Street Art & Urban Creativity | Vol. 11, No. 7, 2025 | ISSN 2183-9956
Street Art & Urban Creativity. Scientific Journal
https://doi.org/10.62161/sauc.v11.6110

RECONOCIMIENTO DE OBJETOS E INTELIGENCIA ARTIFICIAL CONVERSACIONAL EN CONTEXTOS DEL MUNDO REAL:

Mejorar la experiencia en los museos mediante sistemas interactivos

ADRIÁN ORTIZ RAMÍREZ[1], ÁLVARO ILLANA SÁNCHEZ[1], MARTA SALAS GARCÍA[1]

[1] UNIVERSIDAD FRANCISCO DE VITORIA, ESPAÑA

PALABRAS CLAVE

Inteligencia artificial
IA generativa
Patrimonio cultural
Museos
Detección de objetos
Recuperación de información
Conciencia del contexto
RAG

RESUMEN

Este proyecto busca mejorar la experiencia museística de los visitantes, superando los métodos tradicionales de acceso a la información. Presenta un sistema interactivo que combina la detección de objetos en tiempo real con la generación aumentada por recuperación (Retrieval Augmented Generation) para ofrecer una guía conversacional inmersiva, personalizada y sensible al contexto.

Los resultados evidencian una comprensión espacial y conversacional precisa, así como una mejora significativa en la veracidad y la relevancia de las respuestas generadas frente a las de un LLM estándar. Este proyecto demuestra el potencial del sistema para ofrecer un acceso dinámico y atractivo al patrimonio cultural.

RECIBIDO: 31 / 08 / 2025
ACEPTADO: 08 / 10 / 2025

1. Introducción

El comienzo del siglo XXI se ha caracterizado por avances tecnológicos sin precedentes, impulsados por el crecimiento de la inteligencia artificial (IA) en un mundo cada vez más digitalizado. Estas tecnologías son más que simples herramientas. Están transformando la forma en que las personas interactúan con la información, el aprendizaje y la forma en que experimentan el mundo que les rodea. Este estudio surge de la fascinación por ese potencial transformador y explora específicamente cómo la IA puede revolucionar sectores profundamente arraigados en la tradición, pero ávidos de innovación, como el sector del patrimonio cultural. Este trabajo investiga el diseño, el desarrollo y la validación de un novedoso sistema que aprovecha las soluciones modernas de IA para servir como guía personal e inteligente que enriquece la experiencia de los visitantes de los museos.

1.1. El problema: la experiencia estática del museo

Desde la creación del primer museo conocido, que data de alrededor del año 530 a. C., estas instituciones se han dedicado a la preservación y difusión del conocimiento y la cultura. Los museos, como custodios de la historia y la cultura, albergan inmensos repositorios de conocimiento. Sin embargo, el sector del patrimonio cultural se enfrenta al reto de hacer que ese conocimiento sea accesible y atractivo para el público.

Históricamente, los métodos para comunicar este conocimiento han sido fundamentalmente estáticos. Las exposiciones tradicionales de los museos se basaban en herramientas de comunicación pasivas, como etiquetas de obras de arte, carteles explicativos y guías. Estos métodos, aunque de alguna manera informativos, a menudo no logran involucrar plenamente a los visitantes a nivel personal, ya que ofrecen una narrativa única que no se adapta a la curiosidad individual. Para mejorar la interactividad y la inmersión, estas instituciones adoptaron posteriormente tecnologías como las audioguías, que aparecieron por primera vez en 1952. Aunque supusieron una mejora, seguían siendo estáticas en cuanto a la información y, por lo tanto, inadaptables, lo que limitaba el ritmo y la autonomía con grabaciones estáticas. Ahora, en el siglo XXI, han comenzado a aparecer soluciones digitales como las pantallas interactivas y la realidad aumentada (RA). Aunque son más atractivas, interactivas e inmersivas, estas soluciones siguen teniendo dificultades para adaptarse a los intereses individuales, ya que a menudo siguen narrativas estáticas que limitan la exploración personal.

La estaticidad informativa que se encuentra en las soluciones actuales y tradicionales es lo que el presente trabajo pretende superar mediante un enfoque más dinámico y personalizado.

1.2. Tendencias del sector: en busca de una personalización eficaz

La industria del turismo es un importante motor económico en todo el mundo. Europa, por ejemplo, recibió el 51,7 % del turismo internacional total en 2024, lo que supuso aproximadamente el 10 % del PIB medio de los países europeos, con países como España y Croacia muy por encima de esta media. Una tendencia notable de esta industria ha sido la aparición del Turismo 4.0, que se centra en liberar el potencial de la innovación para crear experiencias turísticas enriquecedoras. Como impulsora autodefinida de los Objetivos de Desarrollo Sostenible (ODS), la organización Turismo 4.0 respalda la necesidad de esta innovación mediante sus informes recurrentes.

Como principales actores del turismo cultural, los museos informan anualmente sobre las tendencias de innovación que reflejan este cambio. Según el Barómetro de Innovación en Museos 2021 de Turismo 4.0, el 80 % de los museos consideraban importantes las nuevas tecnologías, y el 72 % de sus iniciativas de inteligencia de datos tenían como objetivo mejorar la experiencia de los visitantes. La adopción de la inteligencia artificial aumentó del 3 % al 14 %, el uso de audioguías del 5 % al 31 % y el de las aplicaciones móviles y web del 49 % al 70 %, lo que pone de manifiesto un claro cambio hacia la personalización y la interactividad. No obstante, desde el punto de vista de los visitantes, los estudios muestran una tendencia preocupante: la interacción

con las obras de arte sigue siendo breve, con una media de 27-29 segundos de interés por pieza en las últimas dos décadas (Smith et al., 2017). Estos patrones ponen de relieve la creciente necesidad de experiencias personalizadas e interactivas que vayan más allá de la interacción tradicional con los museos. Por lo tanto, a partir de estas tendencias, hemos desarrollado una prueba de concepto (PoC) para comprobar cómo la IA puede ofrecer interacciones adaptativas y personalizadas a los visitantes de museos.

1.3. Escenario y alcance: una prueba de concepto en el Louvre

El PoC presentado en este artículo tiene lugar en el Departamento de Antigüedades Griegas, Etruscas y Romanas del Museo del Louvre, más concretamente en unas cuantas esculturas ubicadas en dos salas: la «Salle des Caryatides» y la «Salle de la Vénus de Milo», como se ilustra en la Figura 1. Al ser el museo más visitado del mundo y líder mundial en escala e innovación, el Museo del Louvre se convirtió naturalmente en el escenario clave para nuestro PoC. En concreto, la base de datos del Louvre, disponible públicamente y con más de 500.000 obras de arte, convirtió a este museo en el escenario ideal para un proyecto basado en la IA.

Figura 1. Esculturas de la Salle des Caryatides y la Salle de la Vénus de Milo utilizadas en la prueba de concepto.

Fuente: Museo del Louvre, 2025

Este trabajo se enmarca en los ámbitos de la educación, la tecnología y la cultura, y se centra específicamente en la innovación en las experiencias museísticas. Está alineado con los Objetivos de Desarrollo Sostenible de la Agenda 2030 (Asamblea General de las Naciones Unidas, 2015), concretamente con el objetivo 4.7, centrado en la educación mediante la promoción de la cultura.

En cuanto al alcance técnico, este trabajo se centra en el diseño, desarrollo y validación de un sistema de guía museística conversacional basado en inteligencia artificial. El proceso abarca la investigación industrial y tecnológica, la planificación del sistema y la arquitectura, el entrenamiento de los modelos de prueba de concepto y la validación del rendimiento.

1.4. Objetivos: de la visión a los objetivos medibles

Para alcanzar el objetivo general de este proyecto, el trabajo se centra en varios objetivos interrelacionados que abordan la personalización, la calidad de la información y la gestión del conocimiento. Estos objetivos desglosan el propósito general del proyecto en aspectos específicos y, lo que es más importante, medibles y verificables.

Uno de los principales objetivos del proyecto es permitir la recopilación en tiempo real de contexto individualizado. Dado que los seres humanos se basan principalmente en la vista, el sistema permitirá identificar las obras de arte cercanas mediante modelos de detección de objetos entrenados con conjuntos de datos específicos del museo. Además, incorporará información auxiliar, como la ubicación de los visitantes, las interacciones previas y otras preferencias, en la ecuación contextual. Igualmente, es importante la capacidad del sistema para generar respuestas precisas y contextualmente relevantes. Al aprovechar los conjuntos de datos gestionados por los museos mediante un proceso de generación aumentada por recuperación, el sistema puede ofrecer respuestas significativas y basadas en hechos, lo que reduce el riesgo de confabulaciones o alucinaciones. Por último, el proyecto busca establecer un proceso dinámico de integración de conocimientos para mantener la precisión del sistema y su relevancia a lo largo del tiempo. Dado

que los museos actualizan constantemente sus colecciones, es esencial diseñar un proceso automatizado para la ingesta, el preprocesamiento y la indexación de la nueva información, garantizando que el sistema se mantenga actualizado, coherente y fiable.

La eficacia y la fiabilidad del sistema se validarán mediante métricas cuantificables, como la precisión del reconocimiento visual, la relevancia de la información recuperada y la coherencia del canal de integración de conocimientos.

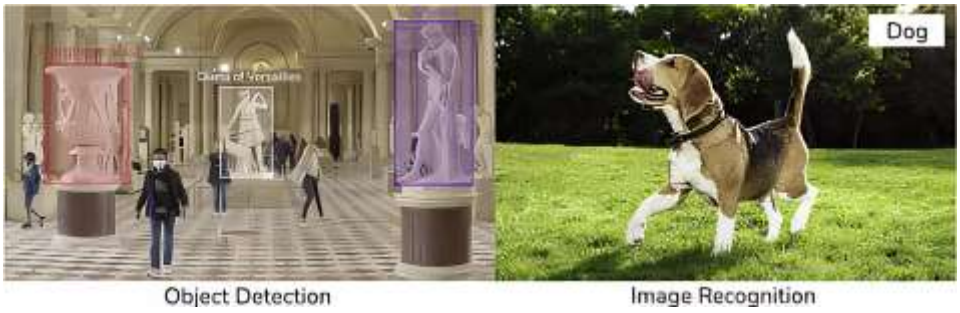
2. Estado actual

La recopilación de contexto, la recuperación de información, la generación de respuestas y la integración de conocimientos son áreas que se han estudiado y explorado ampliamente a lo largo de los años. En el contexto de los museos y el patrimonio cultural, estas tecnologías se están estudiando y aplicando para mejorar la experiencia de los visitantes.

2.1. Visión artificial

Bajo el paraguas de la visión artificial, tareas como el reconocimiento de imágenes y la detección de objetos desempeñan un papel fundamental al ayudar a las máquinas a percibir y comprender la información visual. Como se ilustra en la Figura 2, el reconocimiento de imágenes asigna una única etiqueta a toda la imagen, mientras que la detección de objetos identifica y localiza varios objetos dentro de la misma imagen mediante cuadros delimitadores. Concretamente en los museos, el reconocimiento de imágenes se ha estudiado ampliamente para tareas de clasificación automatizada de obras de arte. La detección de objetos, por otro lado, aún no se ha implementado ampliamente en entornos de patrimonio cultural. Entre los pocos usos actuales de la detección de objetos en este sector, podemos distinguir la exploración de objetos en pinturas y el seguimiento de visitantes (Breitner y Bandung, 2024; Meyer et al., 2024). Sin embargo, en nuestro caso de uso, la detección de objetos tiene mayor potencial que el reconocimiento de imágenes.

Figura 2. Detección de objetos frente a reconocimiento de imágenes



Fuente: Elaboración propia, 2025.

Actualmente, las aplicaciones de guía de museos que incluyen visión artificial, como Ask Mona (2025) o Smartify (2025), se centran en procesar una sola imagen de una obra de arte a la vez, en lugar de permitir la detección en tiempo real de varias obras de arte simultáneamente. Aquí es precisamente donde destaca el reconocimiento de imágenes. Por el contrario, este proyecto pretende superar esas limitaciones haciendo hincapié en la detección de objetos en tiempo real, lo que permite al sistema identificar varias obras de arte a la vez en su contexto.

Para lograr la detección de objetos en tiempo real, se han evaluado las soluciones disponibles. Dado que el proyecto tiene como objetivo funcionar en tiempo real, prioriza la eficiencia por encima de la precisión SOTA. Algunas soluciones existentes incluyen You Only Look Once (YOLO) (Redmon & Farhadi, 2017), Faster Region-based Convolutional Neural Networks (Faster R-CNN) (Ren et al., 2015) y Single Shot Multibox Detectors (SSD)(Liu et al., 2016). Entre estas soluciones, YOLO supera a otras arquitecturas en eficiencia y precisión para objetos grandes, mientras que Faster R-CNN obtiene excelentes resultados en objetos más pequeños, pero con tiempos de ejecución más bajos.

2.2. Generación aumentada por recuperación

La adopción generalizada de los grandes modelos de lenguaje (LLM) subraya su importancia. A pesar de los avances, siguen siendo propensos a las alucinaciones debido a su dependencia de la memoria paramétrica, lo que los hace poco fiables en ámbitos especializados. En los museos, por ejemplo, los resultados de los LLM suelen mostrar desajustes culturales, lo que puede dar lugar a interpretaciones erróneas o distorsionadas de las obras de arte. Estudios recientes indican que estas discrepancias pueden alcanzar hasta el 65 % en el ámbito del patrimonio cultural (Bu et al., 2025). Para abordar este problema, las técnicas de ingeniería de prompts han comenzado a captar la atención. Una solución emergente es la generación aumentada por recuperación (RAG), una de las técnicas más estudiadas. La RAG, introducida en (Sahoo et al., 2024), integra la memoria paramétrica y no paramétrica para basar las respuestas en hechos recuperados, y su eficacia en los museos se ha demostrado en (Loffredo y De Santo, 2024; Vastakas, 2024). Su adopción se observa incluso en sectores críticos como la medicina y la seguridad, en proyectos como los de Du et al. (2024) y Wu et al. (2024).

Más allá de reducir las alucinaciones, el RAG ha demostrado ser especialmente valioso, ya que permite mantener actualizado el conocimiento al incorporar nueva información. Este es un aspecto crucial, ya que los LLM, por sí solos, solo pueden generar información hasta el límite de su entrenamiento. En los museos, donde los catálogos, registros y otros archivos curatoriales se actualizan constantemente, el RAG puede garantizar que la información permanezca precisa y relevante.

2.3. Soluciones actuales

En el estado actual de la técnica, Ask Mona destaca como la solución más completa. Su aplicación móvil personaliza la experiencia de los visitantes al combinar la IA conversacional con el reconocimiento de obras de arte, lo que les permite escanearlas y recibir respuestas contextuales y personalizadas. El sistema integra contenidos de algunos de los museos más famosos del mundo y ya cuenta con la confianza de más de 150 organizaciones a nivel mundial. Sin embargo, la función de reconocimiento de imágenes limita la cantidad de contexto que el sistema puede capturar, lo que obliga a los usuarios a recurrir a la aplicación con frecuencia al explorar nuevas obras de arte. Nuestra prueba de concepto aborda esta limitación incorporando la funcionalidad de transmisión de contexto y sustituyendo el reconocimiento de imágenes por la detección de objetos, lo que permite al sistema adaptarse de forma autónoma al contexto cambiante de los visitantes en tiempo real.

Otras soluciones, como Smartify (2025) y Nubart (2025), abordan aspectos de nuestros requisitos para una experiencia personalizada y sensible al contexto. Por ejemplo, Smartify ofrece reconocimiento de imágenes y Nubart proporciona escaneo de códigos QR, pero ambos carecen de personalización conversacional y siguen dependiendo de los sistemas de audioguía convencionales.

2.4. Propuesta de valor diferencial

Este estudio se diferencia por integrar la detección de objetos en tiempo real con la recuperación de información contextual, ofreciendo una experiencia museística inmersiva y personalizada mediante medios visuales y conversacionales. Si bien algunas soluciones existentes proporcionan información general sobre las obras de arte, a menudo carecen de conciencia contextual y de adaptabilidad a la curiosidad. Al combinar la detección de objetos en tiempo real con la generación aumentada por recuperación, nuestra solución puede recuperar dinámicamente documentos relevantes de una base de datos vectorial, considerando no solo la consulta del visitante, sino también su entorno. Si bien la tecnología RAG ha demostrado su relevancia en sectores como la medicina y la seguridad, su aplicación puede adaptarse a otros campos en los que el acceso a información precisa resulta esencial. La falta de uso documentado en espacios culturales, donde la información es crucial, sugiere una posible necesidad de sus beneficios. Este proyecto innova en la forma de interactuar con las colecciones de los museos, ofreciendo a los visitantes una guía

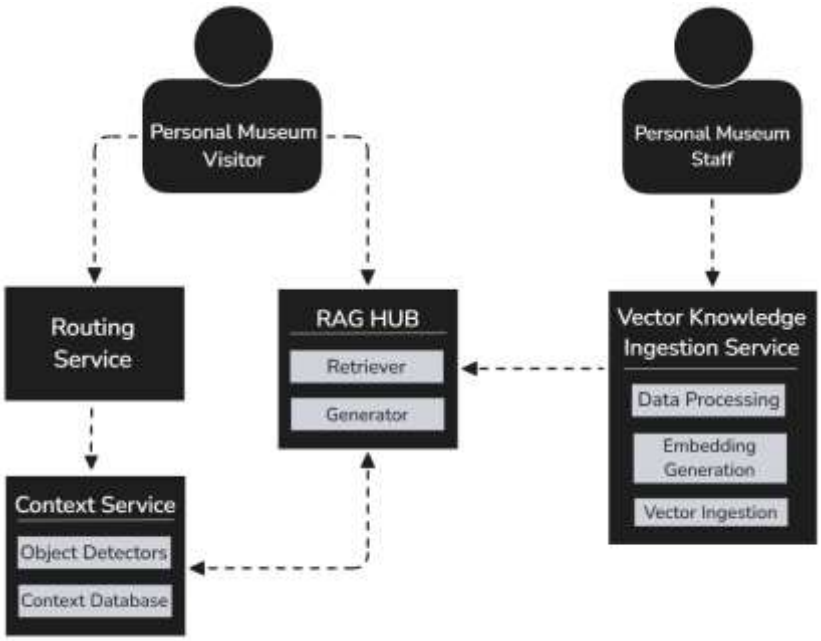
inteligente y dinámica capaz de responder a sus preguntas de forma similar a como lo haría un humano, considerando su entorno en tiempo real.

El verdadero valor de este proyecto reside en su capacidad para transformar la forma en que las personas interactúan con la cultura y la historia. Este enfoque permite a cada usuario desempeñar un papel activo en su propio proceso de aprendizaje, adaptándose a su ritmo y a sus antecedentes. En esencia, este proyecto no se limita a la tecnología, sino que utiliza la innovación para conectar a las personas con la cultura, preservar el patrimonio cultural y hacer que el conocimiento sea más atractivo y fácilmente accesible.

3. Diseño del sistema

El diseño del sistema se ha organizado en una arquitectura cliente-servidor modular y escalable que conecta la entrada visual en tiempo real con la recuperación de conocimientos y el razonamiento del modelo lingüístico, proporcionando respuestas contextuales al usuario. La Figura 3 presenta un esquema de alto nivel que destaca los módulos principales y las relaciones entre ellos.

Figura 3. Arquitectura de alto nivel del sistema.



Fuente: Elaboración propia, 2025.

El *servicio de enrutamiento* recibe el flujo de vídeo y la ubicación del visitante, y luego aplica un enrutamiento basado en la ubicación para reenviar el flujo de vídeo al *detector de objetos* entrenado para la sala correspondiente. *El servicio de contexto* procesa las detecciones, las vincula con la sesión en curso y con el historial de conversaciones, y las almacena en la *base de datos de contexto,* poniendo este contexto a disposición de otros módulos y registrando las respuestas finales para garantizar la continuidad. Paralelamente, el *servicio de ingestión de conocimiento vectorial,* mantenido y alimentado por el personal del museo recopila y limpia las fuentes de la colección, las divide en fragmentos semánticamente coherentes, ejecuta la generación de incrustaciones y completa la ingestión vectorial en la base de datos vectorial. En el momento de la consulta, el *RAG Hub* utiliza la pregunta del visitante junto con las detecciones recientes y el historial para recuperar y reclasificar los documentos relevantes del almacén vectorial; luego, su *generador* produce respuestas fundamentadas que se devuelven al cliente y se vuelven a escribir en el *servicio de contexto*.

Este flujo permite ofrecer una orientación en tiempo real, sensible a la vista y a la sesión, al tiempo que sigue siendo escalable para múltiples visitantes simultáneos.

3.1. Servicio de enrutamiento

Para una comunicación en tiempo real y con baja latencia, el *servicio de enrutamiento* divide la entrada del visitante en pequeños paquetes y los organiza en colas, lo que permite un procesamiento por hilos eficiente.

El flujo contextual del visitante consiste principalmente en fotogramas de vídeo de lo que ve en tiempo real y en la ubicación física actual del visitante dentro del museo. Una vez que el cliente comienza a enviar paquetes de información contextual, el servicio enruta dinámicamente sus fotogramas de vídeo al modelo de detección de objetos adecuado, entrenado específicamente para la sala donde se encuentra el visitante.

El diseño, basado en colas y en el enrutamiento sensible a la ubicación, proporciona una alta escalabilidad y admite múltiples clientes simultáneamente, al tiempo que mantiene tiempos de respuesta bajos. El diseño simplifica la distribución de las cargas de trabajo al adaptarse a distintos espacios físicos.

3.2. Servicio de contexto

El servicio de contexto se encarga de procesar la información visual de los visitantes, generar predicciones sobre las obras de arte que ven y almacenarla junto con el historial de conversaciones, de modo que pueda reutilizarse durante la recuperación y la generación. Su principal funcionalidad es gestionar la información contextual del visitante relevante para su visita actual, de modo que las respuestas finales del LLM sean lo más precisas posible respecto de dicho contexto. Este servicio se organiza en tres componentes principales.

3.2.1. Base de datos contextual

La base de datos contextual almacena información relevante de la sesión del visitante. Para cada visitante, se conserva un identificador único y se vincula a las predicciones generadas, al historial de consultas anteriores del usuario y a las respuestas del sistema, junto con sus respectivas colas. Este diseño relacional garantiza la coherencia entre las entidades, mantiene la información estrictamente vinculada a la sesión y permite reutilizar los datos almacenados en los procesos de recuperación y generación.

3.2.2. Gestor de contexto

El gestor de contexto se utiliza para garantizar que toda la información contextual relevante esté disponible de forma coherente durante la sesión del visitante, y se elimina de forma segura al finalizar la sesión para proteger la privacidad del visitante. Almacena las predicciones del detector de objetos y los intercambios conversacionales, lo que permite al canal RAG considerar lo que el visitante ha visto recientemente y ofrecer respuestas más precisas y relevantes. Al mismo tiempo, registra los intercambios conversacionales de los visitantes para que los diálogos puedan continuar de forma natural sin perder el hilo de los mensajes anteriores. La centralización de estas tareas en un único servicio evita que los microservicios individuales tengan que gestionar la lógica de la base de datos, lo que simplifica la arquitectura, aumenta la modularidad y protege el manejo de los datos. En última instancia, el *gestor de contexto* es el que permite al sistema mantener la continuidad, la coherencia y la conciencia contextual a lo largo de la experiencia del visitante.

3.2.3. Detector de objetos

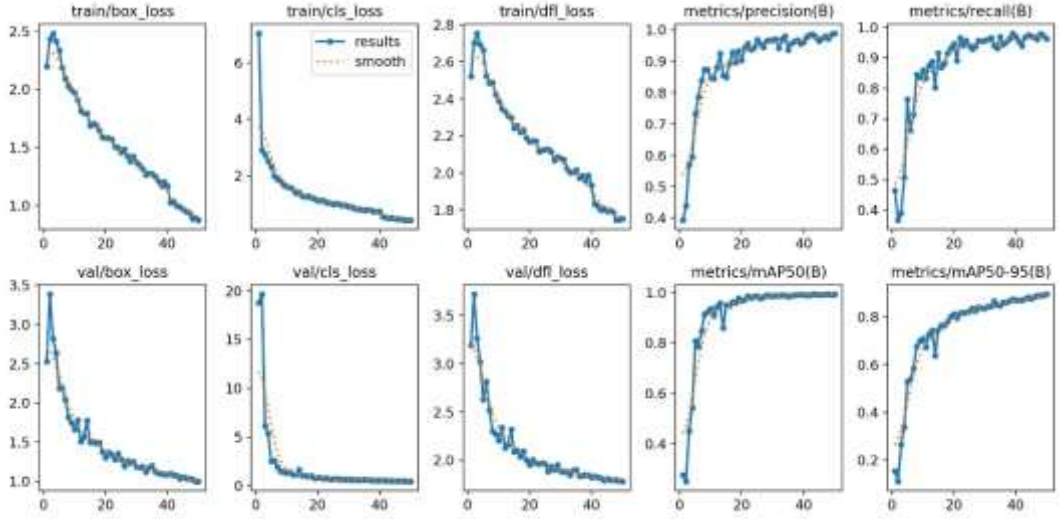
El detector de objetos se encarga de procesar los fotogramas de vídeo recibidos del *servicio de enrutamiento* y de generar predicciones sobre las obras de arte que se ven. Una vez que se le envía el fotograma, el modelo YOLO calcula las predicciones, que posteriormente se almacenan en la *base de datos contextual*, donde pueden combinarse con el historial de conversaciones. Este componente desempeña un papel fundamental al permitir respuestas basadas en la visión, ya que vincula la información visual en tiempo real del visitante con la información contextual gestionada por el resto del servicio.

Dado que la colección del Louvre incluye más de 500.000 objetos, de los cuales aproximadamente 35.000 se exhiben de forma permanente, el sistema debe ser escalable. Aunque experimentos como YOLO9000 han demostrado la capacidad de gestionar miles de clases (Redmon y Farhadi, 2017), el aumento del número de categorías en las arquitecturas convolucionales suele afectar al rendimiento (Luo et al., 2018). Para solucionar este problema, el *servicio de enrutamiento* integra el enfoque de enrutamiento con reconocimiento de ubicación mencionado anteriormente, por lo que se entrenan varios modelos con distintos conjuntos de imágenes de obras de arte ubicadas en distintas salas del museo.

El Museo del Louvre proporciona un acceso estructurado a su colección mediante puntos finales JSON, ofreciendo tanto información como imágenes de las obras de arte (Museo del Louvre, 2025). Sin embargo, estas imágenes no eran suficientes para generar un conjunto de datos de entrenamiento. Por lo tanto, se obtuvo material adicional en forma de vídeos de YouTube con el permiso de sus creadores correspondientes (Robben, 2019). El proceso de etiquetado incluyó técnicas tanto manuales como automáticas. Inicialmente, se utilizó un modelo de detección de conjunto abierto denominado Grounding Dino (Liu et al., 2024) para extraer cada imagen junto con una descripción de clase y detectar el objeto correspondiente. Este enfoque planteó problemas al tratar esculturas similares en la misma imagen. Para solucionarlo, las imágenes con varias esculturas se etiquetaron manualmente para evitar sesgos en los datos de entrenamiento, utilizando la herramienta de código abierto CVAT (CVAT, 2025) para anotar los vídeos en formato YOLO. El proceso de etiquetado fue precedido por una fase de aumento, en la que se dotó a las clases de mayor diversidad mediante desenfoque, saturación, brillo y rotaciones suaves para simular situaciones del mundo real. Además, se utilizaron imágenes no etiquetadas de otras salas del museo como imágenes de fondo, lo que ayudó a evitar falsos positivos. Tras este proceso, el conjunto de datos incluyó alrededor de 525 instancias anotadas, divididas en subconjuntos de entrenamiento (70 %), validación (20 %) y prueba (10 %), manteniendo una distribución equilibrada de clases entre los subconjuntos.

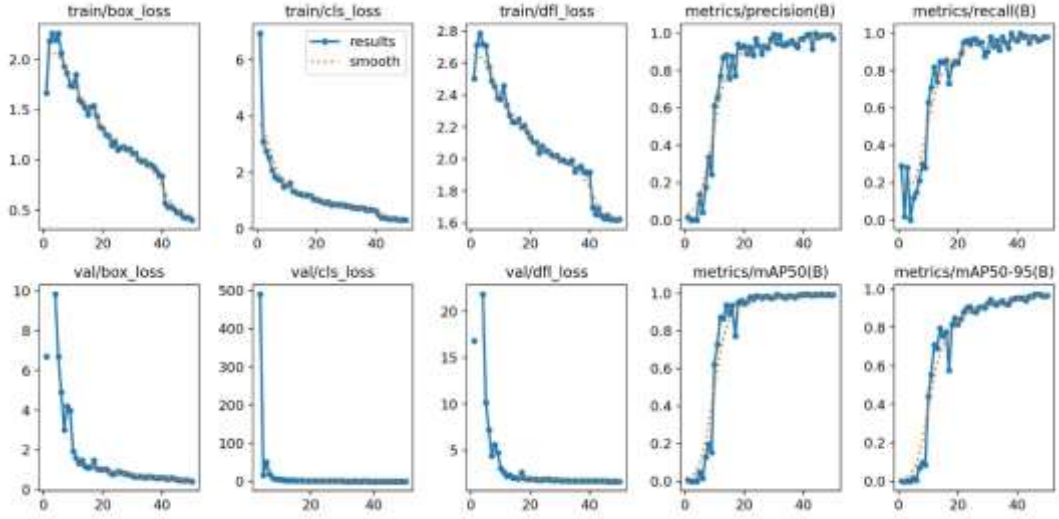
El proceso de entrenamiento comenzó seleccionando un modelo YOLO preentrenado de la gama oficial de modelos. Estos modelos basados en la detección están preentrenados en el conjunto de datos COCO (Lin et al., 2014) y están disponibles en distintas versiones y tamaños. Se eligió YOLOv10 (Wang et al., 2024) como base para el ajuste fino, ya que ofrece un aumento significativo del rendimiento con respecto a las versiones anteriores, mientras que mantiene un coste computacional inferior al de las versiones posteriores. Entre los tamaños disponibles, se seleccionó YOLOv10m.pt como una opción razonable en términos de recursos, con un equilibrio entre velocidad y precisión. El proceso de ajuste se aplicó a un par de modelos: el modelo Room 1, entrenado en cuatro esculturas mediante una combinación de etiquetado automático y manual, y el modelo Room 2, entrenado en una sola escultura con etiquetado automático. Las principales métricas de entrenamiento y evaluación del rendimiento a lo largo de las épocas se muestran en las Figuras 4 y 5, mientras que las curvas de confianza F1 de ambos modelos se muestran en la Figura 6. A partir de estos resultados, se observa que los modelos generan predicciones con precisión y confianza, sin un sobreajuste significativo en el conjunto de entrenamiento. Además, la Figura 7 resume la evaluación de las métricas de rendimiento del entrenamiento de YOLO en ambas salas.

Figura 4. Métricas de entrenamiento del modelo YOLOv10m a lo largo de las épocas (Sala 1).



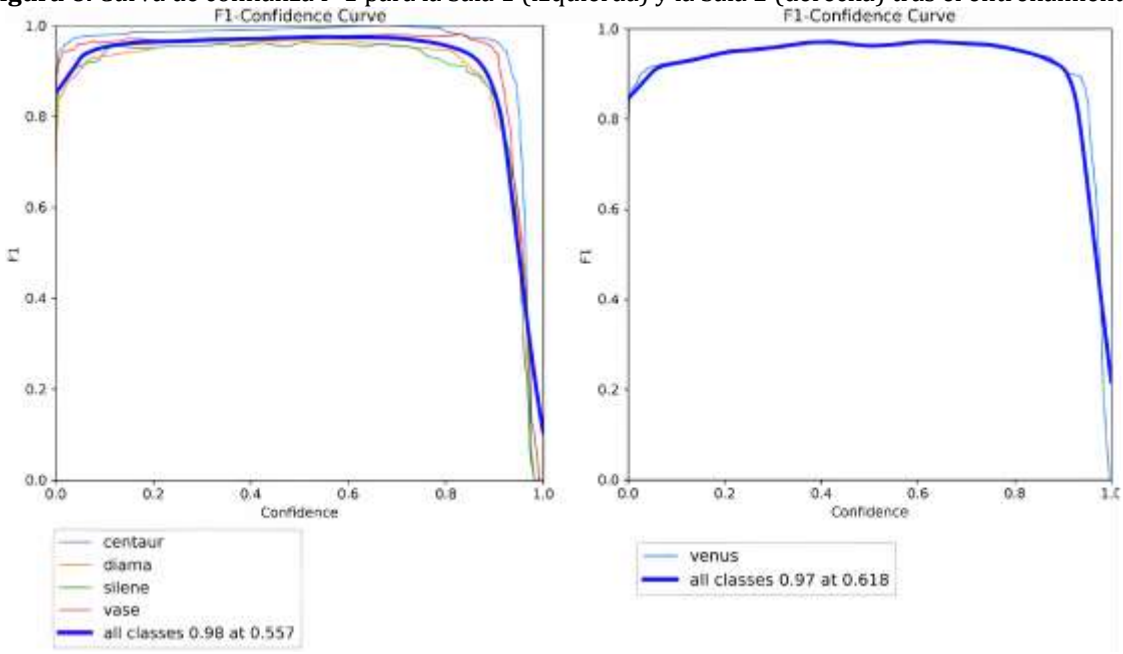
Fuente: Autor, 2025.

Figura 5. Métricas de entrenamiento del modelo YOLOv10m a lo largo de épocas (Sala 2).



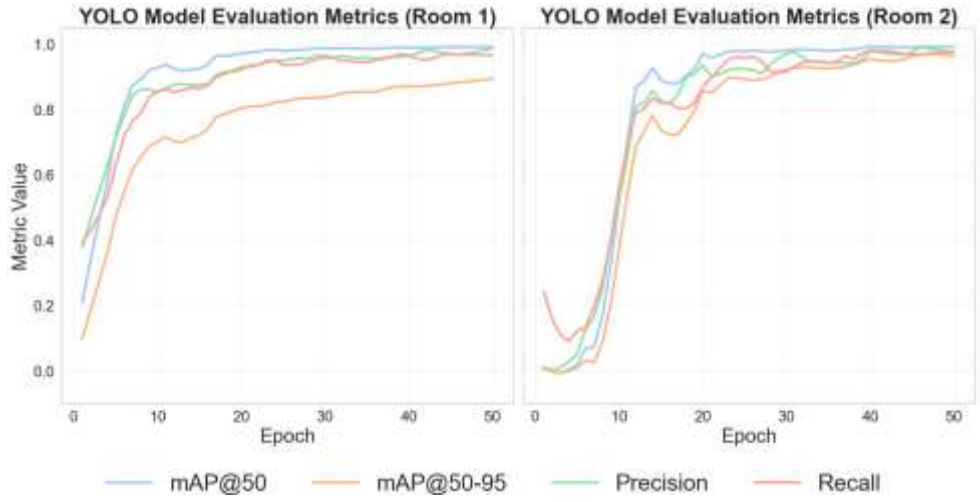
Fuente: Elaboración propia, 2025

Figura 6. Curva de confianza F-1 para la Sala 1 (izquierda) y la Sala 2 (derecha) tras el entrenamiento.



Fuente: Elaboración propia, 2025.

Figura 7. Evaluación de las métricas de rendimiento del entrenamiento de YOLO para la Sala 1 y la Sala 2.



Fuente: Elaboración propia, 2025.

3.3. Servicio de ingestión de conocimiento vectorial

Para preparar los documentos relacionados con el museo para su recuperación y generación, este servicio recopila, procesa y almacena la información en una base de datos vectorial. El flujo de trabajo comienza con la recopilación de documentos de las fuentes disponibles, continúa con el procesamiento de datos para su limpieza y su división en fragmentos semánticamente coherentes, y prosigue con la generación de representaciones vectoriales mediante modelos de incrustación. Una vez preparados, tanto los fragmentos como las incrustaciones se ingieren en la base de datos, lo que permite realizar búsquedas por similitud en toda la colección.

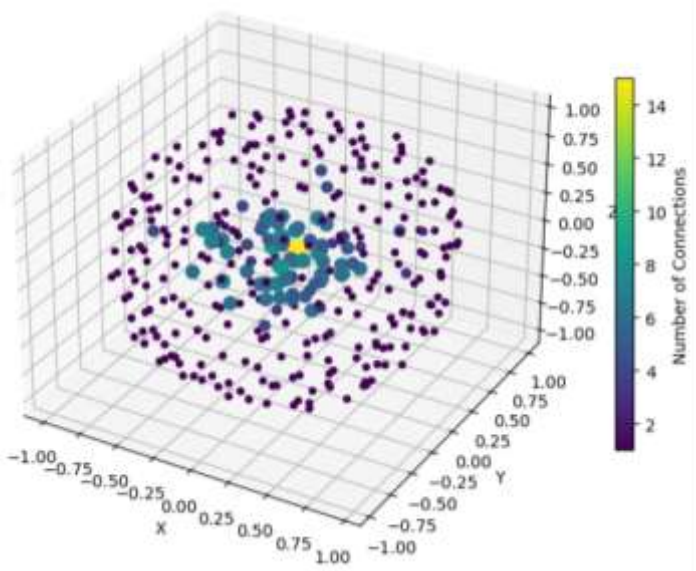
3.3.1. Recopilación de documentos

La calidad de los datos almacenados en la base de datos vectorial determina la eficacia de las futuras tareas de recuperación. Tanto la estructura como la importancia contextual del corpus de documentos son fundamentales para obtener resultados precisos. La información relevante se

recopila de catálogos de colecciones, libros históricos o contenidos dirigidos a los visitantes, pero, en la mayoría de los casos, el acceso a los datos está limitado. Aunque el Louvre ofrece acceso a su colección a través de puntos finales JSON públicos, que incluyen imágenes y descripciones, estos materiales no eran suficientes para satisfacer los requisitos del sistema. Por lo tanto, se exploraron fuentes adicionales, como libros de dominio público sobre el museo y artículos de Wikipedia. A pesar de los debates sobre su fiabilidad, Wikipedia ofrece un contenido bien organizado y semánticamente denso, y su uso ya se ha estudiado para aplicaciones similares de PLN (Yano y Kang, 2008), lo que sirvió como un excelente escenario de prueba para esta prueba de concepto.

Para apoyar la selección de artículos de Wikipedia relacionados con el Louvre que se añadirán a la base de datos vectorial, se realizó una búsqueda en profundidad a partir del artículo principal del museo, «Louvre». A partir de ahí, se seleccionaron las quince páginas más enlazadas y, en cada paso posterior, se redujo el número de enlaces a conservar hasta alcanzar una profundidad máxima de tres. La Figura 8 muestra el gráfico de enlaces final, en el que el tamaño y el color de los nodos representan el número de conexiones entrantes; el nodo amarillo central corresponde al artículo sobre el ala Lescot del Museo del Louvre, que resulta ser el más citado cuando se parte del artículo principal sobre el Louvre.

Figura 8. Gráfico de enlaces de artículos de Wikipedia a partir de «Louvre».



Fuente: Elaboración propia, 2025.

Aunque no todos los artículos del gráfico están directamente relacionados con el Museo del Louvre (por ejemplo, el muro de Carlos V en Gibraltar), resultó una herramienta útil para identificar rápidamente los que sí lo están. Aparte de estos artículos, se incluyeron algunos menos populares, pero igualmente relevantes para la demostración del proyecto. Por ejemplo, el artículo sobre el jarrón Borghese. Este jarrón se exhibe en la sala de las cariátides y se ha incluido en las clases del modelo de detección de objetos de este proyecto, por lo que se ha añadido cierta información sobre él a la base de datos vectorial con fines de validación.

3.3.2. Preprocesamiento de datos

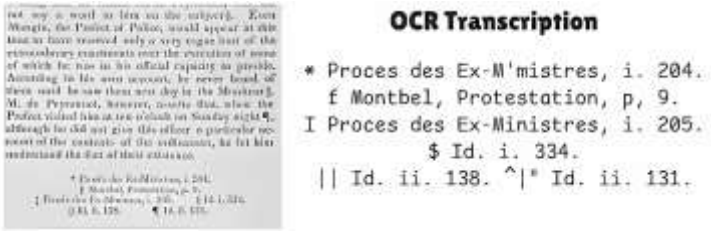
Los datos brutos recopilados en la fase de adquisición requirieron un preprocesamiento para garantizar su idoneidad para la incrustación y la ingestión. Las técnicas de preprocesamiento varían según el tipo de información. En este caso, la información se encontraba en cuatro formatos diferentes: códigos HTML de las páginas de artículos de Wikipedia y tres tipos de PDF: PDF verdaderos, PDF basados en imágenes y PDF con capacidad de búsqueda, que se refieren a la forma en que se codifica y se accede al texto en un archivo.

De estos formatos, los PDF basados en imágenes y los PDF con capacidad de búsqueda requerían técnicas de OCR para extraer el texto. Sin embargo, el OCR resultó menos preciso en los textos históricos, en los que los caracteres siguen tipografías no estándar y los libros pueden haberse deteriorado con el paso del tiempo. Tras el análisis de los datos brutos de estos PDF, se descartaron ambos formatos por la insuficiente calidad del texto. La Figura 9 ilustra una transcripción OCR fallida de un libro de principios del siglo XIX, mientras que la Figura 10 muestra información irrelevante extraída de fuentes históricas, como las notas al pie. Aunque el entrenamiento de un modelo OCR con impresiones antiguas podría haber mejorado la precisión en los textos históricos (Springmann y Lüdeling, 2016), este enfoque no se llevó a cabo debido a limitaciones de alcance. En su lugar, se adoptó la información de los formatos PDF verdaderos y HTML. Los datos HTML requerían una limpieza estructural que incluía la eliminación de etiquetas de idioma y de marcadores de referencia (por ejemplo, «[]»), y la conservación únicamente del contenido principal del cuerpo. A modo de ejemplo, la Figura 11 muestra el código HTML original de una página de Wikipedia y el texto limpio correspondiente que se utilizó para la ingestión. Una vez que los datos brutos se refinaron en un conjunto claro de textos largos, se pudo proceder a la fragmentación.

Figura 9. Transcripción OCR fallida de la carta de Henry Milton.

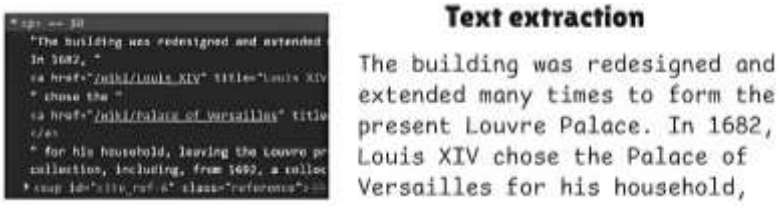
Fuente: Elaboración propia, 2025.

Figura 10. Extracción de información irrelevante de «Paris, and Its Historical Scenes».



Fuente: Elaboración propia, 2025.

Figura 11. Extracción del contenido del artículo de Wikipedia.



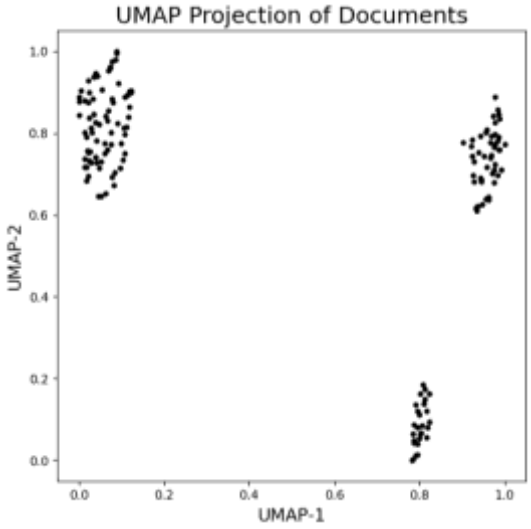
Fuente: Elaboración propia, 2025.

El componente *Chunk Loader* del proyecto gestiona la división de textos largos en unidades de información semánticamente significativas, denominadas documentos. Este paso es fundamental en el contexto de la generación aumentada por recuperación, ya que los modelos lingüísticos tienen límites de tokens y son sensibles a la longitud de la entrada (Levy et al., 2024). Además, las estrategias de fragmentación dependen de la estructura y el contenido de los datos, lo que las convierte no solo en una tarea técnica, sino también, en ocasiones, en un proceso de decisión sensible al dominio. Se llevó a cabo un análisis detallado de los posibles métodos y se concluyó que la división recursiva del texto por caracteres era la mejor opción para nuestro propósito. Este

método combina la simplicidad, de la que carecen los métodos basados en transformadores, con resultados prometedores de recuperación. Mientras que la fragmentación de tamaño fijo no va más allá de una longitud definida, la división recursiva del texto por caracteres adapta su longitud en función de la estructura del texto, dando prioridad a los límites naturales, como las frases y los párrafos, y utilizando separadores (por ejemplo, [«\n\n», «\n», « », «.»]). LangChain, un marco para crear aplicaciones relacionadas con LLM (LangChain, 2025), proporciona este componente divisor y permite ajustar el tamaño de los fragmentos y los parámetros de superposición. El tamaño de los fragmentos limita el número de caracteres de un documento, mientras que la superposición especifica el número de caracteres que se repiten desde el final del último fragmento.3.3.3. Generación de incrustaciones

Los modelos de incrustación, comúnmente denominados bi-codificadores, desempeñan un papel clave en la recuperación semántica, ya que representan el significado intrínseco de los textos como vectores densos. Se prefirió un modelo pequeño y eficiente para priorizar el rendimiento. Se aprovechó el marco Sentence-Transformers con el modelo «all-MiniLM-L6-v2», basado en la arquitectura MiniLM (Wang et al., 2020) y derivado de BERT (Devlin et al., 2019). Este modelo mapea las oraciones en un espacio vectorial denso de 384 dimensiones con una longitud máxima de entrada de 256 tokens. Por lo tanto, se adaptó la fragmentación para mantenerla dentro de los 1.024 caracteres. Para gestionar la vectorización, se creó un servidor de API de incrustación para recibir fragmentos del cargador de fragmentos, generar incrustaciones y cargarlas en la base de datos vectorial. La misma API también procesa las consultas de los usuarios desde el *RAG Hub,* incrustándolas en la búsqueda semántica mediante cálculos de similitud. Para la visualización, se aplicó la aproximación y proyección uniforme de variedades (UMAP)(McInnes et. al., 2018) para reducir la dimensionalidad de las incrustaciones y mostrar grupos de documentos, como se muestra en la Figura 12, que incluye ejemplos de la Mona Lisa, la catedral de Notre Dame y la Venus de Milo.

Figura 12 . Espacio vectorial del sistema, incluyendo 3 clases.



Fuente: Elaboración propia, 2025.

3.3.4. Ingestión de vectores

Después de recopilar, refinar, dividir e integrar los datos históricos, el sistema debe garantizar que los documentos procesados sean accesibles para su recuperación en respuesta a las consultas de los usuarios. Esta funcionalidad la proporciona la base de datos vectorial, conectada directamente a la API de integración. Cuando la API de integración recibe un documento del *cargador de fragmentos* y genera su integración correspondiente, tanto el documento como la integración se almacenan de inmediato en la base de datos.

Se seleccionó Chroma como base de datos vectorial. Funciona con colecciones de documentos que pueden configurarse para utilizar distintos algoritmos de indexación, pero se basa en el índice

Hierarchical Navigable Small World (HNSW) para la búsqueda aproximada del vecino más cercano de forma predeterminada. Para evitar entradas duplicadas, el sistema genera identificadores únicos para cada registro a partir del hash del contenido de los documentos incrustados. Si un documento ya existe en la base de datos, se rechaza la inserción.

3.4. RAG Hub

El *RAG Hub* se encarga de responder a las preguntas de los visitantes en lenguaje natural, integrando no solo la consulta en sí, sino también la información contextual relacionada con el objeto detectado. Implementa una arquitectura de generación aumentada por recuperación (Retrieval Augmented Generation) que combina la recuperación de información de memorias externas (no paramétricas) con la generación de respuestas. El sistema se estructura en dos componentes: un módulo de recuperación, que obtiene información relevante de fuentes de conocimiento externas, y un módulo generador, que emplea una memoria paramétrica para generar respuestas. Este diseño evita las limitaciones de conocimiento del LLM y minimiza los errores de desinformación al basar sus respuestas en la información recuperada, con el objetivo de equilibrar la relevancia de las respuestas con la eficiencia.

3.4.1. Recuperador

El módulo *recuperador* coordina la recuperación, la reclasificación y la preparación del contexto. Su proceso comienza tan pronto como se recibe una solicitud de un cliente, que incluye no solo una pregunta, sino también las detecciones de objetos más recientes y los mensajes históricos (información contextual). El sistema debe determinar primero la relevancia de los objetos detectados respecto de la pregunta. Por ejemplo, un usuario podría estar mirando una estatua, pero preguntar algo que no tiene relación con ella, por lo que, si la recuperación se realizara teniendo en cuenta el objeto, se recopilarían documentos erróneos. Para mitigar esto, se requiere un mecanismo de validación que evalúe la alineación semántica entre la pregunta y el objeto detectado.

Para calcular la similitud entre las etiquetas de los objetos y las preguntas de los usuarios, se empleó un modelo de codificación cruzada, pero su rendimiento fue desigual. Por lo tanto, se mejoró el enfoque mediante la reutilización del modelo generador para reformular la pregunta del usuario (Ma et al., 2023). Al añadir los mensajes anteriores, los objetos detectados y la consulta actual, el modelo generó una pregunta reformulada con mayor probabilidad de recuperar documentos relevantes y contextualizados.

Para la estrategia de recuperación, se consideraron varios métodos. Entre los más populares se encuentra el BM25 (Robertson y Zaragoza, 2009), un método de recuperación basado en palabras clave que busca documentos que coincidan con las palabras de la consulta, evaluando la similitud al otorgar mayor peso a las palabras poco comunes y a la repetición, al tiempo que considera la longitud del texto. Entre los métodos más sofisticados se encuentran la búsqueda semántica (Karpukhin et al., 2020), los codificadores cruzados (Rosa et al., 2022) y la relevancia máxima marginal (MMR) (Carbonell y Goldstein, 1998, pp. 335-336), entre otros. El módulo *Retriever* utiliza una combinación de todos los mencionados anteriormente para seleccionar los documentos más prometedores. Como parte del proceso de recuperación de la base de datos vectorial, se siguió un enfoque híbrido (Bruch et al., 2022), en el que cada algoritmo de recuperación (búsqueda semántica, BM25 y MMR) recuperaba un número k de documentos y, posteriormente, se eliminaban duplicados. Cabe señalar que la búsqueda semántica y el MMR utilizan la API de incrustación. La coherencia proviene del uso del mismo modelo de codificador bidireccional tanto en las incrustaciones de documentos como en las de consultas.

Una vez obtenida una amplia lista de documentos candidatos, se aplica un paso de reclasificación. La reclasificación consiste en ordenar los resultados de la más a la menos relevante. Si bien la recuperación ya proporciona una clasificación inicial, la reclasificación permite el uso de modelos más costosos desde el punto de vista computacional, pero también más precisos. Los codificadores cruzados, aunque excelentes para la comprensión semántica, no se suelen emplear en la primera fase de la recuperación por su alto coste computacional. Se trata de

modelos que toman una consulta y un documento como entrada combinada y producen una puntuación de relevancia muy precisa que refleja la coincidencia entre ambos. Los codificadores cruzados se utilizan comúnmente para puntuar la lista de documentos recuperados y seleccionar una lista más breve de los más relevantes.

Después de la reclasificación y la conservación de los resultados más prometedores, se realiza un paso final antes de enviarlos al generador. Según (Liu et al., 2024), los LLM tienden a perderse en la parte central de contextos largos. LangChain (2025) proporciona un componente que reordena los documentos para que los más relevantes se sitúen al principio y al final del contexto del LLM, y los menos relevantes en la parte central.

3.4.2. Generador

El módulo *Generador* corresponde a la API del LLM en la arquitectura del sistema. Este componente aloja un LLM para generar respuestas. Se seleccionó el marco Ollama (Ollama, 2025) por su perfecta integración con LangChain (2025) y su compatibilidad con modelos de código abierto que pueden implementarse localmente, algunos de los cuales incluyen cientos de miles de millones de parámetros, lo que implica un costo elevado desde el punto de vista computacional. Para mantener una solución funcional a nivel local, se exploraron modelos con entre 1000 y 8000 millones de parámetros. Tras una evaluación con un conjunto de datos predefinido, se seleccionó el modelo «qwen2.5-3b» como el más adecuado para esta prueba de concepto.

Además, cuando el *Retriever* envía al *Generador* los documentos, las conversaciones anteriores y la consulta, también selecciona una plantilla de prompt para usarla. Aparte de una plantilla predeterminada, el componente LLM API también incorpora otras plantillas que el usuario puede elegir, lo que proporciona respuestas más personalizadas. Por último, cuando se genera la respuesta, el *RAG Hub* la envía de vuelta al cliente, que a su vez la envía al *Context Handler* para almacenarla como un nuevo mensaje anterior.

4. Evaluación y resultados

Se evaluaron varios escenarios para mejorar y verificar el rendimiento del sistema, desde la selección de modelos y la justificación de técnicas hasta la formalización de conjuntos de pruebas y de datos. Las pruebas, validaciones y resultados se presentan en los principales módulos del sistema, lo que refleja la naturaleza secuencial de la arquitectura, desde el *servicio de contexto*, pasando por *el servicio de ingestión de conocimiento vectorial*, hasta el *centro RAG*, que transmite el rendimiento final del sistema completo.

4.1. Servicio de contexto

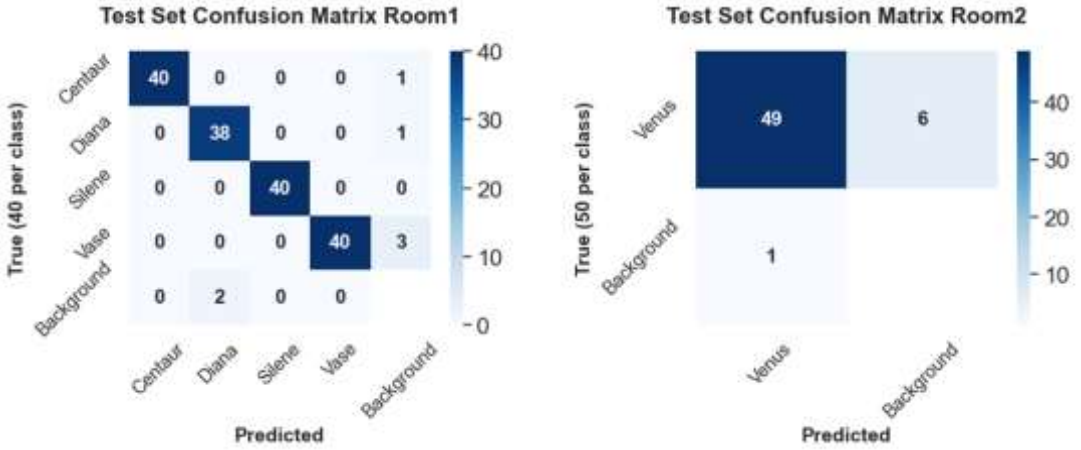
Aunque la mayor parte de las pruebas se centró en los componentes RAG, se realizaron evaluaciones adicionales para validar la precisión de la comunicación del sistema. Cabe señalar que la selección de parámetros se realizó automáticamente para el entrenamiento de los modelos de detección de objetos. El marco de entrenamiento YOLO proporciona una forma automática de seleccionar los valores más adecuados para el optimizador, la tasa de aprendizaje y el impulso. Además, la creación del conjunto de datos siguió los procedimientos estándar de aumento y distribución de imágenes. Esto dio lugar a resultados iniciales satisfactorios desde el primer entrenamiento. Por lo tanto, solo el proceso de validación fue significativo tanto en las comunicaciones como en el entrenamiento de detección de objetos.

Como se muestra en la Figura 7, se entrenaron dos modelos de detección de objetos con esculturas de dos salas distintas del Museo del Louvre. Este proceso de entrenamiento fue seguido de una fase de validación. La fase de validación incluyó realizar inferencias mediante un conjunto de pruebas para cada modelo y evaluar su rendimiento y capacidad de generalización.

Las matrices de confusión, que se muestran en la Figura 13, indican que ambos modelos se benefician de una buena capacidad de generalización general sin confusión entre clases, donde el segundo modelo podría beneficiarse de un conjunto de entrenamiento de mayor diversidad para reducir los falsos negativos, y donde el error más común entre ellos es la pérdida de detección. No

obstante, ambos modelos demuestran un gran rendimiento, con altos valores de precisión, recuperación y puntuación F1 en todas las clases, como se resume en la tabla 1.

Figura 13. Rendimiento de la detección de objetos en los conjuntos de pruebas Room1 y Room2.



Fuente: Autor, 2025.

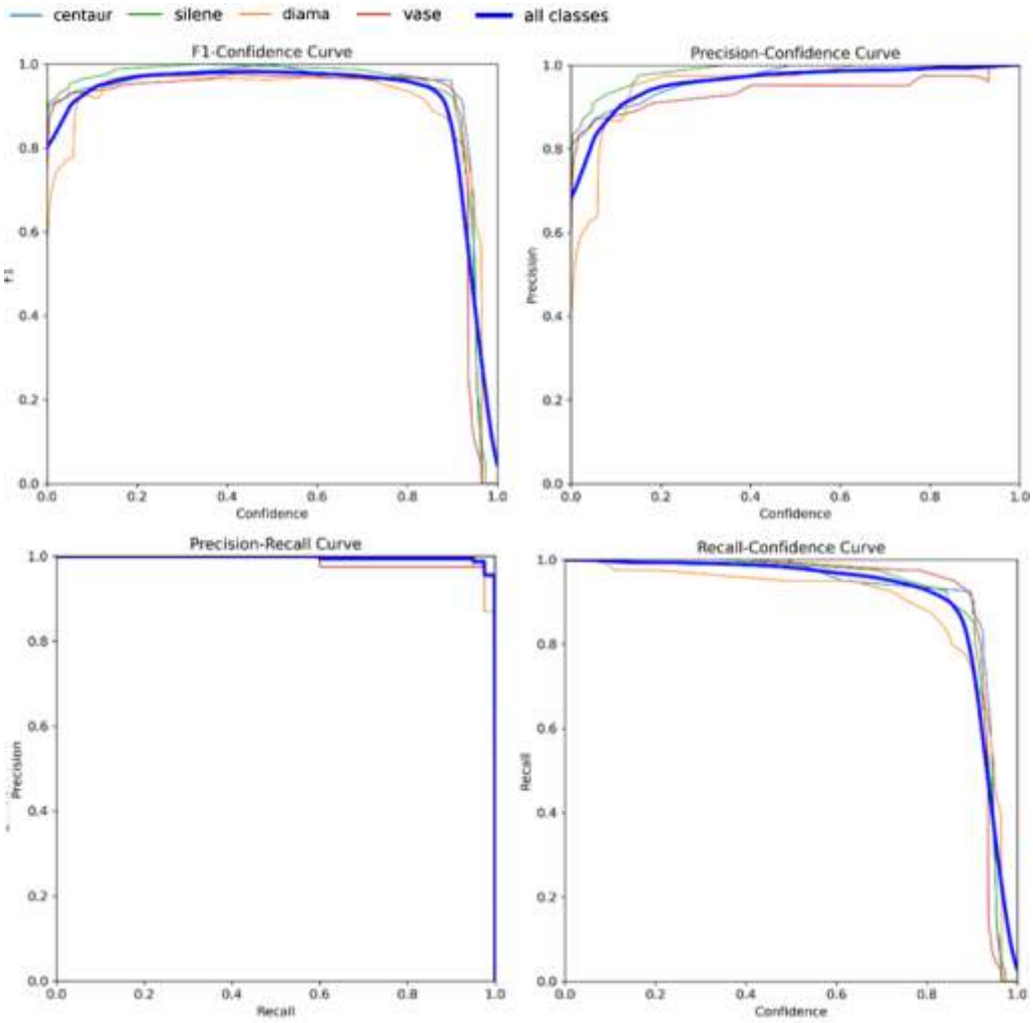
Tabla 1. Métricas promedio macro para ambos modelos de detección de objetos.

Modelo	Precisión	Recuperación	Puntuación F1
Sala 1 (Centaur, Diana, Silene, Vase)	0,988	0,970	0,978
Sala 2 (Venus)	0,980	0,891	0,933

Fuente: Elaboración propia, 2025.

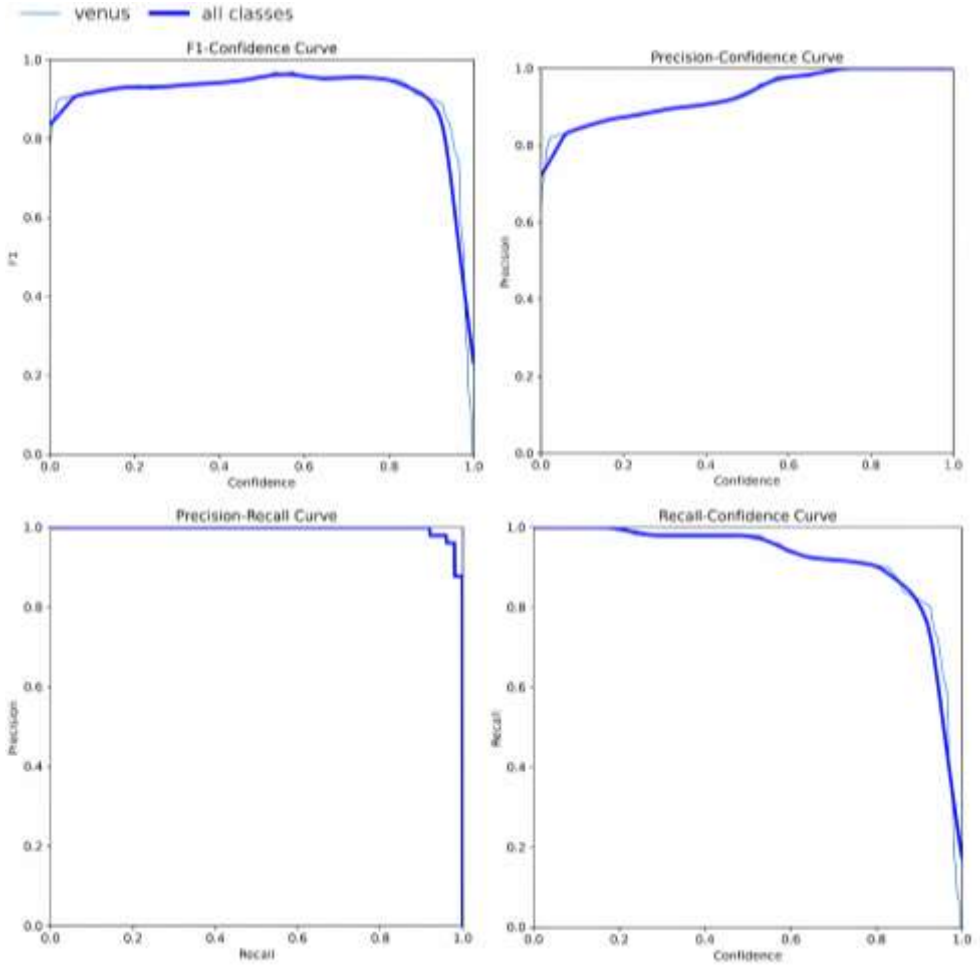
Las curvas de evaluación de la Sala 1 y la Sala 2 se presentan en las Figuras 14 y 15, respectivamente, que ilustran que los modelos alcanzan un rendimiento sólido. Estos gráficos sugieren un equilibrio sólido entre los falsos positivos y los falsos negativos (precisión-recuerdo) y un compromiso seguro entre la fiabilidad de la predicción y la confianza. En ambos modelos, la medida F1 sugiere que funcionan mejor en el intervalo de confianza de 0,4-0,6.

Figura 14. Curvas de evaluación del rendimiento del modelo YOLO en la sala 1.



Fuente: Elaboración propia, 2025.

Figura 15. Curvas de evaluación del rendimiento del modelo YOLO en la sala 2.



Fuente: Elaboración propia, 2025.

El sistema no solo debe recopilar información contextual, sino también generar respuestas basadas en ella. Por lo tanto, el sistema realiza una recuperación a partir de interacciones anteriores y de la vista actual del visitante y, a continuación, genera respuestas basadas en la información recuperada y en dichas interacciones.

Para verificar que esto es así, el sistema se ha sometido a una serie de pruebas centradas en su capacidad para hacer referencia a conversaciones anteriores y en el contexto visual pertinente. En las tablas 2 y 3 se muestran ejemplos de esas pruebas.

Tabla 2. Escenario de generación contextualizada (mensajes del historial).

Modelo: qwen2.5-3b-instruct + nuestro sistema		
Consulta del usuario	**Vista actual de los objetos**	**Mensajes históricos**
¿Qué estaba preguntando antes?		**Usuario**: Háblame del jarrón Borghese. **Asistente**: El jarrón Borghese es un cráter de mármol del siglo I a. C.
Consulta reescrita:	¿Qué estaba preguntando antes?	
Respuesta: Me preguntabas por el jarrón Borghese y te informé de que es un cráter de mármol del siglo I a. C.		

Fuente: Elaboración propia, 2025.

Tabla 3. Escenario de generación contextualizado (objetos a la vista).

Modelo: qwen2.5-3b-instruct + nuestro sistema		
Consulta del usuario	**Vista actual de los objetos**	**Mensajes históricos**
¿Cuándo se creó esto?	Jarrón Borghese	**Usuario**: **Asistente**:
Consulta reescrita:	¿Cuándo se creó el jarrón Borghese?	
Respuesta: Se cree que el jarrón Borghese fue esculpido en Atenas hacia la segunda mitad del siglo I a. C.		

Fuente: Elaboración propia, 2025.

En el primer escenario, el sistema identifica correctamente que no es necesario reescribir la información para su recuperación, ya que esta pregunta no requiere información externa. Por lo tanto, cuando llega la fase de generación de la respuesta, el *generador* recopila correctamente el contexto de los mensajes históricos previos para responder.

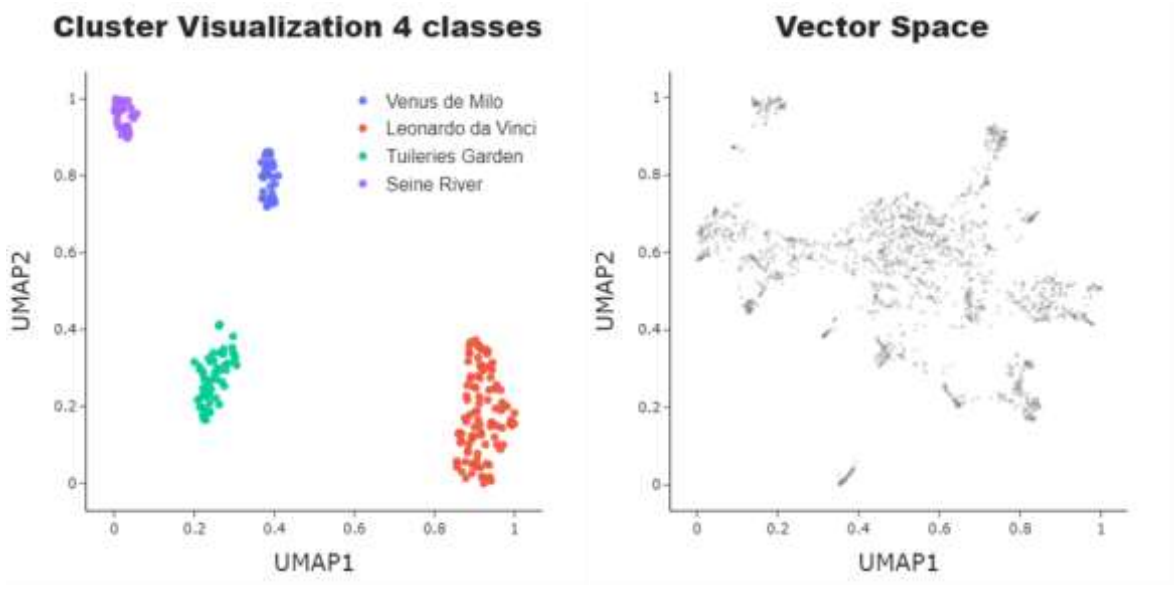
En el segundo escenario, el sistema identifica la necesidad de reescribir la pregunta, ya que es probable que la consulta del visitante se refiera a lo que ve en ese momento. Por lo tanto, añade la vista actual del usuario a la consulta para crear una nueva consulta contextualizada. Siguiendo este método, el sistema puede generar respuestas contextualizadas a partir del contexto visual previamente recopilado y de las conversaciones anteriores.

4.2. Servicio de ingestión de conocimiento vectorial

Los documentos recuperados se integran automáticamente en la base de conocimientos, lo que garantiza la coherencia y la calidad. Un espacio vectorial bien estructurado debe agrupar los documentos de la misma clase.

La Figura 16 muestra que, al incluir documentos de cuatro clases diferentes, la representación del espacio vectorial se distribuye claramente en grupos bien separados de documentos similares. Como era de esperar de contenidos que incluyen múltiples clases, la representación vectorial final incluye algunos grupos claros de documentos con documentos de enlace entre ellos, lo que crea una disposición estructurada. Aunque la calidad de la incrustación se puede ver claramente en el primer gráfico, se recopilaron algunas métricas para confirmar la precisión general del proceso de integración del conocimiento, como se resume en la tabla 4.

Figura 16. Representación del espacio vectorial con 4 clases y del espacio vectorial final.



Fuente: Elaboración propia, 2025.

Tabla 4. Métricas de calidad de la vectorización en vectores de alta dimensión. Evaluación de un subconjunto etiquetado de documentos y de un espacio vectorial no etiquetado.

Métrica	Subconjunto (4 clases)	Espacio completo
Precisión@5 (promedio)	0,955	-
Puntuación de silueta (coseno)	0,609	-
Distancia de vecindad k=5 (coseno promedio)	0,395	0,380

Fuente: Elaboración propia, 2025.

La calidad de la disposición del espacio vectorial se midió mediante dos métricas específicas para el subconjunto etiquetado y una métrica independiente para ambos conjuntos. Una precisión alta @5 indica que los cinco vecinos más cercanos de un documento suelen pertenecer a la misma clase, lo que sugiere que la similitud semántica se captura con precisión. Una puntuación de silueta alta indica que los grupos de documentos similares están bien separados. Una distancia de vecindad baja, medida entre 0 y 2, indica que los vecinos están próximos en el espacio coseno. Estos resultados validan la calidad de los procesos de limpieza, fragmentación e incrustación.

En lo que respecta a la coherencia, las fases de limpieza y fragmentación son deterministas. Además, los cálculos del modelo de incrustación se validaron mediante la reincrustación del mismo archivo de texto varias veces, lo que arrojó un error cuadrático medio (MSE) final de 0.

4.3. RAG Hub

Se creó un conjunto de datos de referencia con 126 preguntas y respuestas sobre el Louvre para evaluar diferentes modelos lingüísticos bajo las restricciones del sistema. Se probaron varios modelos de código abierto, en su mayoría con entre 1.000 y 8.000 millones de parámetros, y se utilizaron algunos más grandes como comparaciones de referencia, como se muestra en la tabla 5.

Tabla 5. Evaluación del modelo en métricas clave de generación (conjunto de datos del Louvre, sin recuperación, ordenado por COMET).

Modelo	BLUE	BERTScore (F1)	Levenshtein	COMET	SummaCZS
deepseek-r1-671b	0,345	0,942	0,714	0,797	0,337
llama-3.1-70b-instruct	0,234	0,933	0,665	0,781	0,095
qwen2.5-3b-instruct	0,152	0,920	0,616	0,775	-0,389
llama-3.1-8b-instruct	0,224	0,929	0,669	0,775	-0,106
qwen2.5-1.5b-instruct	0,203	0,921	0,661	0,759	-0,358
qwen3-1.7b-instruct	0,260	0,923	0,669	0,758	-0,483
qwen2.5-7b-instruct	0,199	0,920	0,631	0,754	-0,328
llama3.2-3b	0,131	0,909	0,575	0,714	-0,351
gemma-3-4b-it	0,258	0,913	0,583	0,680	0,232
gemma-3-27b-it	0,254	0,913	0,579	0,674	0,206
qwen3-4b-instruct	0,209	0,905	0,578	0,671	-0,314
gemma-3-1b-it	0,247	0,911	0,573	0,668	0,173
deepseek-r1-1.5b	0,095	0,897	0,523	0,665	-0,521
mistral-nemo-12b-instruct	0,166	0,879	0,482	0,586	-0,188

Fuente: Elaboración propia, 2025.

De todos los modelos pequeños evaluados, qwen2.5-3b obtuvo la mejor puntuación en COMET. Su fluidez semántica, junto con sus bajos requisitos de recursos, lo convirtió en una opción

adecuada para un entorno con limitaciones de hardware, a pesar de su precisión factual relativamente baja, como se mide mediante la métrica SummaCZS (Laban et al., 2022). Esta limitación lo convirtió en una valiosa referencia para evaluar posteriormente el efecto del aumento de la recuperación.

A continuación, se utilizó el mismo conjunto de datos para probar estrategias de recuperación. Se evaluaron combinaciones de búsqueda semántica (Cosine Similarity), BM25 y MMR utilizando ms-marco-MiniLM-L-6-v2 como evaluador. La tabla 6 resume las puntuaciones brutas de cada método y sus combinaciones.

Tabla 6 . Evaluación del método de recuperación (puntuaciones brutas).
Utilizando el codificador cruzado como evaluador con k = 10 mejores documentos y pesos iguales para las recuperaciones híbridas.

Media (promedio)	Media (promedio)	Máx. (promedio)	Desviación estándar (media)
Semántico+BM25+MMR	2,9182	6,8420	2,0081
Semántica+BM25	2,7143	6,8335	2,1603
Semántica+MMR	2,4240	6,7895	2,3418
BM25+MMR	2,3796	6,7131	2,3214
Semántico	1,7783	6,7636	2,9004
MMR	0,6364	6,4247	3,3224
BM25	-0,3909	6,0480	4,0369

Fuente: Elaboración propia, 2025.

A partir de estos resultados, se observó que la combinación de métodos de recuperación mejora el rendimiento y que la combinación de los tres proporciona el mejor equilibrio entre calidad y consistencia. La recuperación híbrida se probó más a fondo ajustando las ponderaciones del conjunto para controlar la influencia de cada método. La tabla 7 presenta las métricas de evaluación para distintas configuraciones de ponderación.

Tabla 7. Métricas de evaluación para distintas configuraciones de ponderación de recuperación (búsqueda semántica, BM25, MMR).
Utilizando el codificador cruzado para la reclasificación, k = 10 mejores documentos y reordenación de contexto largo.

Configuración	BLEU	BERTScore (F1)	COMET	SummaCZS
[0,1; 0,6; 0,3]	0,116	0,919	0,749	**0,099**
[0,3; 0,6; 0,1]	0,111	0,918	0,748	0,095
[0,6; 0,3; 0,1]	0,109	0,917	0,742	0,094
[0,3; 0,3; 0,3]	0,108	0,917	0,746	0,085
[0,1; 0,3; 0,6]	0,107	0,915	0,735	0,060

Fuente: Elaboración propia, 2025.

Aunque la recuperación semántica suele predominar en los procesos modernos, estos resultados muestran que el sistema se beneficia enormemente del BM25, probablemente debido a la naturaleza del conjunto de datos, que contiene una gran cantidad de datos concretos. Utilizando la configuración de mejor rendimiento, se evaluó la calidad de la generación con distintas cantidades de documentos recuperados, como se muestra en la tabla 8.

Tabla 8. Métricas de evaluación para distintos valores de k.
Utilizando la configuración con mejor rendimiento, con reordenación cruzada del codificador y reordenación de contexto largo.

N Mejor	BLEU	BERTScore (F1)	COMET	SummaCZS
k5	0,107	0,917	0,739	0,069
k10	0,116	0,919	0,749	0,099
k15	0,110	0,917	0,740	**0,133**
k20	0,105	0,917	0,741	0,092

Fuente: Elaboración propia, 2025.

El efecto del número de documentos recuperados depende de la capacidad del LLM para gestionar contextos largos. En este caso, qwen2.5-3b logró sus mejores resultados con alrededor de 15 documentos, tras lo cual el rendimiento disminuyó.

A partir de estos resultados de recuperación, se comparó el sistema con el modelo base qwen2.5-3b-instruct sin recuperación. Como se muestra en la tabla 9, el sistema aumentó sustancialmente el valor de SummaCZS, lo que confirma las mejoras en la veracidad. Cabe destacar que el modelo de 3.000 millones de parámetros superó el rendimiento factual de llama-3.1-70b.

Tabla 9. Comparación métrica del rendimiento de qwen2.5-3b-instruct con (k = 15) y sin recuperación de información.

Configuración	SummaCZS	COMET	BERTScore (F1)
Con recuperación	**0,133**	0,740	0,917
Sin recuperación	-0,389	0,775	0,920

Fuente: Elaboración propia, 2025.

Una ligera disminución en la adecuación y la fluidez, medida por COMET, así como una caída marginal en la similitud semántica, medida por la puntuación de BERT, sugiere que los modelos pequeños pueden verse abrumados por contextos extensos. Esto a menudo producía respuestas más prolijas con información redundante ocasional. Para investigar más a fondo, también se probaron modelos más grandes, como se muestra en la Tabla 10.

Tabla 10. Comparación métrica del rendimiento de modelos más grandes con y sin recuperación de información.

Configuración	SummaCZS	COMET	BERTScore (F1)
gemini-2.0-flash Recuperación	**0,245**	**0,785**	**0,934**
gemini-2.0-flash Sin recuperación	0,195	0,707	0,918
mistral-nemo-12b-instruct Recuperación	**0,263**	**0,603**	**0,891**
mistral-nemo-12b-instruct Sin recuperación	-0,188	0,586	0,879

Fuente: Elaboración propia, 2025.

Estos resultados confirman que el sistema mejora significativamente la precisión factual, al tiempo que mejora ligeramente la fluidez y la similitud semántica. El resultado es, por lo tanto, un proceso de generación más fiable, adecuado para proporcionar información precisa en el contexto de un museo. Es importante señalar que estos resultados no constituyen un límite superior, ya que el conjunto de datos de evaluación incluye preguntas cuyas respuestas no siempre están disponibles en la base de datos vectorial.

Por último, se validó la relevancia de la recuperación para evaluar la fundamentación de las respuestas. Utilizando el codificador cruzado ms-marco-MiniLM-L-6-v2, se evaluó la relevancia en distintos tamaños de recuperación con pesos híbridos equilibrados. Los resultados, presentados

en la tabla 11, muestran que la relevancia Score@1 se mantiene constante, mientras que las puntuaciones medias disminuyen y la desviación estándar aumenta a medida que se recuperan más documentos. Las puntuaciones Top1 constantes de 6,8 indican una fuerte alineación con las consultas.

Tabla 11. Métricas de evaluación de la recuperación en diferentes valores k utilizando ponderaciones equilibradas para la recuperación híbrida.

Estrategia	Media (promedio)	Puntuación@1 (media)	Desviación estándar (media)
Semántica + bm25 + mmr (k=5)	4,159	6,713	1,695
Semántica + bm25 + mmr (k=10)	2,918	6,842	2,008
Semántica + bm25 + mmr (k=15)	2,064	6,848	2,203

Fuente: Elaboración propia, 2025.

5. Conclusiones y debate

Este proyecto ha profundizado en el estado actual de la accesibilidad al conocimiento en el sector museístico, comparando las mejoras actuales con los métodos tradicionales del sector y con soluciones modernas aplicadas en otros ámbitos con fines similares. La solución está diseñada para centrarse en ofrecer una experiencia individual, personal y contextual a los visitantes del museo.

Tras su desarrollo, la solución incorpora un módulo de reconocimiento de objetos para comprender lo que los visitantes ven. Implica hacer que el sistema responda de forma natural y precisa a las preguntas, al tiempo que comprende lo que el visitante ve. Por último, el sistema incorpora un módulo orientado a de nuevos conocimientos precisos, de modo que las respuestas queden actualizadas con la información más reciente y relevante de la que dispone la institución museística.

El resultado del proyecto incluye una característica hasta ahora inexistente en el sector museístico (respuestas en tiempo real, sensibles a la vista), y mejora significativamente la veracidad de las respuestas generadas por los LLM, lo que las hace más fiables.

5.1. Consideraciones éticas y consentimiento informado

Las consideraciones éticas y de consentimiento informado son factores esenciales al implementar sistemas basados en la IA para la interacción con los visitantes de los museos. Aunque este trabajo no detalla la implementación completa de dichos procesos, garantiza la minimización de la retención de datos, como lo demuestra el *Context Handler*, que elimina todos los datos de las conversaciones al finalizar la sesión, preservando así la privacidad del usuario. La comunicación transparente sobre el uso de los datos y la supervisión humana responsable siguen siendo fundamentales para fomentar la confianza y la aceptación. Los esfuerzos futuros deben alinearse con marcos legales como el Reglamento General de Protección de Datos (Parlamento Europeo y Consejo, 2016) y las directrices éticas (Comisión Europea, 2019).

5.2 Dimensiones socioculturales

El trabajo puede contribuir a una experiencia museística más participativa, inclusiva y sensible a las diferencias culturales, en consonancia con los valores contemporáneos.

Las humanidades digitales fomentan el uso responsable de la IA para promover la inclusión, la diversidad cultural y el acceso equitativo (Güven et al., 2025). Para alinearse con estos principios, el sistema añadiría voz para una interacción más intuitiva e inclusiva (texto a voz y voz a texto), podría admitir la interacción multilingüe y proporcionaría respuestas en un lenguaje sencillo, calibradas según los conocimientos previos y el dominio del idioma de los visitantes. En el momento de la ingestión, se espera que el personal del museo seleccione y revise los materiales

para mitigar sesgos, diversificando los documentos, registrando las fuentes y los niveles de confianza e incorporando las aportaciones de la comunidad antes de que el contenido entre en el almacén vectorial.

La museología crítica insta a ir más allá de las narrativas estáticas de los museos para fomentar una participación activa, dialógica y reflexiva, lo cual se ajusta a este enfoque al pedir al sistema que apoye la participación activa de los visitantes en lugar del consumo pasivo (Boulakal y Hadi, 2025; Lundgren et al., 2019). El diseño de interacciones conversacionales que inviten a la reflexión y de narrativas diversas mejoraría el empoderamiento de los visitantes y la conciencia cultural (Damiano et al., 2022). El sistema debería incorporar indicaciones de diálogo que animen a los visitantes a reflexionar y a compartir perspectivas; podría habilitar mecanismos de retroalimentación para recopilar aportaciones con vistas a una mejora iterativa y ofrecer respuestas basadas en múltiples fuentes para presentar interpretaciones variadas. En conjunto, estas características convertirían la orientación en un diálogo dinámico e inclusivo que respeta la pluralidad cultural y promueve el pensamiento crítico.

5.3. Orientaciones técnicas futuras

En la capa de interacción, la calidad de la respuesta también podría mejorar mediante el análisis de plantillas de indicaciones y el ajuste de los parámetros de muestreo y de decodificación. El *servicio de contexto* podría integrar modelos de estimación de profundidad y, considerando la posición de los objetos dentro del marco, mejorar la relevancia de las obras de arte detectadas, mientras que una arquitectura de detección de objetos más flexible también permitiría mover las obras de arte entre salas sin reducir el rendimiento del sistema.

En el *servicio de ingestión de conocimiento vectorial*, la migración de Chroma a una base de datos vectorial gestionada y de nivel de producción, como Pinecone, podría aumentar la eficiencia; la ingestión también podría enriquecer automáticamente los documentos con metadatos, y un modelo OCR especializado ampliaría la gama de fuentes PDF utilizables.

La calidad de la recuperación podría mejorar mediante la fragmentación agencial de la relevancia semántica y modelos especializados para la incrustación y la reclasificación en arte/historia. La calidad de la respuesta también podría beneficiarse de la ingeniería de prompts, del ajuste de parámetros o del ajuste fino. Técnicas como ReAct también podrían ayudar a incluir funcionalidades como el acceso a Internet, evitar la recuperación si no es necesaria o la búsqueda por palabras clave.

Por último, ampliar el modelo con más obras de arte e información podría proporcionar una mejor simulación de un escenario real.

Referencias

Ask Mona. (2025). *Ask Mona*. https://www.askmona.fr

Breitner, A. R., & Bandung, Y. (2024). Development of visitor interest detection and tracking system in the museums. *Journal of Sustainable Engineering: Proceedings Series, 2*(1), 7–12. https://doi.org/10.35793/joseps.v2i1.1279

Bruch, S., Gai, S., & Ingber, A. (2023). An analysis of fusion functions for hybrid retrieval. *ACM Transactions on Information Systems, 42*(1), 1–35. https://doi.org/10.1145/3596512

Boulakal, F., & Hadi, W. M. E. (2025). Cultural & Knowledge Spaces: the Immersive Museums as a Challenge for KO and the Digital Humanities. *Informatio*, *30*(1), e205. https://doi.org/10.35643/info.30.1.10

Bu, F., Wang, Z., Wang, S., & Liu, Z. (2025). An investigation into value misalignment in LLM-generated texts for cultural heritage. *arXiv*, *1*. https://doi.org/10.48550/arXiv.2501.02039

Carbonell, J., & Goldstein, J. (1998). The use of MMR, diversity-based reranking for reordering documents and producing summaries. En *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 335–336). ACM. https://doi.org/10.1145/290941.291025

Cetinić, E., Lipić, T., & Grgić, S. (2018). Fine-tuning convolutional neural networks for fine art classification. *Expert Systems with Applications, 114*, 107–118. https://doi.org/10.1016/j.eswa.2018.07.026

CVAT. (2025). *CVAT: Computer Vision Annotation Tool* [Computer software]. https://github.com/opencv/cvat

Damiano, R., Kuflik, T., Wecker, A. J., Striani, M., Lieto, A., Bruni, L. E., Kadastik, N., & Pedersen, T. A. (2022). Exploring values in museum artifacts in the SPICE project: A preliminary study. En *Adjunct Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization (UMAP '22 Adjunct)* (pp. 391–396). Association for Computing Machinery. https://doi.org/10.1145/3511047.3537662

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. En *Proceedings of NAACL-HLT 2019* (pp. 4171–4186). Association for Computational Linguistics. https://aclanthology.org/N19-1423/

Du, X., Zheng, G., Wang, K., Feng, J., Deng, W., Liu, M., Chen, B., Peng, X., Ma, T., & Lou, Y. (2024). Vul-RAG: Enhancing LLM-based vulnerability detection via knowledge-level RAG. *arXiv*, *1*. https://doi.org/10.48550/arXiv.2406.11147

European Commission. (2019). *Ethics guidelines for trustworthy AI*. Independent High-Level Expert Group on Artificial Intelligence set up by the European Commission. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

European Parliament & Council. (2016). *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation)*. Official Journal of the European Union, L119, 1–88. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32016R0679

Fortuna-Cervantes, J. M., Soubervielle-Montalvo, C., Puente-Montejano, C. A., Pérez-Cham, O. E., & Peña-Gallardo, R. (2024). Evaluation of CNN models with transfer learning in art media classification in terms of accuracy and class relationship. *Computación y Sistemas, 28*(1), 233–244. https://doi.org/10.13053/cys-28-1-4895

Güven, Ç., Alishahi, A., Brighton, H., Nápoles, G., Olier, J. S., Šafář, M., Postma, E., Shterionov, D., De Sisto, M., & Vanmassenhove, E. (2025). *AI in support of diversity and inclusion*. *arXiv*. https://doi.org/10.48550/arXiv.2501.09534

Karpukhin, V., Oguz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., & Yih, W.-t. (2020). Dense passage retrieval for open-domain question answering. En *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 6769–6781). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.550

LangChain. (2025). *LangChain* [Computer software]. https://github.com/langchain-ai/langchain

Levy, M., Jacoby, A., & Goldberg, Y. (2024). Same task, more tokens: The impact of input length on the reasoning performance of large language models. En *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 15339–15353). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.acl-long.818

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., & Zitnick, C. L. (2014). Microsoft COCO: Common objects in context. En *European conference on computer vision (ECCV 2014)* (pp. 740–755). Springer. https://doi.org/10.1007/978-3-319-10602-1_48

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., & Berg, A. C. (2016). SSD: Single shot multibox detector. En *European Conference on Computer Vision (ECCV 2016)* (pp. 21–37). Springer. https://doi.org/10.1007/978-3-319-46448-9_2

Liu, N. F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., & Liang, P. (2024). Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics, 12*, 157–173. https://aclanthology.org/2024.tacl-1.9/

Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Jiang, Q., Li, C., Yang, J., Su, H., Zhu, J., & Zhang, L. (2024). Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection. *arXiv*. https://arxiv.org/abs/2303.05499

Loffredo, R., & De Santo, M. (2024). Using ontologies for LLM applications in cultural heritage. In *CEUR Workshop Proceedings* (Vol. 3865). https://ceur-ws.org/Vol-3865/06_paper.pdf

Lundgren, L., Stofer, K., Dunckel, B., Krieger, J., Lange, M., & James, V. (2019). Panel-based exhibit using participatory design elements may motivate behavior change. *Journal of Science Communication*, *18*(02), A03. https://doi.org/10.22323/2.18020203

Luo, C., Li, X., Wang, L., He, J., Li, D., & Zhou, J. (2018). How does the data set affect CNN-based image classification performance? En *2018 5th International Conference on Systems and Informatics (ICSAI)* (pp. 361–366). IEEE. https://doi.org/10.1109/ICSAI.2018.8599448

Ma, X., Gong, Y., He, P., Zhao, H., & Duan, N. (2023). Query rewriting for retrieval-augmented large language models. En *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 5303–5315). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.emnlp-main.322

McInnes, L., Healy, J., Saul, N., & Großberger, L. (2018). UMAP: Uniform manifold approximation and projection. *Journal of OpenSource Software, 3*(29), 861. https://doi.org/10.21105/joss.00861

Meyer, L. S., Engel Aaen, J., Tranberg, A. R., Kun, P., Freiberger, M., Risi, S., & Løvlie, A. S. (2024). Algorithmic ways of seeing: Using object detection to facilitate art exploration. En *CHI '24: Proceedings of the CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery. https://doi.org/10.1145/3613904.3642157

Museo del Louvre. (2025). Collections site JSON documentation [Website]. Retrieved January 20, 2025, from https://collections.louvre.fr/en/page/documentationJSON

Nubart. (2025). *Nubart* [Mobile application]. https://www.nubart.eu

Ollama. (2025). *Ollama* [Computer software]. https://github.com/ollama/ollama

Redmon, J., & Farhadi, A. (2017). YOLO9000: Better, faster, stronger. En *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 6517–6525). IEEE. https://doi.org/10.1109/CVPR.2017.690

Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster R-CNN: Towards real-time object detection with region proposal networks. *Advances in Neural Information Processing Systems, 28*. https://arxiv.org/abs/1506.01497

Robben, H. [Wanderlust Travel Videos]. (2019, 31 de mayo). Louvre Museum Paris – Mona Lisa – walking tour | 4K [Video]. YouTube. https://youtu.be/6vuFh6NNa70

Robertson, S., & Zaragoza, H. (2009). The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval, 3*(4), 333–389. https://doi.org/10.1561/1500000019

Rosa, G. M., Bonifacio, L. H., Jeronymo, V., Abonizio, H. Q., Fadaee, M., Lotufo, R. A., & Nogueira, R. (2022). *In defense of cross-encoders for zero-shot retrieval*. arXiv. https://arxiv.org/abs/2212.06121

Sahoo, P. K., Sharma, N., Mehta, P., Kumar, S., Garg, A., … & Pratama, M. (2024). A systematic survey of prompt engineering in large language models: Techniques and applications. *arXiv*. https://doi.org/10.48550/arXiv.2402.07927

Smartify. (2025). *Smartify* [Mobile application]. https://smartify.org

Smith, J. K., & Smith, L. F. (2001). Spending time on art. *Empirical Studies of the Arts, 19*(2), 229–236. https://doi.org/10.2190/5MQM-JWH6-V2P4-7DLK

Smith, J. K., Smith, L. F., & Tinio, P. P. L. (2017). Time spent viewing art and reading labels. *Psychology of Aesthetics, Creativity, and the Arts, 11*(1), 77–85. https://doi.org/10.1037/aca0000049

Smith, B., & Troynikov, A. (2024, 3 de julio). Evaluating chunking strategies for retrieval (Chroma Technical Report). Chroma. https://research.trychroma.com/evaluating-chunking-strategies-in-retrieval

Springmann, U., Lüdeling, A., & Ernst, F. (2017). OCR of historical printings with an application to building diachronic corpora: The RIDGES herbal corpus. *Digital Humanities Quarterly, 11*(2). http://www.digitalhumanities.org/dhq/vol/11/2/000291/000291.html

United Nations General Assembly. (2015). *Transforming our world: The 2030 agenda for sustainable development* (A/RES/70/1). https://sustainabledevelopment.un.org/post2015/transformingourworld/publication

Vastakas, L. (2024). *Cultural heritage search with large language models: Enhancing the discoverability of cultural heritage artifacts through large language model-based search systems* [Master's thesis, Linnaeus University]. DiVA portal. https://urn.kb.se/resolve?urn=urn:nbn:se:lnu:diva-132431

Wang, A., Chen, H., Liu, L., Chen, K., Lin, Z., Han, J., & Ding, G. (2024). YOLOv10: Real-time end-to-end object detection. *arXiv*. https://arxiv.org/abs/2405.14458

Wang, W., Wei, F., Dong, L., Bao, H., Yang, N., & Zhou, M. (2020). MiniLM: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. En *NeurIPS 2020*. https://proceedings.neurips.cc/paper/2020/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

Wu, J., Zhu, J., Qi, Y., Chen, J., Xu, M., Menolascina, F., & Grau, V. (2024). Medical graph RAG: Towards safe medical large language model via graph retrieval-augmented generation. *arXiv*. https://doi.org/10.48550/arXiv.2408.04187

Yano, T., & Kang, M. (2008). Taking advantage of Wikipedia in natural language processing. Language Technologies Institute, Carnegie Mellon University. https://www.cs.cmu.edu/~taey/pub/wiki.pdf