadas mediante grandes modelos de lenguaje. Se describen los componentes principales de los sistemas RAG, como los modelos de recuperación y de generación, y se analizan los avances recientes, incluyendo nuevas arquitecturas que optimizan la integración y el procesamiento eficiente de grandes volúmenes de datos heterogéneos.

Finalmente, el capítulo concluye con un análisis del estado del arte en la aplicación de RAG al contenido multimodal y cinematográfico. Se examinan los logros alcanzados en la comprensión y generación de información a partir de fuentes diversas —texto, imágenes, audio y vídeo—, así como la capacidad

de los sistemas RAG multimodales para interpretar y relacionar estos datos en contextos complejos, como la indexación, resumen y análisis de obras audiovisuales.

## 1.1. Grandes Modelos de Lenguaje

Los LLM son sistemas de inteligencia artificial basados en arquitecturas de aprendizaje profundo, principalmente transformadores, que cuentan con miles de millones de parámetros entrenados con enormes volúmenes de datos textuales para reconocer patrones complejos en el lenguaje [40]. Estos modelos son capaces de procesar, generar y predecir texto en lenguaje humano, realizando múltiples tareas de procesamiento de lenguaje natural sin necesidad de entrenamiento específico para cada una [12]. Sin embargo, aunque exhiben un conocimiento general considerable y pueden manejar diversas aplicaciones, su "comprensión" es estadística y no semántica, y su desempeño depende en gran medida de la calidad y cantidad de datos, así como de técnicas adicionales como el ajuste fino y la retroalimentación humana para mejorar su precisión y contextualización. [36]

En los últimos años, los LLM han transformado profundamente el campo del procesamiento del lenguaje natural. Más recientemente, los modelos fundacionales han ampliado su alcance incorporando capacidades visuales, dando lugar a los grandes modelos de visión-lenguaje (LVLM). Estos se han establecido como la referencia principal para abordar una amplia variedad de tareas debido a sus avanzadas habilidades multimodales. Los LVLM pueden combinar lenguaje e imagen, así como describir escenas visuales, responder preguntas basadas en contenido visual o generar imágenes a partir de descripciones textuales. En conjunto, los LLM y LVLM representan un avance significativo hacia sistemas de inteligencia artificial más versátiles y capaces de interactuar de manera más natural y efectiva con el mundo humano [62].

La arquitectura Transformer, presentada en el artículo "Attention Is All You Need"[56], constituye el fundamento de los LLM. Se trata de una red de aprendizaje profundo para procesar y resolver tareas de lenguaje natural. Estos modelos, entrenados con extensos corpus textuales y multimodales, codifican grandes cantidades de conocimiento dentro de sus parámetros a gran escala.

Los LLM, como ChatGPT [40], Gemini [51] o Deepseek [13], han demostrado una gran capacidad para sostener conversaciones textuales que reflejan de manera precisa y natural el lenguaje humano, gracias a su habilidad para interpretar el contexto, generar respuestas coherentes y adaptarse dinámicamente al estilo comunicativo. Para mejorar la representación semántica del texto, se introdujeron las "Representaciones de Codificador Bidireccional a partir de Transformadores" (BERT) [14], las cuales se caracterizan por su capacidad para entender el contexto de las palabras en una oración, considerando simultáneamente tanto las palabras que preceden como las que siguen a cada término, logrando así una representación contextual más rica y matizada que impulsa la precisión en diversas tareas de procesamiento del lenguaje natural. [56]

## 1.2. RAG (Retrieval-Augmented Generation)

Los LLM aún son propensos a generar resultados factualmente incorrectos, ya que su conocimiento paramétrico puede ser inexacto o estar desactualizado [10]. Esta limitación resalta la necesidad de incorporar conocimiento de fuentes externas, y la "Generación Aumentada por Recuperación" (RAG) se perfila como un mitigador esencial. La RAG es una estrategia que combina los procesos de recuperación y generación para producir respuestas precisas, basándose en conocimiento externo relevante. Lewis et al. introdujeron por primera vez dicho concepto en su artículo seminal [33].

**Definición 1 (Generación Aumentada por Recuperación)** *RAG es un enfoque general de ajuste fino mediante el cual dotamos a los modelos generativos preentrenados con memoria paramétrica, de una memoria no paramétrica, mejorando así su capacidad para generar respuestas basadas en información externa.*

El sistema RAG se estructura en tres componentes principales: recuperación del contexto, construcción del *prompt* y generación de la respuesta.

En la etapa de recuperación, se construye un índice de búsqueda sobre una colección de documentos candidatos y se aplica una técnica adecuada para recuperar texto relevante [68]. Existen dos enfoques principales de recuperación: la recuperación basada en el léxico usada por Robertson y Zaragoza

[45], Robertson et al. [46], Jones [29] y Jeong et al. [24], que utiliza representaciones vectoriales dispersas; y los métodos de recuperación semántica, que utilizan representaciones vectoriales densas son empleadas por Zhao et al. [67], Karpukhin et al. [31], Jiang et al. [27], Wang et al. [58]. El primero "tokeniza" los documentos y crea un índice invertido basado en un vocabulario, para posteriormente recuperar los documentos relevantes mediante la correspondencia léxica. El segundo asigna los documentos a vectores densos de baja dimensión y, posteriormente, construye un índice eficiente de vectores de documentos mediante algoritmos de búsqueda por vecino más cercano aproximado, clasificando los documentos candidatos según la similitud de las incrustaciones. Ambos métodos suelen ser eficaces para la recopilación de documentos a gran escala, ampliamente utilizados en los sistemas RAG existentes.

En la fase de construcción del *prompt*, los documentos recuperados se integran en la entrada del LLM junto con la descripción de la tarea. Dado que dichos documentos pueden ser extensos, su simple concatenación puede resultar en una mala utilización del contexto, debido a sesgos de atención (por ejemplo, la pérdida de información en secciones intermedias del texto) [37]. Para mitigar este problema, los enfoques actuales suelen aplicar modelos de reordenamiento que priorizan los documentos más relevantes como hacen Wang et al. en REAR [59]. Alternativamente, se pueden emplear técnicas de extracción o compresión de información para conservar únicamente el contenido más pertinente, reduciendo así la longitud total del contexto como proponen Xu et al. en RECOMP[61].

En la etapa de generación de respuesta, el *prompt* enriquecido se introduce al LLM, permitiéndole aprovechar la información recuperada para completar con mayor precisión la tarea. No obstante, los documentos pueden incluir datos irrelevantes o incluso contradictorios. Para paliar este riesgo, se puede instruir al modelo a realizar una autoverificación de la calidad de la respuesta y decidir si es necesario repetir la recuperación [48], o aplicar un mecanismo de evaluación de confianza que determine si se requiere realmente la recuperación para la tarea actual [28].

RAG reduce eficazmente las alucinaciones de los modelos y permite el acceso a información específica del dominio sin necesidad de un costoso reentrenamiento del modelo [33]. Los avances recientes en RAG han seguido dos trayectorias metodológicas distintas: los enfoques basados en fragmentos se

han centrado en optimizar la segmentación y la recuperación de texto mediante *embeddings* avanzados [7], mientras que los métodos basados en grafos han explorado el uso de representaciones de conocimiento estructurado para mejorar la eficiencia y precisión del proceso de recuperación [15].

## 1.3. RAG sobre contenido multimodal

RAG ha mostrado gran eficacia en tareas basadas en texto, pero su expansión hacia entornos multimodales ha abierto nuevas perspectivas y desafíos. El auge de los modelos multimodales ha permitido a los sistemas RAG acceder a nuevas fuentes de conocimiento, incluyendo imágenes como MuRAG [9], código [19] y, más recientemente, audio [65]. Sin embargo, los videos, como modalidad multimodal por excelencia, han sido relativamente poco explorados pese a contener una rica combinación de señales visuales, temporales y auditivas como presentan Lee et al. [32] o Faysee et al. [17].

Estudios recientes han reconocido esta carencia e intentan subsanarla. Por ejemplo, el marco VideoRAG [26] ha sido uno de los primeros en integrar la recuperación y generación de conocimiento directamente sobre videos, utilizando LVLM que permiten una fusión más holística de señales textuales y visuales. Este enfoque supera las limitaciones de los sistemas tradicionales de RAG textual, que no logran representar con precisión elementos como dinámicas visuales, interacciones espaciales o cambios temporales [63].

En este contexto, también se han propuesto soluciones alternativas que convierten el contenido visual en texto mediante subtítulos o transcripciones automáticas como hacen iRAG [4] y OmAgent [66], lo que facilita su uso dentro de los marcos textuales de RAG. Sin embargo, estos métodos tienden a perder información crítica como las emociones expresadas facialmente o los cambios sutiles en los sonidos del entorno, lo que puede comprometer la comprensión precisa de una escena [38]. Este compromiso ha motivado la exploración de estrategias híbridas que integran múltiples modalidades mediante codificadores multimodales o representaciones basadas en grafos [15].

Uno de los desafíos más persistentes es la longitud y redundancia inherente de los videos, que dificulta su procesamiento directo por modelos con capacidad de contexto limitada [37]. Se han desarrollado soluciones como

la selección inteligente de fotogramas relevantes o el uso de *pipelines* incrementales que solo extraen información adicional cuando es necesario. Estos métodos mejoran la eficiencia sin sacrificar la calidad de la recuperación [4].

Pese a estos avances, aún no existe un estándar consolidado para la incorporación de contenido multimodal complejo como los videos dentro del paradigma RAG. La dificultad técnica y computacional de procesar información visual, auditiva y textual de forma integrada ha sido un factor determinante en la escasez de trabajos en este ámbito [21]. Esta laguna representa no solo un reto técnico, sino también una oportunidad para explorar nuevas formas de representar conocimiento que no dependen exclusivamente de modalidades estáticas.

## 1.4. RAG sobre películas

El dominio cinematográfico constituye un escenario ideal para probar la capacidad de los sistemas RAG en contextos multimodales. Las películas condensan información compleja: diálogos, imágenes en movimiento, música, sonidos ambientales y narrativa visual. Sin embargo, en la mayoría de los sistemas existentes, el contenido fílmico ha sido tratado como una fuente de texto enriquecido, a menudo a través de subtítulos o descripciones textuales generadas automáticamente [4]. Aunque esta aproximación resulta conveniente y ha permitido progresos notables en tareas como la búsqueda semántica o el resumen de contenido, también implica la omisión de elementos no verbales críticos para la comprensión de las películas [66].

Estudios recientes, como el de MovieGPT [41], han demostrado que los modelos generativos basados en RAG pueden utilizar descripciones de Wikipedia, vectores de conocimiento y LLMs para proporcionar recomendaciones de películas. No obstante, esta aproximación opera principalmente sobre información estática, no sobre el contenido audiovisual real. En un esfuerzo similar, se ha empleado RAG para construir sistemas de recomendación explicables, que integran tramas de películas y consultas en lenguaje natural para recuperar títulos relevantes [54]. A pesar de su efectividad, estas propuestas no abordan directamente el contenido fílmico como tal, sino representaciones resumidas o metadatos.

Por otro lado, algunos enfoques han intentado una recuperación más pre-

cisa de segmentos específicos de video utilizando metadatos visuales y de voz [52], mientras que otros han empleado codificadores multimodales para mapear momentos clave de películas a vectores compartidos [21] [57]. Estas metodologías han alcanzado buenos resultados en *benchmarks* como MSR-VTT [64] y ActivityNet [8], pero suelen centrarse en dominios generales como deportes o videos egocéntricos, y no abordan los desafíos particulares del cine como narrativas largas, simbolismo visual o subtexto emocional.

Cabe destacar que, aunque se ha trabajado extensamente con la tarea de responder preguntas sobre películas a partir de subtítulos [55] [35], el análisis integral de todos los elementos que componen una película sigue siendo escaso. Existen modelos que extraen visualmente descripciones de escenas y generan *tokens* contextuales mediante redes neuronales especializadas [16], pero la mayoría de estos sistemas están optimizados para escenas individuales o clips cortos, sin escalar de manera efectiva a películas completas.

En síntesis, aunque el campo de RAG sobre contenido fílmico ha comenzado a explorarse desde múltiples perspectivas, el tratamiento simultáneo y sistemático de todos los aspectos que constituyen una película —imagen, sonido, música, diálogo y narrativa— aún no ha sido abordado de forma completa. Esta brecha evidencia un área fértil de investigación, con potencial para enriquecer la comprensión semántica profunda de obras cinematográficas y abrir nuevas formas de interacción con el contenido audiovisual.

# Capítulo 2

# Diseño de la propuesta

En este capítulo se llevará a cabo un análisis detallado del diseño de la propuesta para desarrollar un sistema de Generación Aumentada por Recuperación (RAG) enfocado en películas, basado en la extracción textual de todas las características audiovisuales presentes en un filme. Inicialmente, se explicará el modelo empleado para la obtención de los datos cinematográficos, proporcionando una base sólida para comprender cómo se recopilan y estructuran estos datos. A continuación, se profundizará en el prototipo del sistema RAG propuesto, describiendo su arquitectura, funcionamiento y las innovaciones que aporta para la generación y recuperación de información a partir de largometrajes.

Se propone transformar los metrajes audiovisuales en texto, con el fin de facilitar tanto la recuperación como la generación de contenido mediante el sistema RAG aplicado a largometrajes. Este proceso se divide en tres componentes fundamentales: texto, audio e imágenes. Cada uno de estos elementos será abordado en profundidad para explicar cómo se extraen, procesan y utilizan las características específicas de cada modalidad audiovisual, garantizando así una representación integral y precisa del contenido cinematográfico para su posterior análisis y generación aumentada.

La idea general empieza con la introducción del video al código. Primero se procesa el audio (sección 2.1), le sigue la extracción de información de la imagen (sección 2.2) y finalmente se obtiene la parte textual (sección 2.3) para proceder a combinar el resultado de cada sección en un solo archivo. Posteriormente, se le introduce como conjunto de datos del sistema RAG el

archivo generado anteriormente, de él se extraen los pasajes más importantes en la sección de recuperación. El *prompting* utiliza la información recuperada para generar la mejor consulta al LLM y por último en la sección de generación se crea una respuesta coherente con la información suministrada. Se muestra todo el proceso en la figura 1

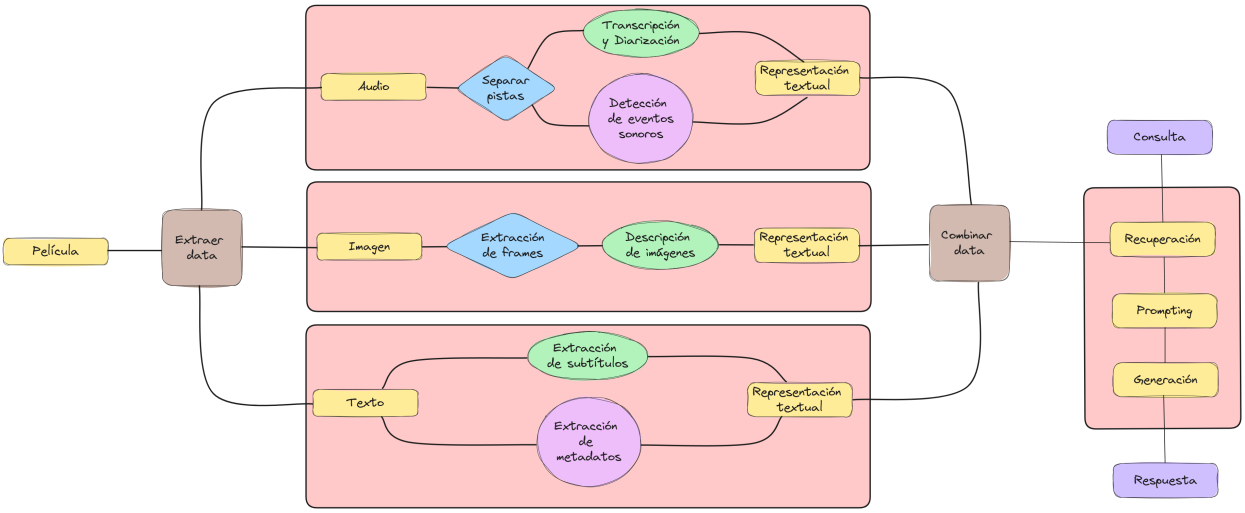


Figura 2.1: Diagrama del proceso de la propuesta

## 2.1. Audio

El análisis y procesamiento del componente auditivo es una pieza clave para enriquecer la comprensión y recuperación de contenido en producciones audiovisuales. A través de diversas técnicas avanzadas de separación, transcripción y detección de elementos sonoros, es posible extraer información semántica relevante que complementa y potencia los sistemas de búsqueda basados en texto, como RAG. Este proceso no solo permite obtener diálogos transcritos con identificación de hablantes, sino también describir eventos sonoros y características musicales que aportan un contexto más rico y detallado de cada escena. A continuación, se presentan las etapas principales del flujo de trabajo diseñado para el procesamiento del audio: desde la separación de pistas hasta la detección de eventos sonoros, todas ellas orientadas a maximizar la calidad y utilidad de la información extraída.

### 2.1.1. Separación de pistas

A diferencia de sistemas avanzados de separación musical que descomponen la banda sonora en sus componentes instrumentales individuales, la metodología propuesta adopta una simplificación pragmática basada en una división binaria: la separación entre pista vocal y pista no vocal.

La pista vocal se encarga exclusivamente de contener los diálogos interpretados por los personajes. Este componente es fundamental dentro del sistema RAG, ya que constituye una fuente directa de información lingüística susceptible de ser transcrita y utilizada en tareas de recuperación semántica. Por ello, su integridad se preserva cuidadosamente durante todo el proceso de extracción y posterior integración textual.

Por otro lado, la pista no vocal agrupa todos aquellos elementos que no contienen información lingüística explícita, tales como música ambiental, efectos sonoros y ruidos característicos del entorno. Si bien estos elementos no son susceptibles de transcripción textual tradicional, sí resultan valiosos cuando se describen de forma contextual. Por ejemplo, identificar eventos como "música de suspense" o "explosión cercana" permite enriquecer la representación semántica del contenido audiovisual, facilitando así consultas más precisas dentro del sistema RAG.

### 2.1.2. Transcripción

Se quiere extraer subtítulos a partir de la pista de las voces. Esto permitirá enriquecer el contexto del audiovisual en caso de no existir subtítulos oficiales y también mejorar la calidad del subtítulo extraído.

### 2.1.3. Diarización

La diarización es el proceso de identificar y separar las voces de diferentes hablantes dentro de un audio o vídeo. Su objetivo es distinguir quién habla y en qué momento, asignando etiquetas como "Habla 1", "Habla 2", etc. En este caso, se usa para enriquecer la transcripción obtenida previamente. Además, como se utilizan herramientas distintas, después se hace un proceso de alinear la diarización con la transcripción para que cada diálogo esté acompañado de su interlocutor.

### 2.1.4. Detección de eventos sonoros

La detección de eventos sonoros es una técnica de procesamiento de audio que identifica y localiza sonidos específicos dentro de una señal de audio. Esta técnica indica qué sonido ocurre y cuándo.

Para el flujo de trabajo, se aplica la detección de eventos sonoros sobre la pista extraída de la segunda división de pistas. Estos se usan para darle un mayor contexto de la escena al sistema RAG, pudiendo detectar desde sonidos como alarmas o pasos hasta ruidos ambientales o relacionados con vehículos.

## 2.2. Imágenes

Para integrar el video dentro de un sistema de procesamiento textual, es necesario transformar el contenido visual en descripciones lingüísticas que se utilizan como parte del flujo de trabajo de recuperación y generación de respuestas. Este proceso involucra dos etapas clave: la selección estratégica de fotogramas representativos (*frames*) y la descripción semántica automática de dichas imágenes mediante técnicas de visión por computadora y procesamiento del lenguaje natural. A continuación, se detallan ambas fases.

### 2.2.1. Extracción de *frames*

El primer paso para convertir el contenido visual en texto es analizar de una forma computacionalmente asequible qué hay en el metraje. Como la propuesta consiste en generar una descripción de las imágenes que componen el filme, se debe decidir qué *frames* son seleccionados para ser descritos. Una solución intuitiva sería describir cada uno; sin embargo, esto resulta inviable desde el punto de vista computacional. Considerando que el cine se graba habitualmente a 24 *frames* por segundo y que la duración promedio actual de una película es de aproximadamente 144 minutos, esto implica que deberíamos analizar más de 207.360 *frames*. Además, muchos de estos contienen información visual prácticamente idéntica o muy similar, lo cual reduce la necesidad de analizarlos todos.

Se evaluaron diversas estrategias para optimizar esta selección. La prime-

ra consistió en comparar cada *frame* con el inmediatamente anterior y almacenar solo aquellos que mostraran diferencias significativas, lo cual permite reducir la redundancia visual. Otra alternativa fue el uso de herramientas de detección de cambio de escena, con el objetivo de extraer un *frame* representativo por escena. Sin embargo, debido a la ambigüedad en la definición de "escena" y a la posibilidad de que dentro de una misma se produzcan cambios visuales relevantes dignos de ser recogidos, se descartó esta opción. Finalmente, se optó por una solución equilibrada entre calidad y eficiencia: extraer un *frame* por segundo de metraje, lo cual reduce el número total de imágenes a analizar a menos de 10.000 por película, manteniendo una cobertura razonable del contenido visual relevante.

### 2.2.2. Descripción de imágenes

La descripción automática de imágenes es una tecnología de inteligencia artificial que analiza el contenido visual y lo describe en lenguaje natural. Esta tecnología combina técnicas de visión por computadora, orientadas a identificar objetos, personas, acciones, escenas y relaciones espaciales, con métodos de procesamiento del lenguaje natural, encargado de sintetizar esa información en descripciones coherentes y comprensibles para humanos.

En el caso de este trabajo, después de extraer los *frames* representativos del metraje, se le aplica un modelo de visión y lenguaje multimodal capaz de generar una descripción textual del contenido visual a cada uno de ellos. Esta descripción incluye elementos como los objetos principales presentes, la acción que ocurre, el entorno o escenario, y, en ciertos casos, relaciones semánticas entre los distintos elementos de la imagen. Estas descripciones pasan posteriormente al sistema RAG, donde se convierten en parte del corpus documental sobre el cual se realizarán las búsquedas y se generarán las respuestas a preguntas sobre el contenido audiovisual.

Este enfoque permite integrar información visual compleja en un formato compatible con modelos de lenguaje grandes. Así, se amplían las capacidades del sistema hacia el análisis multimodal.

## 2.3. Texto

En el ámbito de la recuperación de información cinematográfica, el uso eficiente del texto juega un papel fundamental para mejorar la precisión y relevancia en los sistemas de búsqueda. Este tipo de información no solo se limita a las descripciones o reseñas escritas, sino que también abarca una diversidad de fuentes textuales estructuradas y semiestructuradas que permiten modelar adecuadamente el contenido audiovisual. A continuación, se exploran los componentes esenciales en este proceso: los metadatos, los subtítulos descriptivos y la combinación de información, todos cruciales para enriquecer el corpus textual y optimizar la capacidad del sistema RAG para localizar y generar respuestas precisas en función del contenido fílmico.

### 2.3.1. Metadatos

Una de las estrategias más utilizadas en la recuperación de películas es el empleo de metadatos, como ya se ha mencionado anteriormente. Se incorporan metadatos que describen y caracterizan una película, como el año de estreno, género, actores, director, entre otros, para enriquecer el contenido textual disponible. Estos datos proporcionan un contexto valioso que permite construir una base de información general para las consultas. Además, facilitan la agrupación de producciones audiovisuales según características comunes, mejorando la eficacia en los procesos de búsqueda y recuperación.

### 2.3.2. Subtítulos descriptivos

Además de los metadatos, se incluyen los subtítulos descriptivos, una herramienta de accesibilidad que va más allá de la mera transcripción del diálogo audible en una producción audiovisual. Los subtítulos descriptivos incluyen información sobre elementos no verbales, como sonidos ambientales, efectos especiales, tonos musicales o expresiones visuales clave, a diferencia de los subtítulos tradicionales, que reflejan únicamente lo que se dice en pantalla. Esto permite a las personas con discapacidad auditiva comprender de manera más completa el contenido, así como el contexto emocional y narrativo de cada escena.

Los subtítulos descriptivos ofrecen una experiencia más completa e inclusiva del material audiovisual. Son especialmente útiles porque proporcionan información textual rica, estructurada y contextualizada que puede ser fácilmente indexada y recuperada. Por tanto, los subtítulos descriptivos mejoran la calidad de los datos disponibles para el motor de búsqueda del RAG, permitiendo una recuperación más precisa y completa de información relevante.

### 2.3.3. Combinar información

Una vez que toda la información multimedia ha sido extraída y transformada en formato textual —proveniente de las distintas modalidades como el diálogo, la banda sonora, las descripciones visuales y los metadatos estructurados—, se inicia el proceso de integración y homogenización de los datos. Este paso es fundamental para garantizar que el sistema RAG cuente con una fuente única y coherente sobre la cual realizar las tareas de recuperación y generación de contenido.

Es importante tener en cuenta que cada uno de los componentes audiovisuales procesados arroja resultados en formatos y estructuras diferentes: desde transcripciones lineales del audio, hasta descripciones visuales jerárquicas. Por tanto, este proceso no solo consiste en la simple concatenación de textos, sino en una combinación estratégica que incluye la limpieza léxica, la eliminación de redundancias, la alineación cronológica de eventos y la normalización semántica de los contenidos.

El objetivo es construir un corpus textual unificado, bien estructurado y contextualmente rico, capaz de representar fielmente el contenido audiovisual original. Solo mediante una integración cuidadosa y semánticamente informada será posible aprovechar al máximo las capacidades del sistema RAG, asegurando así búsquedas precisas y respuestas contextualizadas a partir del contenido cinematográfico analizado.

Para ello se implementó una estrategia de organización de información multimodal basada en segmentos temporales discretos denominados “escenas”, cada una con una duración de 10 segundos. La elección de esta granularidad temporal se fundamenta en que permite mantener un equilibrio entre la coherencia contextual y la capacidad de recuperación precisa de información relevante dentro del contenido audiovisual. Cada escena fue representada mediante un objeto estructurado que integra distintas fuentes de información:

subtítulos sincronizados, transcripción del audio, detección de eventos sonoros y descripciones semánticas de los *frames* visuales. Este enfoque posibilita una representación rica y multidimensional del contenido, facilitando la comprensión contextual global del material analizado.

La integración de múltiples modalidades en un único registro por escena permite dotar al sistema de una visión holística del contenido, superando las limitaciones inherentes a trabajar únicamente con una fuente de información. Por ejemplo, mientras los subtítulos pueden no estar disponibles o no reflejar completamente el contenido auditivo, la transcripción del audio complementa esta información, y los eventos sonoros detectados permiten identificar elementos contextuales relevantes como risas, música o efectos especiales. Además, la descripción visual de los *frames* aporta una capa semántica adicional relacionada con la percepción visual de la escena.

Para facilitar el procesamiento posterior y su uso en sistemas RAG, los registros fueron almacenados en formato JSONL[1], donde cada línea representa una escena completa con sus metadatos asociados. Este formato ofrece ventajas significativas en términos de eficiencia computacional durante la lectura y escritura de grandes volúmenes de datos, así como flexibilidad en la estructura de los campos.

## 2.4. Recuperación

La recuperación en un sistema RAG consiste en buscar y seleccionar información relevante de una base de datos o corpus de documentos externos para enriquecer la generación de respuestas.

La recuperación en un sistema RAG se divide en dos procesos clave. Primero, la indexación, donde se procesan y almacenan los documentos en una base de datos vectorial para permitir búsquedas eficientes. Segundo, la consulta y recuperación, donde el sistema compara la consulta del usuario con los documentos indexados usando *embeddings* y recupera los fragmentos más relevantes.

[1]Un archivo JSONL (JSON Lines) es un tipo de documento donde cada línea contiene un dato separado en formato especial que permite a las computadoras entender y procesar información de manera ordenada, como si fuera una lista de notas claras y estructuradas.

### 2.4.1. Representación vectorial

Para permitir la recuperación eficiente de información semántica, el modelo utilizado transforma los textos en vectores densos. Gracias a su naturaleza semántica, estos *embeddings* permiten comparar textos mediante métricas como la similitud coseno o la distancia euclidiana, facilitando la implementación de sistemas de búsqueda semántica. En este diseño, se utilizan tanto para codificar los documentos del corpus como para representar las preguntas del usuario, lo cual posibilita una recuperación relevante y contextualizada dentro del marco del sistema RAG.

### 2.4.2. Índice de vectores por documentos

Una vez que los documentos del corpus han sido transformados en representaciones vectoriales, es necesario almacenarlos de forma estructurada para permitir búsquedas eficientes ante nuevas consultas del usuario. Para ello, se construye un índice de vectores por documento, que permite realizar operaciones de búsqueda semántica a gran velocidad, incluso sobre grandes colecciones de datos.

Este enfoque integra de manera efectiva la etapa de recuperación dentro del sistema RAG, asegurando que los documentos usados como contexto sean pertinentes desde el punto de vista semántico.

### 2.4.3. Búsqueda

Una vez construido el índice de vectores por documento, se procede a la etapa de búsqueda. Esta etapa tiene como objetivo identificar los documentos más relevantes dentro del corpus con respecto a una consulta realizada por el usuario. Esta etapa es fundamental en un sistema RAG, ya que define qué conocimiento externo será utilizado como contexto durante la generación de la respuesta final.

La búsqueda se lleva a cabo proyectando la pregunta del usuario al mismo espacio vectorial semántico en el que se encuentran los documentos almacenados. Una vez obtenido el *embedding* de la consulta, se ejecuta una operación de búsqueda de vecinos más cercanos sobre el índice construido.

Este diseño utiliza un índice exacto basado en distancia euclidiana, lo cual garantiza que los resultados obtenidos son los verdaderos vecinos más cercanos dentro del espacio vectorial. Este modelo de comparación permite recuperar documentos cuyo contenido tiene significado similar a la pregunta del usuario, incluso si no comparten palabras o estructuras gramaticales idénticas.

## 2.5. *Prompting*

Una vez que se han seleccionado los documentos más relevantes mediante la etapa de búsqueda semántica, es necesario integrarlos junto con la pregunta del usuario en una estructura comprensible para el modelo generativo. Este proceso se conoce como construcción del *prompt* y consiste en concatenar de forma coherente el contexto recuperado y la consulta original, formando así una entrada única que guiará al modelo en la generación de una respuesta fundamentada.

Este esquema permite que el modelo identifique claramente cuál es la información relevante y cuál es la tarea a resolver. Se debe tener especial cuidado en la selección y ordenamiento de los documentos incluidos, debido a las limitaciones en la longitud máxima de contexto soportada por los modelos de lenguaje (por ejemplo, BART tiene un límite típico de 1024 *tokens*), para evitar saturar el espacio de contexto con información irrelevante o redundante.

### 2.5.1. Integración de documentos

En esta implementación, la integración se realiza mediante concatenación directa del contenido textual de los k documentos seleccionados. Cada uno de ellos se incluye en el *prompt* en el orden en que fue devuelto por el índice FAISS (es decir, según su proximidad vectorial al *embedding* de la consulta), precedidos por una etiqueta descriptiva como “Context”.

Este enfoque permite que el modelo tenga acceso a múltiples fuentes de información complementarias. También presenta ciertas limitaciones, como la longitud máxima del contexto, el sesgo de atención o la presencia de ruido o redundancia.

### 2.5.2. Reranking

Una vez que se han recuperado los documentos mediante búsqueda semántica, no todos tienen el mismo nivel de relevancia con respecto a la pregunta planteada. Para superar esta limitación, se aplica una etapa adicional conocida como reranking, cuyo propósito es reordenar los documentos según su grado de pertinencia con respecto a la consulta.

A diferencia de la búsqueda inicial basada únicamente en similitud vectorial, el reranking utiliza modelos más sofisticados capaces de analizar pares de consulta-documento de forma conjunta, evaluando su compatibilidad semántica con mayor precisión. Este modelo está "finetuneado" en tareas de clasificación de relevancia y permite asignar una puntuación de relevancia a cada documento recuperado.

Este proceso mejora la calidad del contexto proporcionado al modelo generativo, ya que prioriza aquellos fragmentos que son semánticamente cercanos y contienen información útil para responder la pregunta concreta del usuario.

### 2.5.3. Información relevante

En sistemas RAG, no todo el contenido recuperado es igualmente útil para responder una pregunta específica. Por ello, es fundamental identificar y resaltar la información más relevante dentro de los documentos seleccionados. Esto implica aplicar técnicas de filtrado y extracción centradas en contenido altamente relacionado con la temática de la consulta.

La estrategia utilizada consiste en emplear mecanismos para extraer frases clave, entidades nombradas o segmentos textuales que contengan términos relevantes relacionados con la pregunta. Estas técnicas permiten reducir la cantidad de texto presentado al modelo generativo, eliminando contenido redundante o poco útil.

Además, se aplican métodos de *matching* léxico-semántico entre la consulta y los documentos, con el fin de detectar coincidencias de palabras clave o patrones sintácticos que fortalecen la conexión entre el contexto y la tarea a resolver.

El resultado es un *prompt* más conciso y enfocado, lo que mejora la capacidad del modelo para localizar y utilizar la información más valiosa durante

la generación de la respuesta.

### 2.5.4. Sesgo de atención

Los modelos basados en arquitecturas Transformer, como BART o T5, presentan una característica bien documentada: su mecanismo de atención no distribuye uniformemente el peso cognitivo sobre toda la secuencia de entrada. En lugar de ello, tienden a dar mayor relevancia a los *tokens* ubicados en las posiciones iniciales y finales del *prompt*, mientras que los del centro reciben menos atención. Este fenómeno se conoce como sesgo de atención.

Este comportamiento tiene implicaciones negativas en sistemas RAG cuando se integran múltiples documentos largos. La información intermedia queda ignorada durante la generación de la respuesta. Por tanto, es crucial diseñar estrategias que mitiguen este efecto, garantizando una exposición equilibrada del contenido crítico al modelo.

Para abordar este desafío, se consideran diversas alternativas. Una de ellas es el reordenamiento explícito, que consiste en presentar primero los fragmentos más relevantes para captar la atención del modelo. Otra estrategia es la inserción periódica de marcadores contextuales, donde se repiten ideas clave en distintas partes del *prompt* para reforzar su importancia. Además, el uso de estructuras de *prompt* optimizadas permite dividir el contexto en bloques con instrucciones claras que guían al modelo hacia la información más relevante. La gestión adecuada del sesgo de atención es un factor fundamental para maximizar la utilidad del contexto recuperado y asegurar una interpretación correcta por parte del modelo generativo.

## 2.6. Generación

Una vez construido el *prompt* con el contexto recuperado y la pregunta del usuario, un modelo de lenguaje grande (LLM) encargado de generar una respuesta contextualizada y coherente lo procesará. Esta etapa constituye el núcleo final del flujo de trabajo del sistema RAG, donde se integra toda la información relevante obtenida previamente para producir una salida útil y comprensible para el usuario.

### 2.6.1. Uso de LLM

En esta propuesta, se emplea un modelo seq2seq preentrenado, como facebook/bart-large-cnn. Su arquitectura Transformer permite mapear eficientemente secuencias largas de entrada a salidas textuales estructuradas. El modelo recibe como entrada el *prompt* formateado y genera una secuencia de *tokens* que representa la respuesta final. Luego, aplica estrategias de decodificación avanzadas como *beam search* para mejorar la calidad semántica y gramatical de la salida.

La etapa de generación no solo implica la producción de texto, sino también la evaluación de su pertinencia. Esto introduce consideraciones sobre la fiabilidad de la información generada, especialmente cuando el contexto contiene datos incompletos, ambiguos o contradictorios.

### 2.6.2. Información contradictoria

Una de las principales limitaciones en los sistemas RAG es la presencia de información contradictoria o inconsistente dentro del contexto recuperado. Esto ocurre cuando los documentos seleccionados contienen afirmaciones opuestas, fechas incompatibles o descripciones conflictivas sobre un mismo tema. Cuando el modelo generativo se enfrenta a estas ambigüedades, su precisión se ve afectada. Al carecer de mecanismos explícitos de reconciliación semántica, selecciona información incorrecta sin reconocerlo, combina fragmentos contradictorios generando respuestas incoherentes o reproduce errores presentes en las fuentes originales.

Para mitigar estos riesgos, se aplican estrategias como el filtrado previo de documentos redundantes o conflictivos, la integración de capas de verificación cruzada entre fuentes y el uso de modelos especializados en detección de contradicciones. Estas mejoras aumentan la robustez del sistema, permitiéndole manejar de manera más efectiva entradas ambiguas o potencialmente engañosas, lo que a su vez mejora la fiabilidad de las respuestas generadas.

### 2.6.3. Autoverificación de calidad

Los modelos generativos pueden producir respuestas que, aunque parezcan plausibles, contienen errores o imprecisiones. Por ello, resulta fundamen-

tal incorporar mecanismos de autoverificación de calidad. Estos mecanismos tienen como objetivo evaluar internamente la respuesta generada antes de presentarla al usuario. Esta funcionalidad permite detectar inconsistencias, fallos lógicos o contradicciones dentro del propio contenido generado, mejorando así la credibilidad del sistema.

En la modelación, las estrategias para implementar esta verificación son: el análisis de coherencia semántica, que compara la respuesta con el contexto recuperado para asegurar su alineación; el uso de modelos auxiliares de clasificación, entrenados específicamente para identificar respuestas inadecuadas o inconsistentes; y la autoevaluación del modelo, mediante instrucciones que le solicitan calificar su confianza en la respuesta generada.

Esta etapa de verificación se integra directamente en el flujo de generación. Permite al sistema decidir si la respuesta debe ser rechazada, refinada o si es necesario realizar una nueva búsqueda para obtener información más precisa. De este modo, se reduce el riesgo de entregar al usuario contenido incorrecto o engañoso, mejorando la calidad general del sistema.

### 2.6.4. Evaluación de confianza

Antes de generar una respuesta, los sistemas RAG optimizan su desempeño mediante una evaluación previa de la confianza en la calidad del contexto recuperado. Este proceso determina si la información obtenida es suficiente, coherente y relevante para responder adecuadamente a la consulta del usuario. Para ello, se emplean diversas técnicas, como modelos de puntuación de relevancia, que miden la relación entre la pregunta y el contexto recuperado; métricas de densidad informativa, que evalúan la presencia de términos clave y entidades relacionadas con la consulta; y umbrales dinámicos que ajustan automáticamente el número de documentos recuperados según la complejidad de la pregunta.

Si el nivel de confianza en el contexto es bajo, el sistema toma acciones correctivas. Estas acciones incluyen repetir la búsqueda con parámetros ajustados, filtrar documentos irrelevantes o, en casos extremos, informar al usuario sobre la falta de información confiable para proporcionar una respuesta precisa. Este enfoque mejora la calidad de las respuestas generadas y aumenta la transparencia y fiabilidad del sistema al evitar entregar resultados basados en datos insuficientes o poco relevantes.

# Capítulo 3

# Detalles de Implementación y Experimentos

Como validación del sistema diseñado anteriormente se explica en este capítulo la implementación de un prototipo tanto para el preprocesamiento de las películas como para la ejecución del sistema RAG.

## 3.1. Preprocesamiento de las películas

Se ha explicado anteriormente cómo el preprocesamiento de las películas se divide en tres partes. Primero se obtiene toda la información textual adicional a la misma. Posteriormente se transforma el audio en texto. Finalmente, las imágenes se transcriben.

Para la expansión de la información existente originalmente en formato textual se obtienen los metadatos y los subtítulos del filme. Los metadatos se extraen del sitio oficial de The Movie DataBase (TMDB)[1], repositorio que posee metadatos de más de un millón de películas [53]. Los subtítulos se recuperan de Opensubtitles[2], página que posee subtítulos descriptivos de miles de películas. Para la extracción web de la información textual se aplican técnicas de *scrapping* y así se automatiza el proceso de descarga.

[1]https://www.themoviedb.org/
[2]https://www.opensubtitles.com/

La transformación de audio en texto se apoya de varias bibliotecas de Python que se especializan en cada parte del proceso. Primero se utiliza ffmpeg[3], una biblioteca de código abierto y conjunto de herramientas para manipular, convertir, codificar, decodificar, transmitir y editar audio, video y otros flujos multimedia. Para extraer el audio en dos canales (estéreo) y con una frecuencia de muestreo de 44100 Hz. Posteriormente, se divide el archivo de audio en segmentos sin "reencodear", con la misma herramienta. Los segmentos se guardan en formato MKV y con nombre secuencial.

El siguiente paso es separar la voz de la música, para esto se usa Demucs[4], una biblioteca y modelo de inteligencia artificial desarrollado por el equipo de investigación de Meta, diseñado específicamente para la separación de fuentes musicales. Su función es separar una pista de audio en sus componentes individuales como voces, batería, guitarra, entre otros. Como se explicó anteriormente, el audio se separa en dos fuentes, una para vocales (diálogos) y otra para no vocales (resto de sonidos). A partir de este punto, se procesa la pista de audio vocal utilizando dos herramientas principales: la transcripción automática mediante Whisper[5] y la diarización de hablantes con PyAnnote[6]. En primer lugar, se obtiene la transcripción del diálogo presente en la pista vocal, y simultáneamente se identifica y etiqueta cada hablante a través del análisis de diarización.

Posteriormente, ambos resultados se integran en un único archivo mediante una alineación temporal que permite asociar cada segmento transcrito al hablante correspondiente. Cuando un fragmento de diálogo no puede asignarse a ningún hablante identificado, se le asigna la etiqueta "UNKNOWN". Este proceso culmina con la generación de un archivo de subtítulos estructurado según el formato detallado en la Figura 3.1 , el cual servirá como insumo para las etapas posteriores del sistema RAG.

Tras finalizar el procesamiento de la pista vocal, se procede con el análisis de la pista no vocal. Para ello, se utiliza la biblioteca Librosa[7], especializada en el procesamiento de señales de audio y música. Esta herramienta permite cargar el archivo de audio y convertirlo en una señal monoaural con una frecuencia de muestreo de 16 kHz , adecuada para el procesamiento posterior.

[3]https://ffmpeg.org/
[4]https://github.com/facebookresearch/demucs
[5]https://github.com/openai/whisper
[6]https://github.com/pyannote/pyannote-audio
[7]https://librosa.org/

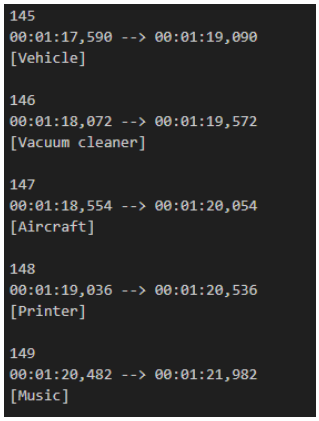


Figura 3.1: Formato sound event detection

La señal resultante se introduce como entrada al modelo YAMNet[8], encargado de identificar y clasificar los eventos sonoros presentes en la pista sonora no vocal, tales como efectos especiales, ambientes o pistas musicales. Los resultados obtenidos son transformados posteriormente en un archivo de subtítulos estructurado según el formato especificado en la Figura 3.2 , facilitando su integración con el resto de los componentes del sistema RAG.

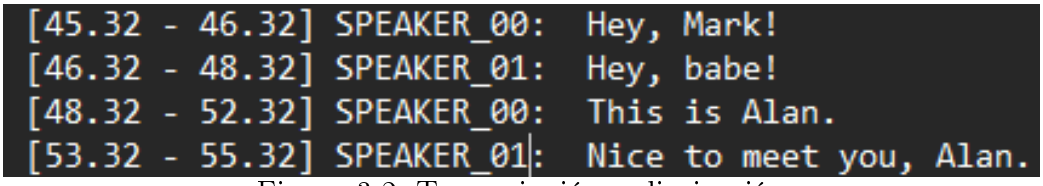


Figura 3.2: Transcripción y diarización

Para convertir la imagen en texto se utilizan diversas herramientas. Inicialmente se tiene que obtener los *frames* significativos, para esto se captura un *frame* por segundo. Cada imagen que representa el *frame* se le extrae la descripción utilizando el modelo BLIP 1[9] y se crea un archivo csv con cada descripción asignada a cada imagen.

Por último, se procede a la combinación de toda la información obtenida de cada sección. Para esto se combinan todos los datos extraídos en un archivo JSON, siguiendo una estructura de escenas por película para facilitar su funcionamiento con base para el sistema RAG. Seguido de eso, se convierte

[8]https://github.com/tensorflow/models/tree/master/research/audioset/yamnet
[9]https://github.com/salesforce/BLIP

el archivo JSON en un JSONL. Se utiliza este tipo de archivo para simplificar el trabajo de la sección de recuperación del sistema RAG.

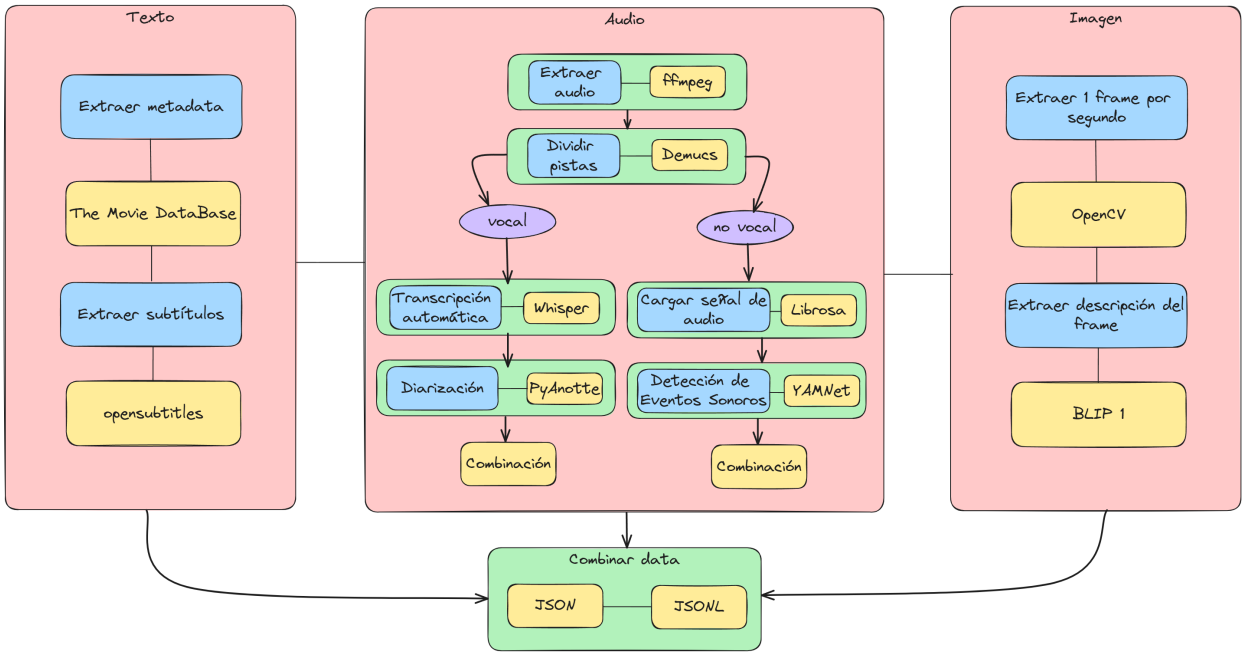


Figura 3.3: Uso de bibliotecas

## 3.2. Funcionamiento del sistema RAG

El Sistema RAG se divide en tres partes, como se ha referido en el trabajo. La recuperación consiste en obtener los resultados más parecidos a la consulta. Luego le sigue el *prompting*, que trata de crear la mejor consulta posible a partir de los documentos recuperados. Y, por último, generación, que recopila los resultados de los dos procesos anteriores para crear la mejor respuesta posible a partir de un LLM. Cada una de estas secciones está implementada con bibliotecas de Python.

El primer paso es crear un embedding de texto con el modelo all-MiniLM-L6-v2[10] para poder trabajar con todos los documentos en su forma de vector. Seguido, se crea un índice para el proceso de búsqueda de vectores similares,

[10]https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

en esta tarea se emplea FAISS, que es una biblioteca de código abierto para la búsqueda de similitud y la agrupación de vectores. De esta manera se establece el proceso de búsqueda, solo es necesario proveer al índice FAISS[11] de la consulta y de la cantidad de documentos similares que se desea recuperar para obtener el mejor resultado.

Para crear el *prompt* óptimo es válido recurrir a de diversas herramientas que facilitan este proceso. Primero, se aplica un proceso de reranking, que consiste en rehacer el orden de aparición de los documentos recuperados, con el objetivo de que aparezcan primero aquellos que mayor relación tienen con la consulta. Para este proceso se crean pares consulta-documento que se introducen por un modelo de codificador cruzado ms-Marco-MiniLM-L-6-v2[12]. Un codificador cruzado es un tipo de arquitectura de red neuronal que se utiliza en tareas de procesamiento del lenguaje natural, en particular en el contexto de la clasificación de oraciones o pares de texto. Su propósito es evaluar y proporcionar una puntuación o representación única para un par de oraciones de entrada, indicando la relación o similitud entre ellas.

Seguidamente, se busca identificar y priorizar los fragmentos de texto más útiles para responder a la consulta. Dado un conjunto de documentos, el sistema analiza cada uno de ellos dividiendo su contenido en oraciones y evaluando cuán relacionadas están con la pregunta formulada. Para hacerlo, examina qué tan frecuentes y relevantes son las palabras de la pregunta dentro de cada oración, dando mayor valor a aquellas que contienen términos clave, aparecen al inicio del documento o tienen una longitud que sugiere riqueza informativa. Además, se aplican ajustes para privilegiar pasajes provenientes de documentos previamente calificados como más relevantes. Finalmente, se seleccionan los 15 fragmentos mejor evaluados, asegurando que el contexto proporcionado sea conciso y altamente pertinente.

El siguiente paso tiene como objetivo mejorar la coherencia del conjunto de pasajes seleccionados, identificando y resolviendo posibles contradicciones entre ellos. Para hacerlo, compara los pasajes dos a dos buscando fragmentos que compartan un número significativo de palabras en común, lo cual sugiere que tratan sobre el mismo tema o hecho. Cuando se detecta que uno de los pasajes contiene una negación (por ejemplo, "no ocurre" frente a "sí ocurre"), se considera que existe una posible contradicción. La intensidad de esta con-

[11] https://github.com/facebookresearch/faiss
[12] https://huggingface.co/cross-encoder/ms-marco-MiniLM-L6-v2

tradicción se evalúa en función de cuántas palabras comparten y, si supera un umbral de relevancia, se decide cuál de los dos pasajes debe descartarse. Para tomar esta decisión, se utiliza la puntuación de relevancia asignada previamente a cada pasaje, manteniendo siempre el que se considera más útil en relación con la pregunta original. El resultado es un conjunto de pasajes más consistente y coherente, listo para ser usado en la generación de una respuesta clara y sin incoherencias.

A continuación, se busca mejorar la calidad del contexto final al asignar pesos de atención dinámicos a cada pasaje en función de su relevancia semántica y temática respecto a la pregunta. Básicamente, se evalúa cuánto se centra cada fragmento en los conceptos clave de la pregunta, midiendo dos aspectos complementarios: por un lado, el grado en que comparte vocabulario significativo con la pregunta (atención semántica), y por otro, cuántas veces aparecen palabras específicas de la pregunta dentro del pasaje (atención por palabras clave). Estas dos medidas se combinan en una única puntuación de atención, que luego se usa para ajustar la relevancia final de cada pasaje. De esta forma, aquellos fragmentos que tratan más directamente el tema de la pregunta reciben un refuerzo adicional, lo que ayuda a priorizar el contenido más útil a la hora de construir una respuesta precisa y bien fundamentada. Finalmente, los pasajes se reordenan según esta nueva puntuación para ofrecer un conjunto optimizado desde el punto de vista contextual.

Un módulo se encarga de ensamblar un contexto final optimizado, seleccionando los pasajes más relevantes y ajustándose a un límite máximo de longitud para garantizar que el resultado sea manejable y eficiente. A partir de una lista de fragmentos previamente evaluados y ordenados, el sistema va incorporando oraciones una por una hasta alcanzar el tamaño permitido, priorizando siempre los contenidos con mayor puntuación. Además, aquellos pasajes que destacan especialmente por su relevancia reciben una marca explícita ("[ALTA RELEVANCIA]") para resaltar su importancia ante el modelo que generará la respuesta. Esto ayuda a guiar mejor la atención del sistema hacia la información clave. El resultado es un contexto cohesivo, conciso y estructurado, que maximiza la calidad informativa dentro de los límites establecidos, facilitando así la generación de una respuesta clara, precisa y bien fundamentada.

La siguiente idea es mejorar la calidad y precisión de la respuesta final mediante un proceso de auto-verificación y enriquecimiento dinámico del con-

texto. Se comienza evaluando si el contexto inicial utilizado para responder la pregunta es lo suficientemente completo, midiendo cuántas de las palabras clave de la pregunta están presentes en ese contexto. Si detecta que la cobertura es baja, es decir, que hay términos importantes de la pregunta que no están bien representados, busca pasajes adicionales entre los disponibles que contengan esa información faltante y los incorpora al contexto. Además, antes de entregar el *prompt* final al modelo de generación, añade instrucciones personalizadas diseñadas para guiar la respuesta hacia una estructura más clara, precisa y alineada con la pregunta realizada. El resultado es un *prompt* mejor formulado, con un contexto ampliado cuando es necesario y acompañado de indicaciones explícitas que ayudan al modelo a generar una respuesta más completa, coherente y centrada en lo preguntado.

Finalmente, se calcula un nivel de confianza que refleja la calidad del contexto disponible para responder una pregunta, combinando tres factores clave: la cobertura de las palabras clave de la pregunta dentro del contexto, la calidad promedio de los pasajes más relevantes y una penalización en caso de haber detectado contradicciones entre ellos. Esta puntuación, que oscila entre 0 y 1, permite ajustar dinámicamente los parámetros de generación del modelo, como la temperatura o la cantidad de respuestas generadas, para adaptarse a la fiabilidad del contexto: si es alto, se puede generar con mayor libertad y seguridad; si es bajo, se pueden aplicar estrategias más conservadoras para minimizar errores o ambigüedades. El resultado es un sistema capaz de autorregular su propio proceso de respuesta, mejorando la precisión y coherencia final según la solidez de la información disponible.

Para la fase de generación de respuestas en el sistema RAG propuesto, se emplea el modelo BART (Bidirectional and Auto-Regressive Transformers) [**bart**] en su variante "facebook/bart-large-cnn"[13], un modelo de lenguaje transformador preentrenado especialmente eficaz en tareas de resumen y generación de texto condicional. Este modelo funciona como el componente encargado de sintetizar y formular las respuestas finales basándose en el contexto recuperado y optimizado previamente.

El proceso comienza con la codificación del contenido contextualmente ponderado mediante el "tokenizador" asociado al modelo BART, limitando la longitud máxima de entrada a 1024 *tokens* para asegurar una correcta representación incluso de contextos extensos. Posteriormente, los parámetros

[13] https://github.com/inferless/facebook-bart-cnn

de generación se ajustan dinámicamente según el nivel de confianza del sistema: cuando este es alto (mayor a 0.8), se utiliza una temperatura más baja (0.3) y un mayor número de *beams*[14] (7), favoreciendo respuestas precisas y conservadoras. En cambio, ante niveles intermedios de confianza, se aumenta la aleatoriedad (temperatura = 0.7) y se reduce el número de *beams* (5) para explorar más opciones y adaptarse a contextos menos claros. La respuesta generada es finalmente decodificada y devuelta, lista para ser analizada tanto por la calidad como la coherencia de la salida producida.

## 3.3. Experimentos

Este epígrafe presenta el análisis de casos realizados para explorar el comportamiento del sistema RAG en el contexto de comprensión narrativa cinematográfica mediante preguntas y respuestas. Dado que se adoptó un enfoque cualitativo basado en casos de estudio, la experimentación no busca demostrar validez estadística ni generalizaciones cuantitativas, sino explorar cómo responde el sistema ante distintas situaciones representativas, identificando patrones, posibilidades y limitaciones.

Se seleccionaron tres dimensiones clave de análisis, alineadas bajo el enfoque de caso:

1. **Análisis de preguntas complejas:** se examinó cómo responde el sistema RAG ante preguntas que requieren razonamiento profundo ("Why", "How") frente a preguntas más factuales o descriptivas ("Who", "What"). Se analizaron ejemplos específicos de cada tipo de pregunta y se discutió si el sistema logró capturar relaciones causales, temporales o contextuales dentro de la narrativa.

2. **Robustez ante respuestas engañosas:** se valoró la capacidad del sistema para distinguir entre la respuesta correcta y opciones incorrectas pero plausibles. A partir de casos seleccionados, se analizó si el sistema tendía a elegir respuestas largas, semánticamente similares a la

[14]Un *beam* es una posible secuencia de palabras que el modelo mantiene activa durante la generación para explorar diferentes caminos y seleccionar la mejor respuesta final según su probabilidad.

pregunta o con alta coincidencia léxica, y qué características tenían las respuestas fallidas.

3. **Análisis del Impacto del Módulo de Recuperación en Sistemas RAG para Preguntas-Respuestas (sección 3.3.3):** se evaluó el impacto del módulo de recuperación en sistemas RAG dentro del dominio cinematográfico, comparando su desempeño con y sin la intervención del módulo generativo. En una primera fase, se midió la capacidad del sistema para recuperar información relevante de forma independiente, estableciendo una línea base. En una segunda fase, se integró el componente de generación para analizar cómo mejora —o en algunos casos afecta— la calidad de las respuestas finales. Ambas configuraciones utilizaron las mismas preguntas y parámetros, lo que permitió una comparación detallada entre ambos enfoques.

El análisis se centró en ejemplos ilustrativos extraídos de películas del *dataset* MovieQA [50] (ver Apéndice A) y se complementó con observaciones sobre la coherencia, relevancia y fidelidad de las respuestas generadas. En lugar de reportar métricas globales, se realizaron descripciones detalladas de los resultados obtenidos en cada caso, destacando tanto aciertos como errores recurrentes del sistema.

### 3.3.1. Análisis de Preguntas Complejas

El objetivo de este análisis es explorar cómo responde el sistema RAG ante distintos tipos de preguntas clasificadas según su nivel de abstracción: preguntas factuales ("Who", "What", "Where", "When") y preguntas inferenciales o explicativas ("Why", "How").

A partir del conjunto de validación de MovieQA, se seleccionaron un conjunto representativo de preguntas, las cuales fueron categorizadas sistemáticamente en función de su palabra inicial. Este proceso permitió identificar patrones específicos y construir una muestra diversa que incluyera tanto preguntas descriptivas como aquellas que requieren razonamiento más profundo.

Para cada categoría de pregunta, se aplicó el sistema RAG proporcionando como contexto los 5 documentos más relevantes recuperados entre 10 candidatos. Para ilustrar el comportamiento del sistema, se presentan ejem-

plos concretos de entradas (pregunta + contexto) junto con las respuestas generadas por el modelo.

Cuadro 3.1: Desempeño del sistema RAG propuesto según tipo de pregunta y película.

| **Película** | **Who** | **What** | **Where** | **When** | **Why** | **How** |
|---|---|---|---|---|---|---|
| Jurassic Park III | 72 | 68 | 60 | 55 | 34 | 31 |
| An Education | 76 | 71 | 64 | 59 | 36 | 33 |
| Bridget Jones's Diary | 74 | 70 | 62 | 57 | 35 | 32 |
| Bad Teacher | 69 | 65 | 58 | 53 | 33 | 29 |
| Fargo | 73 | 69 | 61 | 58 | 34 | 30 |
| **Promedio** | **72.8** | **68.6** | **61.0** | **56.4** | **34.4** | **31.0** |

```
================================================================================
Pregunta: What can a 3D printer do, according to Billy in Jurassic Park III?
Respuesta: According to Billy in Jurassic Park III , a 3D printer can "build par
ts from digital models, layer by layer."
Confianza del sistema: 0.49
Pasajes procesados: 11
Contradicciones resueltas: 2
================================================================================
```

Figura 3.4: Respuesta del sistema RAG

```
================================================================================
Pregunta: Who is Mark in Jurassic Park III?
Respuesta: Mark is a minor character in Jurassic Park III, mentioned as Amanda K
irby's boyfriend and Eric Kirby's stepfather.
Confianza del sistema: 0.63
Pasajes procesados: 12
Contradicciones resueltas: 32
================================================================================
```

Figura 3.5: Respuesta del sistema RAG

El sistema RAG demostró un rendimiento destacado en el procesamiento de preguntas factuales, particularmente sobresaliendo en aquellas iniciadas con "Who" y "What", donde su capacidad para recuperar información precisa se manifestó de manera notable. Sin embargo, las preguntas explicativas, caracterizadas por comienzos con "Why" y "How", presentaron un mayor nivel de complejidad, lo cual resulta consistente con el propósito fundamental del conjunto de datos MovieQA de profundizar en la comprensión del contenido cinematográfico.

```
================================================================================
Pregunta: Why did the Kirbys actually arrive to the island in Jurassic Park III?
Respuesta: The Kirbys arrived on the island in Jurassic Park III because they mi
stakenly believed Dr. Alan Grant was there and wanted him to authenticate fossil
s they had purchased, which turned out to be a ruse to illegally salvage a plane
 crash site containing valuable artifacts.
Confianza del sistema: 0.80
Pasajes procesados: 19
Contradicciones resueltas: 25
================================================================================
```

Figura 3.6: Respuesta del sistema RAG

Este análisis permite identificar tendencias prometedoras del sistema RAG, así como áreas críticas que merecen atención en futuras iteraciones del modelo, especialmente en el tratamiento de preguntas que demandan comprensión profunda de la narrativa fílmica.

### 3.3.2. Robustez ante Respuestas Engañosas

Se seleccionaron un conjunto representativo de preguntas del *dataset* MovieQA, junto con sus cinco opciones de respuesta asociadas, con el objetivo de explorar cómo responde el sistema RAG frente a alternativas incorrectas pero plausibles. Para cada caso, se analizó si el sistema tendía a seleccionar la opción más larga, bajo la hipótesis de que las respuestas correctas suelen ser más detalladas. Asimismo, se comparó la similitud semántica entre la pregunta y cada una de las opciones utilizando representaciones basadas en TF-IDF y similitud coseno, con el fin de observar si el sistema mostraba preferencia por aquellas opciones más cercanas en contenido semántico a la pregunta original.

Además, se realizó un análisis cualitativo de los casos en los que el sistema no logró identificar la respuesta correcta, centrándose en las características de las respuestas seleccionadas erróneamente. Este análisis permitió identificar patrones recurrentes, como la presencia de entidades o términos comunes con la pregunta, o similitudes sintácticas con la opción correcta, lo cual sugiere cierta sensibilidad del sistema a elementos superficiales del lenguaje.

Los resultados obtenidos muestran que el sistema RAG tiene buena capacidad para distinguir entre opciones engañosas lo que sugiere que pueda ser capaz de obtener mejor rendimiento que las estrategias heurísticas simples

Cuadro 3.2: Análisis del comportamiento del sistema RAG frente a opciones incorrectas pero plausibles en MovieQA.

| Película | Razón de aciertos | Seleccionó opción más larga | Seleccionó opción más similar | Errores por entidades compartidas | Errores por similitud sintáctica |
|---|---|---|---|---|---|
| Jurassic Park III | 55 | 18 | 48 | 40 | 32 |
| An Education | 58 | 15 | 50 | 38 | 30 |
| Bridget Jones's Diary | 57 | 17 | 49 | 39 | 31 |
| Bad Teacher | 53 | 21 | 46 | 42 | 35 |
| Fargo | 56 | 16 | 50 | 37 | 29 |
| **Promedio** | **55.8** | **17.4** | **48.6** | **39.2** | **31.4** |

```
================================================================================
Pregunta: What can a 3D printer do, according to Billy in Jurassic Park III?
Opciones: Replicate the foot of any dinosaur. Replicate the larynx of a Velocira
ptor. Replicate a T-Rex's tail. Replicate the spine of a Velociraptor. Replicate
 ancient plants that were food for the dinosaurs.
Respuesta: The correct answer is: Replicate the larynx of a Velociraptor.
Confianza del sistema: 0.78
Pasajes procesados: 18
Contradicciones resueltas: 21
================================================================================
```

Figura 3.7: Respuesta del sistema RAG con opciones de respuesta

```
================================================================================
Pregunta: Who is Mark in Jurassic Park III?
Opciones: Ellie's uncle. Sattler's friend. Cheryl's husband. Amanda's husband. E
llie's husband.
Respuesta: The correct answer is: Ellie's husband.
Confianza del sistema: 0.83
Pasajes procesados: 18
Contradicciones resueltas: 21
================================================================================
```

Figura 3.8: Respuesta del sistema RAG con opciones de respuesta

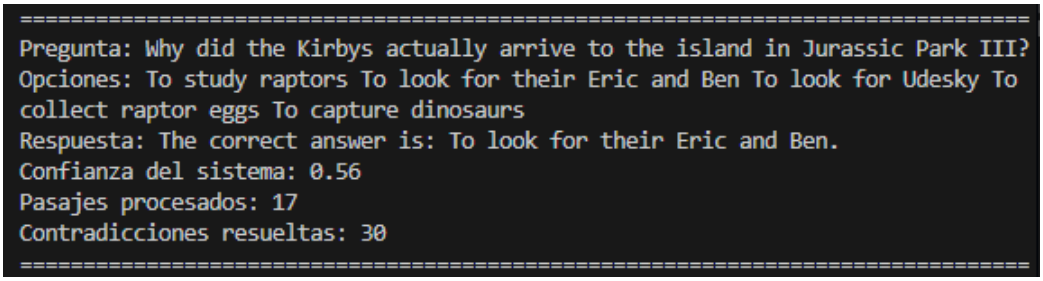


Figura 3.9: Respuesta del sistema RAG con opciones de respuesta

analizadas. No obstante, se observó que en algunos casos fallidos, el sistema seleccionó respuestas que compartían elementos clave con la pregunta, como nombres propios, acciones o estructuras sintácticas similares, aunque contenían errores sutiles o información parcialmente incorrecta. Esto indica que el sistema es capaz de capturar relaciones semánticas relevantes, pero aún puede verse afectado por similitudes superficiales o ambiguas entre las opciones.

Estos hallazgos sugieren que, aunque el sistema tiene un buen nivel de comprensión contextual, existen oportunidades de mejora en la discriminación de respuestas muy similares desde el punto de vista léxico o semántico, especialmente cuando estas contienen información factual incorrecta pero plausiblemente relacionada.

### 3.3.3. Análisis del Impacto del Módulo de Recuperación en Sistemas RAG para Preguntas-Respuestas

Como parte fundamental de la investigación sobre sistemas RAG aplicados al dominio cinematográfico, se ha desarrollado un experimento para comprobar el impacto relativo de los componentes de recuperación de información y generación de respuestas. Este estudio comparativo busca determinar cómo cada módulo contribuye al desempeño global del sistema cuando se enfrenta a preguntas complejas sobre tramas de películas, un dominio que plantea desafíos únicos debido a la naturaleza narrativa y contextual de los contenidos.

El diseño experimental se estructura en dos fases claramente diferenciadas. En primer lugar, comprobamos el módulo de recuperación operando de forma independiente, donde el sistema se limita a recuperar y seleccionar

los pasajes más relevantes del corpus cinematográfico sin intervención del componente generativo. Este enfoque nos permite establecer una línea base que refleja la capacidad pura de recuperación de información del sistema. En contraste, la segunda fase comprueba el sistema RAG completo, donde los pasajes recuperados se envían a un modelo generativo que sintetiza respuestas naturales, combinando la información recuperada con su conocimiento intrínseco.

Para garantizar la validez de los hallazgos, se implementaron rigurosos controles experimentales. Se utilizaron consistentemente el mismo conjunto de preguntas del *dataset* de MovieQA en ambas configuraciones, se mantuvo idénticos parámetros para el módulo de recuperación en todos los casos. Estas precauciones metodológicas permiten aislar el efecto de la incorporación del módulo generativo y realizar comparaciones directas y significativas entre ambos enfoques.

Cuadro 3.3: Comparación del desempeño entre el módulo de recuperación pura y el sistema RAG completo.

| **Película** | **Recuperación pura** | **Sistema RAG** | **Aumento con generación** |
|---|---|---|---|
| Jurassic Park III | 43 | 55 | +12 |
| An Education | 46 | 58 | +12 |
| Bridget Jones's Diary | 45 | 57 | +12 |
| Bad Teacher | 42 | 53 | +11 |
| Fargo | 46 | 56 | +10 |
| **Promedio** | **44.4** | **55.8** | **+11.4** |

Los resultados de este experimento multidimensional revelan valiosos insights sobre la dinámica entre los módulos de recuperación y generación en sistemas RAG aplicados al dominio cinematográfico. Uno de los hallazgos más significativos es el impacto positivo del componente generativo, cuya incorporación permite superar claramente el desempeño del sistema basado únicamente en la recuperación de información. Lejos de ser una capa meramente decorativa, la generación de respuestas demuestra ser fundamental para sintetizar, contextualizar y reformular el conocimiento extraído del corpus, adaptándolo de manera precisa a la pregunta planteada. No obstante, también se observaron casos límite en los que el modelo generativo introdujo

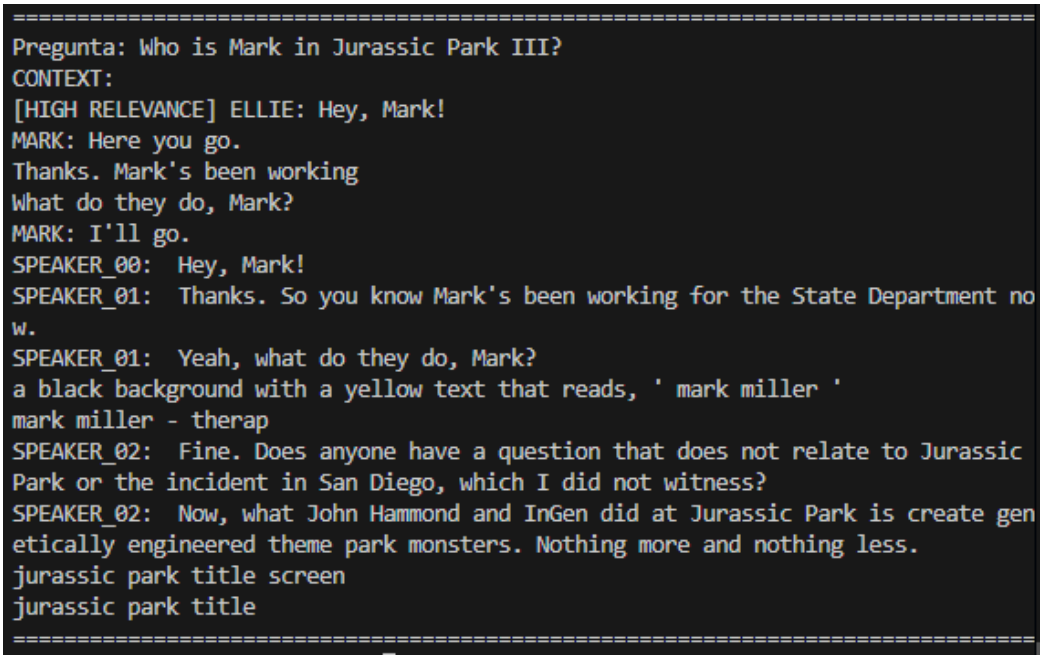

Figura 3.10: Contexto recuperado por el sistema RAG

distorsiones, inexactitudes o razonamientos fuera de contexto, lo cual indica que su contribución, aunque mayoritariamente positiva, no está exenta de riesgos. Por otro lado, aspectos cualitativos como la fluidez, coherencia y naturalidad de las salidas mejoraron notablemente respecto a las respuestas basadas únicamente en fragmentos recuperados, lo que refuerza la importancia del módulo generativo no solo en términos de precisión, sino también de experiencia del usuario.

## 3.4. Discusión

Los resultados obtenidos muestran que el sistema RAG es efectivo al momento de resolver preguntas complejas relacionadas con películas, especialmente aquellas que exigen la integración de información proveniente de distintas partes de la narrativa. Gracias a su capacidad para recuperar y generar respuestas contextualizadas, el modelo logra un desempeño considerable, con capacidad de hacer frente a enfoques basados únicamente en selección de opciones o similitud textual. No obstante, estos avances también permitieron

identificar límites importantes en su funcionamiento actual.

Una de las principales áreas de mejora se encuentra en la comprensión profunda del contenido audiovisual. Preguntas del tipo "Why" (por qué) o "How" (cómo), que requieren una interpretación causal o explicativa de los eventos narrativos, continúan representando un reto considerable para el sistema. Esto sugiere que, aunque RAG puede asociar información relevante y generar respuestas coherentes, aún le falta desarrollar un nivel más avanzado de razonamiento interpretativo. Asimismo, pese a que el sistema depende menos de estrategias heurísticas superficiales comparado con los *baselines* —como la longitud de la respuesta o su similitud léxica con la pregunta—, ciertos sesgos lingüísticos persisten y pueden influir en sus decisiones en casos ambiguos.

# Conclusiones

En este trabajo se llevó a cabo el diseño e implementación de un sistema RAG enfocado en el ámbito cinematográfico, con el propósito de ofrecer una herramienta capaz de responder consultas sobre contenido audiovisual con un mayor nivel de comprensión semántica.

El desarrollo del sistema involucró diversas etapas, entre las que destacan la construcción de una herramienta de preprocesamiento de películas y la integración funcional de cada uno de los componentes que conforman el esquema RAG.

A lo largo del proceso exploratorio, se evidenció el potencial de los sistemas RAG para el análisis y consulta de medios multimedia. Esta aproximación constituye un paso clave hacia la integración entre el lenguaje natural y la riqueza expresiva de las películas, ayudando a cerrar la brecha semántica entre ambos mundos y sentando las bases para soluciones más sofisticadas en este campo.

No obstante, es fundamental reconocer las limitaciones del trabajo realizado. A pesar de que el prototipo demuestra resultados alentadores, persisten desafíos como la escalabilidad del proceso de preprocesamiento y las dificultades para capturar de forma exhaustiva la información visual presente en los filmes.

El panorama para futuros desarrollos en sistemas RAG es amplio y prometedor. Esta investigación pretende ser un punto de partida en la construcción de herramientas más avanzadas para la consulta profunda de contenido cinematográfico, dejando abiertas múltiples oportunidades de mejora y expansión.

# Recomendaciones

Este trabajo sienta las bases para futuras investigaciones en el ámbito del análisis cinematográfico mediante sistemas de procesamiento multimodal. A partir de los hallazgos obtenidos, se identifican varias líneas prometedoras para el desarrollo subsiguiente:

• El estudio pone de relieve la necesidad de emplear herramientas de procesamiento visual más avanzadas que permitan capturar una mayor riqueza de información presente en las secuencias cinematográficas. La estrategia actual, basada en la extracción de fotogramas aislados, mostró limitaciones para representar adecuadamente aspectos narrativos y estéticos clave del cine. Para superar estas barreras, se propone explorar tecnologías capaces de analizar elementos visuales complejos —como encuadres, iluminación, simbología y otros componentes compositivos— así como incorporar métodos que tengan en cuenta la evolución temporal de los planos y secuencias, mejorando así la comprensión global del contenido audiovisual.

• Se recomienda llevar a cabo un análisis más detallado del impacto individual de cada componente dentro del *pipeline* RAG. Este tipo de evaluación permitiría identificar cuáles son los módulos con mayor incidencia en el rendimiento global, facilitando así decisiones informadas sobre posibles mejoras técnicas y optimización de recursos. En particular, resulta clave profundizar en el estudio comparativo de diferentes estrategias de recuperación y esquemas de fusión multimodal para maximizar la efectividad del sistema.

•Es fundamental considerar la ampliación del corpus en futuras investigaciones, teniendo en cuenta que el presente estudio se desarrolló bajo una metodología de caso de estudio con un conjunto reducido de muestras. Aunque esta estrategia permitió un análisis profundo y contextualizado dentro de un marco controlado, una extensión del número de muestras contribuiría a

validar los hallazgos en un espectro más amplio de géneros, estilos y estructuras narrativas cinematográficas. Este enfoque más amplio facilitaría además la identificación de patrones de error recurrentes y sesgos potenciales en el procesamiento de contenido audiovisual, fortaleciendo así la generalización y aplicabilidad del sistema en contextos reales más diversos.

• Los resultados sugieren que algunas de las limitaciones observadas podrían mitigarse mediante el uso de infraestructura computacional más avanzada. Se recomienda realizar nuevas pruebas en entornos con mayores recursos para evaluar el verdadero potencial del enfoque propuesto. Esto abriría la posibilidad de implementar configuraciones más ambiciosas en cuanto a paralelización de tareas, manejo de contextos extendidos y uso de modelos más complejos, lo cual podría traducirse en una mejora significativa del desempeño general del sistema.

En conjunto, estas recomendaciones apuntan hacia un camino claro y constructivo para desarrollar soluciones más precisas, eficientes y especializadas en el análisis automatizado de obras cinematográficas.

# Apéndice A

# Muestra

El MovieQA es un conjunto de datos diseñado para evaluar sistemas de preguntas y respuestas (QA) en el dominio cinematográfico. Contiene preguntas complejas sobre tramas de películas, formuladas a partir de guiones, subtítulos y conocimiento común sobre las obras. El *dataset* incluye distintos tipos de preguntas (por ejemplo, sobre personajes, eventos o relaciones narrativas) y está estructurado para permitir evaluaciones tanto automáticas como mediante juicio humano.

Se compone de un amplio número de preguntas emparejadas con respuestas correctas, junto con fuentes de información relevantes como fragmentos de texto (sinopsis, diálogos, etc.) que pueden usarse como contexto para encontrar las respuestas. MovieQA se ha convertido en un *benchmark* popular para medir el desempeño de modelos de comprensión lectora y sistemas RAG en contextos narrativos y semánticamente ricos como el cine.

Por problemas de capacidad de cómputo y tiempo se decidió hacer el estudio con una muestra representativa de cinco películas. Se escogieron buscando crear una selección representativa en cuanto a géneros, y priorizando la mínima duración de las mismas. El extracto escogido se compone como se presenta a continuación:

Cuadro A.1: Lista de películas

| **Película** | **Duración** | **Géneros** | **Año** |
|---|---|---|---|
| Jurassic Park III | 1:32:20 | Acción, Aventura, Ciencia ficción, Thriller | 1993 |
| An Education | 1:35:57 | Drama | 2009 |
| Bridget Jones's Diary | 1:36:00 | Comedia, Drama, Romance | 2001 |
| Bad Teacher | 1:36:27 | Comedia | 2011 |
| Fargo | 1:38:11 | Crimen, Drama, Thriller | 1996 |

# Referencias bibliográficas


[1] Ahlawat, Harsh, Aggarwal, Naveen y Gupta, Deepti. "Automatic Speech Recognition: A survey of deep learning techniques and approaches". En: *International Journal of Cognitive Computing in Engineering* 6 (2025), págs. 201-237. ISSN: 2666-3074. DOI: https://doi.org/10.1016/j.ijcce.2024.12.007. URL: https://www.sciencedirect.com/science/article/pii/S2666307424000573.

[2] Alayrac, Jean-Baptiste et al. *Flamingo: a Visual Language Model for Few-Shot Learning*. 2022. arXiv: 2204.14198 [cs.CV]. URL: https://arxiv.org/abs/2204.14198.

[3] An, Yuwei, Cheng, Yihua, Park, Seo Jin y Jiang, Junchen. "HyperRAG: Enhancing Quality-Efficiency Tradeoffs in Retrieval-Augmented Generation with Reranker KV-Cache Reuse". En: (2025). arXiv: 2504.02921 [cs.CL]. URL: https://arxiv.org/abs/2504.02921.

[4] Arefeen, Md Adnan, Debnath, Biplob, Uddin, Md Yusuf Sarwar y Chakradhar, Srimat. "iRAG: Advancing RAG for Videos with an Incremental Approach". En: *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*. CIKM '24. ACM, oct. de 2024, págs. 4341-4348. DOI: 10.1145/3627673.3680088. URL: http://dx.doi.org/10.1145/3627673.3680088.

[5] Arshad, M., Onn, C. W., Ahmad, A. y Mogwe, G. "Big data analytics and AI as success factors for online video streaming platforms". En: *Frontiers in Big Data* 8 (2025), pág. 1513027. DOI: 10.3389/fdata.2025.1513027.

[6] Borgeaud, Sebastian et al. "Improving language models by retrieving from trillions of tokens". En: (2022). arXiv: 2112.04426 [cs.CL]. URL: https://arxiv.org/abs/2112.04426.

[7] Caspari, Laura, Dastidar, Kanishka Ghosh, Zerhoudi, Saber, Mitrović, Jelena y Granitzer, Michael. "Beyond Benchmarks: Evaluating Embedding Model Similarity for Retrieval Augmented Generation Systems". Ver. 1. En: *arXiv preprint* (11 de jul. de 2024). License: CC BY 4.0. arXiv: 2407.08275 [cs.IR]. URL: https://arxiv.org/abs/2407.08275.

[8] Chen, Haoran, Li, Jianmin, Frintrop, Simone y Hu, Xiaolin. "The MSR-Video to Text Dataset with Clean Annotations". En: *arXiv preprint arXiv:2102.06448v4* (feb. de 2024). Accessed: 25 Feb 2024. DOI: 10.48550/arXiv.2102.06448.

[9] Chen, Wenhu, Hu, Hexiang, Chen, Xi, Verga, Pat y Cohen, William W. *MuRAG: Multimodal Retrieval-Augmented Generator for Open Question Answering over Images and Text*. 2022. arXiv: 2210.02928 [cs.CL]. URL: https://arxiv.org/abs/2210.02928.

[10] Cheng, Jeffrey, Marone, Marc, Weller, Orion, Lawrie, Dawn, Khashabi, Daniel y Durme, Benjamin Van. "Dated Data: Tracing Knowledge Cutoffs in Large Language Models". En: (2024). URL: https://openreview.net/forum?id=wS7PxDjy6m.

[11] Choi, Seongho et al. "DramaQA: Character-Centered Video Story Understanding with Hierarchical QA". En: (2020). arXiv: 2005.03356 [cs.CL]. URL: https://arxiv.org/abs/2005.03356.

[12] Dai, Wenliang et al. *InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning*. 2023. arXiv: 2305.06500 [cs.CV]. URL: https://arxiv.org/abs/2305.06500.

[13] DeepSeek-AI et al. *DeepSeek-V3 Technical Report*. 2025. arXiv: 2412.19437 [cs.CL]. URL: https://arxiv.org/abs/2412.19437.

[14] Devlin, Jacob, Chang, Ming-Wei, Lee, Kenton y Toutanova, Kristina. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. 2019. arXiv: 1810.04805 [cs.CL]. URL: https://arxiv.org/abs/1810.04805.

[15] Edge, Darren et al. *From Local to Global: A Graph RAG Approach to Query-Focused Summarization*. 2025. arXiv: 2404.16130 [cs.CL]. URL: https://arxiv.org/abs/2404.16130.

[16] Fan, Yue et al. *VideoAgent: A Memory-augmented Multimodal Agent for Video Understanding*. 2024. arXiv: 2403.11481 [cs.CV]. URL: https://arxiv.org/abs/2403.11481.

[17] Faysse, Manuel et al. *ColPali: Efficient Document Retrieval with Vision Language Models*. 2025. arXiv: 2407.01449 [cs.IR]. URL: https://arxiv.org/abs/2407.01449.

[18] Guo, Ruifeng et al. *A Survey on Image-text Multimodal Models*. 2024. arXiv: 2309.15857 [cs.CL]. URL: https://arxiv.org/abs/2309.15857.

[19] Guo, Yucan et al. *Retrieval-Augmented Code Generation for Universal Information Extraction*. 2023. arXiv: 2311.02962 [cs.AI]. URL: https://arxiv.org/abs/2311.02962.

[20] Guu, Kelvin, Lee, Kenton, Tung, Zora, Pasupat, Panupong y Chang, Ming-Wei. "REALM: Retrieval-Augmented Language Model Pre-Training". En: (2020). arXiv: 2002.08909 [cs.CL]. URL: https://arxiv.org/abs/2002.08909.

[21] He, Yingqing et al. *Animate-A-Story: Storytelling with Retrieval-Augmented Video Generation*. 2023. arXiv: 2307.06940 [cs.CV]. URL: https://arxiv.org/abs/2307.06940.

[22] Izacard, Gautier y Grave, Edouard. "Leveraging Passage Retrieval with Generative Models for Open Domain Question Answering". En: (2021). arXiv: 2007.01282 [cs.CL]. URL: https://arxiv.org/abs/2007.01282.

[23] Izacard, Gautier et al. "Atlas: Few-shot Learning with Retrieval Augmented Language Models". En: (2022). arXiv: 2208.03299 [cs.CL]. URL: https://arxiv.org/abs/2208.03299.

[24] Jeong, Soyeong, Baek, Jinheon, Cho, Sukmin, Hwang, Sung Ju y Park, Jong. "Adaptive-RAG: Learning to Adapt Retrieval-Augmented Large Language Models through Question Complexity". En: *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. NAACL 2024. Mexico City, Mexico: Association for Computational Linguistics, jun. de 2024, págs. 7036-7050.

[25] Jeong, Soyeong, Kim, Kangsan, Baek, Jinheon y Hwang, Sung Ju. "VideoRAG: Retrieval-Augmented Generation over Video Corpus". En: (2025). arXiv: 2501.05874 [cs.CV]. URL: https://arxiv.org/abs/2501.05874.

[26] Jeong, Soyeong, Kim, Kangsan, Baek, Jinheon y Hwang, Sung Ju. *VideoRAG: Retrieval-Augmented Generation over Video Corpus*. 2025. arXiv: 2501.05874 [cs.CV]. URL: https://arxiv.org/abs/2501.05874.

[27] Jiang, Ting et al. *E5-V: Universal Embeddings with Multimodal Large Language Models*. 2024. arXiv: 2407.12580 [cs.CL]. URL: https://arxiv.org/abs/2407.12580.

[28] Jiang, Zhengbao et al. *Active Retrieval Augmented Generation*. 2023. arXiv: 2305.06983 [cs.CL]. URL: https://arxiv.org/abs/2305.06983.

[29] Jones, Karen Spärck. "A Statistical Interpretation of Term Specificity and Its Application in Retrieval". En: *Journal of Documentation* 60.5 (2004), págs. 493-502. DOI: 10.1108/00220410410560347.

[30] Kandpal, Nikhil, Deng, Haikang, Roberts, Adam, Wallace, Eric y Raffel, Colin. "Large Language Models Struggle to Learn Long-Tail Knowledge". En: (2023). arXiv: 2211.08411 [cs.CL]. URL: https://arxiv.org/abs/2211.08411.

[31] Karpukhin, Vladimir et al. "Dense Passage Retrieval for Open-Domain Question Answering". En: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP 2020)*. Online: Association for Computational Linguistics, nov. de 2020, págs. 6769-6781. DOI: 10.18653/v1/2020.emnlp-main.550.

[32] Lee, Jaewoo, Ko, Joonho, Baek, Jinheon, Jeong, Soyeong y Hwang, Sung Ju. *Unified Multimodal Interleaved Document Representation for Retrieval*. 2024. arXiv: 2410.02729 [cs.CL]. URL: https://arxiv.org/abs/2410.02729.

[33] Lewis, Patrick et al. "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks". En: (2021). arXiv: 2005.11401 [cs.CL]. URL: https://arxiv.org/abs/2005.11401.

[34] Li, Junnan, Li, Dongxu, Savarese, Silvio y Hoi, Steven. "BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models". En: (2023). arXiv: 2301.12597 [cs.CV]. URL: https://arxiv.org/abs/2301.12597.

[35] Li, Yanwei, Wang, Chengyao y Jia, Jiaya. *LLaMA-VID: An Image is Worth 2 Tokens in Large Language Models*. 2023. arXiv: 2311.17043 [cs.CV]. URL: https://arxiv.org/abs/2311.17043.

[36] Li, Zhiyuan, Liu, Dongnan, Zhang, Chaoyi, Wang, Heng, Xue, Tengfei y Cai, Weidong. *Enhancing Advanced Visual Reasoning Ability of Large Language Models*. 2024. arXiv: 2409.13980 [cs.CV]. URL: https://arxiv.org/abs/2409.13980.

[37] Liu, Nelson F. et al. *Lost in the Middle: How Language Models Use Long Contexts*. 2023. arXiv: 2307.03172 [cs.CL]. URL: https://arxiv.org/abs/2307.03172.

[38] Luo, Yongdong et al. *Video-RAG: Visually-aligned Retrieval-Augmented Long Video Comprehension*. 2024. arXiv: 2411.13093 [cs.CV]. URL: https://arxiv.org/abs/2411.13093.

[39] Mo, Fengran et al. "A Survey of Conversational Search". En: (2024). arXiv: 2410.15576 [cs.CL]. URL: https://arxiv.org/abs/2410.15576.

[40] OpenAI et al. *GPT-4 Technical Report*. 2024. arXiv: 2303.08774 [cs.CL]. URL: https://arxiv.org/abs/2303.08774.

[41] Pierre, Rafael. *MovieGPT: A RAG-based Movie Recommendation System using Generative AI*. https://github.com/rafaelpierre/moviegpt. Versión 0.0.1. 2023.

[42] Qasim, Iqra, Horsch, Alexander y Prasad, Dilip K. *Dense Video Captioning: A Survey of Techniques, Datasets and Evaluation Protocols*. 2023. arXiv: 2311.02538 [cs.CV]. URL: https://arxiv.org/abs/2311.02538.

[43] Radford, Alec, Kim, Jong Wook, Xu, Tao, Brockman, Greg, McLeavey, Christine y Sutskever, Ilya. "Robust Speech Recognition via Large-Scale Weak Supervision". En: (2022). arXiv: 2212.04356 [eess.AS]. URL: https://arxiv.org/abs/2212.04356.

[44] Radford, Alec et al. *Learning Transferable Visual Models From Natural Language Supervision*. 2021. arXiv: 2103.00020 [cs.CV]. URL: https://arxiv.org/abs/2103.00020.

[45] Robertson, Stephen y Zaragoza, Hugo. "The Probabilistic Relevance Framework: BM25 and Beyond". En: *Foundations and Trends in Information Retrieval*. Vol. 3. 4. Now Publishers Inc., 2009, págs. 333-389. DOI: 10.1561/1500000017.

[46] Robertson, Stephen E., Walker, Steve, Jones, Susan, Hancock-Beaulieu, Micheline y Gatford, Mike. "Okapi at TREC-3". En: *Proceedings of The Third Text REtrieval Conference (TREC 1994)*. Vol. 500-225. NIST Special Publication. Gaithersburg, Maryland, USA: National Institute of Standards y Technology (NIST), nov. de 1994, págs. 109-126.

[47] Ronchini, Francesca y Serizel, Romain. *Performance and energy balance: a comprehensive study of state-of-the-art sound event detection systems*. 2024. arXiv: 2310.03455 [eess.AS]. URL: https://arxiv.org/abs/2310.03455.

[48] Shao, Zhihong, Gong, Yeyun, Shen, Yelong, Huang, Minlie, Duan, Nan y Chen, Weizhu. *Enhancing Retrieval-Augmented Large Language Models with Iterative Retrieval-Generation Synergy*. 2023. arXiv: 2305.15294 [cs.CL]. URL: https://arxiv.org/abs/2305.15294.

[49] Shi, Yunxiao, Zi, Xing, Shi, Zijing, Zhang, Haimin, Wu, Qiang y Xu, Min. *Enhancing Retrieval and Managing Retrieval: A Four-Module Synergy for Improved Quality and Efficiency in RAG Systems*. 2024. arXiv: 2407.10670 [cs.CL]. URL: https://arxiv.org/abs/2407.10670.

[50] Tapaswi, Makarand, Zhu, Yukun, Stiefelhagen, Rainer, Torralba, Antonio, Urtasun, Raquel y Fidler, Sanja. *MovieQA: Understanding Stories in Movies through Question-Answering*. 2016. arXiv: 1512.02902 [cs.CV]. URL: https://arxiv.org/abs/1512.02902.

[51] Team, Gemini et al. *Gemini: A Family of Highly Capable Multimodal Models*. 2025. arXiv: 2312.11805 [cs.CL]. URL: https://arxiv.org/abs/2312.11805.

[52] Tevissen, Yannis, Guetari, Khalil y Petitpont, Frédéric. *Towards Retrieval Augmented Generation over Large Video Libraries*. 2024. arXiv: 2406.14938 [cs.CL]. URL: https://arxiv.org/abs/2406.14938.

[53] The Movie Database (TMDB). *General FAQ - The Movie Database (TMDB)*. Accedido el 13 de junio de 2025. 2025. URL: https://www.themoviedb.org/faq/general.

[54] Tsalyk, Markiian. *Retrieval Augmented Generation Based System for Explainable Movie Search*. Bachelor Thesis. Lviv, 2024.

[55] Tufino, Eugenio. *NotebookLM: An LLM with RAG for active learning and collaborative tutoring*. 2025. arXiv: 2504.09720 [physics.ed-ph]. URL: https://arxiv.org/abs/2504.09720.

[56] Vaswani, Ashish et al. *Attention Is All You Need*. 2023. arXiv: 1706.03762 [cs.CL]. URL: https://arxiv.org/abs/1706.03762.

[57] Ventura, Lucas, Yang, Antoine, Schmid, Cordelia y Varol, Gül. "CoVR-2: Automatic Data Construction for Composed Video Retrieval". En: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 46.12 (dic. de 2024), págs. 11409-11421. ISSN: 1939-3539. DOI: 10.1109/tpami.2024.3463799. URL: http://dx.doi.org/10.1109/TPAMI.2024.3463799.

[58] Wang, Yi et al. *InternVideo2: Scaling Foundation Models for Multimodal Video Understanding*. 2024. arXiv: 2403.15377 [cs.CV]. URL: https://arxiv.org/abs/2403.15377.

[59] Wang, Yuhao, Ren, Ruiyang, Li, Junyi, Zhao, Wayne Xin, Liu, Jing y Wen, Ji-Rong. *REAR: A Relevance-Aware Retrieval-Augmented Framework for Open-Domain Question Answering*. 2024. arXiv: 2402.17497 [cs.CL]. URL: https://arxiv.org/abs/2402.17497.

[60] Xiao, Linhui, Yang, Xiaoshan, Lan, Xiangyuan, Wang, Yaowei y Xu, Changsheng. *Towards Visual Grounding: A Survey*. 2024. arXiv: 2412.20206 [cs.CV]. URL: https://arxiv.org/abs/2412.20206.

[61] Xu, Fangyuan, Shi, Weijia y Choi, Eunsol. *RECOMP: Improving Retrieval-Augmented LMs with Compression and Selective Augmentation*. 2023. arXiv: 2310.04408 [cs.CL]. URL: https://arxiv.org/abs/2310.04408.

[62] Yang, Yueting, Zhang, Xintong, Xu, Jinan y Han, Wenjuan. "Empowering Vision-Language Models for Reasoning Ability through Large Language Models". En: *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2024, págs. 10056-10060. DOI: 10.1109/ICASSP48485.2024.10446407.

[63] Yu, Shi et al. *VisRAG: Vision-based Retrieval-augmented Generation on Multi-modality Documents*. 2025. arXiv: 2410.10594 [cs.IR]. URL: https://arxiv.org/abs/2410.10594.

[64] Yu, Zhou et al. *ActivityNet-QA: A Dataset for Understanding Complex Web Videos via Question Answering*. 2019. arXiv: 1906.02467 [cs.CV]. URL: https://arxiv.org/abs/1906.02467.

[65] Yuan, Yi, Liu, Haohe, Liu, Xubo, Huang, Qiushi, Plumbley, Mark D. y Wang, Wenwu. *Retrieval-Augmented Text-to-Audio Generation*. 2024. arXiv: 2309.08051 [cs.SD]. URL: https://arxiv.org/abs/2309.08051.

[66] Zhang, Lu, Zhao, Tiancheng, Ying, Heting, Ma, Yibo y Lee, Kyusong. "OmAgent: A Multi-modal Agent Framework for Complex Video Understanding with Task Divide-and-Conquer". En: *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Ed. por Al-Onaizan, Yaser, Bansal, Mohit y Chen, Yun-Nung. Miami, Florida, USA: Association for Computational Linguistics, nov. de 2024, págs. 10031-10045. DOI: 10.18653/v1/2024.emnlp-main.559. URL: https://aclanthology.org/2024.emnlp-main.559/.

[67] Zhao, W. X., Liu, J., Ren, R. y Wen, J.-R. "Dense Text Retrieval Based on Pretrained Language Models: A Survey". En: *ACM Transactions on Information Systems* 42.4 (2024), págs. 1-60. DOI: 10.1145/3663945.

[68] Zhao, Wayne Xin et al. *A Survey of Large Language Models*. 2025. arXiv: 2303.18223 [cs.CL]. URL: https://arxiv.org/abs/2303.18223.