MÁSTER EN FORMACIÓN PERMANENTE EN INTELIGENCIA ARTIFICIAL

MultimodalBioQA: Un framework agéntico con explicabilidad incorporada que integra grandes modelos de visión y lenguaje con procesamiento de lenguaje natural para *question answering* biomédico multimodal

Presentado por:

Johanna Angulo

Dirigido por:

Víctor Yeste

CURSO ACADÉMICO 2024-2025

Resumen

Los sistemas de *question answering* biomédico actuales enfrentan limitaciones en explicabilidad, lo que obstaculiza su adopción en entornos clínicos donde la interpretabilidad es fundamental para la toma de decisiones médicas basadas en evidencia.

Se desarrolló **MultimodalBioQA**, un sistema multimodal que integra procesamiento textual y visual con capacidades de explicabilidad (XAI). El módulo textual implementa búsqueda híbrida combinando bases de datos vectoriales locales (PubMedBERT 768D) con consultas en tiempo real a PubMed API, utilizando GPT-4o para extracción de evidencia científica a nivel de oración. El módulo visual emplea un modelo LLaVA-LLaMA 3 8B con *fine-tuning* LoRA especializado para análisis de imágenes médicas.

La contribución principal es el sistema de explicabilidad integrado que combina cuatro métodos complementarios: GradCAM, *Attention Maps*, *Integrated Gradients* y mapeo concepto-región con *bounding boxes* mediante técnicas de segmentación automática. Esta integración permite generar respuestas fundamentadas con trazabilidad completa desde la evidencia hasta la conclusión.

La arquitectura integra múltiples componentes: LLaVA-LLaMA 3 8B, GPT-4o, BiomedNLP - PubMedBERT, *Segment Anything Model* (SAM), bases de datos vectoriales RAG, y APIs UMLS/MeSH y PubMed. El sistema se implementa como una arquitectura multiagente con LlamaIndex, coordinada mediante un *Writer Agent* central que sintetiza respuestas diferenciadas según el tipo de consulta biomédica.

La evaluación en competencias internacionales demostró rendimiento competitivo: los módulos textual y visual del sistema obtuvieron desempeño *top 10* en BioASQ Tarea 13B e ImageCLEFmed Caption 2025, confirmando su efectividad en comparación con sistemas de instituciones de investigación consolidadas a nivel mundial.

Este trabajo contribuye al desarrollo de sistemas de IA médica interpretable mediante la integración de explicabilidad multimodal desde el diseño, estableciendo una base técnica para futuras herramientas de apoyo clínico que combinen precisión con transparencia en el análisis de información biomédica.

Palabras clave: Question Answering Biomédico, IA Multimodal, XAI, IA Generativa, IA Agéntica, PLN Biomédico, Visual Transformers, Multimodal Transformers, RAG, IA Médica

Abstract

Current biomedical question answering systems face limitations in explainability, which hinders their adoption in clinical environments where interpretability is fundamental for evidence-based medical decision-making.

We developed **MultimodalBioQA**, a multimodal system that integrates textual and visual processing with explainability capabilities (XAI). The textual module implements hybrid search combining local vector databases (PubMedBERT 768D) with real-time PubMed API queries, utilizing GPT-4o for sentence-level scientific evidence extraction. The visual module employs a LLaVA-LLaMA 3 8B model with specialized LoRA fine-tuning for medical image analysis.

The primary contribution is an integrated explainability system that combines four complementary methods: GradCAM, Attention Maps, Integrated Gradients, and concept-region mapping with bounding boxes through automatic segmentation techniques. This integration enables the generation of grounded responses with complete traceability from evidence to conclusion.

The architecture integrates multiple components: LLaVA-LLaMA 3 8B, GPT-4o, BiomedNLP - PubMedBERT, Segment Anything Model (SAM), RAG vector databases, and UMLS/MeSH and PubMed APIs. The system is implemented as a multi-agent architecture with LlamaIndex, coordinated through a central Writer Agent that synthesizes differentiated responses according to biomedical query type.

Evaluation in international competitions demonstrated competitive performance: the system's textual and visual modules achieved Top 10 results in BioASQ Task 13B and ImageCLEFmed Caption 2025, confirming their effectiveness compared to systems from established research institutions worldwide.

This work contributes to the development of interpretable medical AI systems through the integration of multimodal explainability from design, establishing a technical foundation for future clinical support tools that combine precision with transparency in biomedical information analysis.

Keywords: Biomedical Question Answering, Multimodal AI, XAI, Generative AI, Agentic AI, Biomedical NLP, LLM, Visual Transformers, Multimodal Transformers, RAG, Health AI

Índice

Índice de figuras

Índice de cuadros

1. Introducción

La integración de sistemas de inteligencia artificial en el dominio biomédico ha experimentado un crecimiento exponencial en la última década, transformando fundamentalmente la manera en que se accede, procesa y sintetiza el conocimiento médico (Warner et al., 2024). En este contexto, los sistemas de `question answering` (Q&A) biomédico han emergido como una tecnología crítica para democratizar el acceso a información científica especializada, permitiendo que profesionales de la salud, investigadores y estudiantes formulen consultas en lenguaje natural y obtengan respuestas precisas basadas en evidencia científica actualizada (Kell et al., 2024).

La medicina moderna se caracteriza por su naturaleza inherentemente multimodal, donde la toma de decisiones clínicas requiere la integración de información textual (literatura científica, historiales clínicos, guías de práctica) e información visual (imágenes radiológicas, histopatológicas, microscópicas) (Simon et al., 2024). Sin embargo, los sistemas de Q&A biomédico tradicionales han operado predominantemente en modalidades aisladas, limitando su capacidad para proporcionar análisis integrales que reflejen la realidad clínica multidimensional (Warner et al., 2024). Esta fragmentación de modalidades representa una limitación significativa para la medicina de precisión, donde la síntesis de evidencia multimodal es fundamental para diagnósticos precisos y planificación terapéutica personalizada. La necesidad de sistemas que puedan procesar simultáneamente consultas textuales complejas y análisis de imágenes médicas especializadas es imperativa para avanzar hacia herramientas de apoyo clínico verdaderamente integradas (Tariq et al., 2025).

Paralelamente al desarrollo de capacidades multimodales, la comunidad médica y regulatoria ha identificado la explicabilidad como un requisito fundamental para la adopción responsable de sistemas de IA en entornos clínicos (Cálem et al., 2024). A diferencia de otros dominios donde los modelos `black-box` pueden ser aceptables, la medicina exige transparencia interpretativa que permita a los profesionales comprender, validar y confiar en las recomendaciones generadas por sistemas automatizados (Guidotti et al., 2018; Markus et al., 2021). La explicabilidad en IA médica trasciende la mera interpretación de resultados; constituye un imperativo ético y regulatorio que determina la viabilidad de la implementación clínica. En esta área los sistemas no solo deben proporcionar respuestas precisas, sino también articular el razonamiento subyacente, identificar las fuentes de evidencia y permitir la verificación independiente de las conclusiones generadas (Markus et al., 2021). Esta exigencia se intensifica en el contexto multimodal, donde la integración de información textual y visual requiere mecanismos de explicabilidad sofisticados que puedan elucidar las contribuciones relativas de cada modalidad al resultado final (Borys et al., 2023).

Por lo tanto, el presente sistema pretende ser un aporte en el avance hacia la resolución de esta problemática.

2. Motivación

La convergencia entre inteligencia artificial y medicina representa uno de los desafíos más apremiantes y prometedores en la actualidad. Mi trabajo en **MultimodalBioQA** emerge de una convicción profunda: que las decisiones clínicas respaldadas por sistemas interpretables pueden transformar radicalmente los resultados para los pacientes, especialmente en contextos donde el tiempo de diagnóstico es crítico.

La medicina de precisión asistida por IA no es meramente una innovación tecnológica; es una oportunidad de democratizar el acceso a diagnósticos de calidad y reducir las disparidades en la atención médica. Cada algoritmo que desarrollamos, cada modelo que entrenamos, cada sistema de explicabilidad que implementamos, tiene el potencial de traducirse en intervenciones más tempranas, tratamientos más efectivos y, en última instancia, en preservar vidas.

Mi trayectoria académica y profesional ha sido, en retrospectiva, una preparación deliberada hacia este momento de convergencia. Los fundamentos en ciencias de la computación, la experiencia en procesamiento de lenguaje natural, el trabajo con datos biomédicos masivos, y la investigación independiente en explicabilidad de modelos, han confluido en una visión clara: contribuir al desarrollo de sistemas de IA médica que no solo sean precisos, sino también transparentes y confiables para los profesionales de la salud y los pacientes.

El desarrollo de **MultimodalBioQA** representa más que un logro técnico; encarna una filosofía de investigación centrada en el impacto humano. Creo profundamente que la explicabilidad integrada no es solo un requisito técnico, sino un imperativo ético que reconoce que, en medicina, entender el "por qué"detrás de una recomendación puede ser tan crucial como la recomendación misma. Esta investigación me ha confirmado que la intersección entre IA y medicina biomédica es donde mi contribución puede ser más significativa. La complejidad inherente de los datos médicos multimodales, la necesidad crítica de interpretabilidad en decisiones clínicas, y el potencial transformador de estas tecnologías, constituyen un espacio donde siento que la excelencia técnica a la que aspiro se encuentra con el propósito social. Mi objetivo es consolidar una línea de investigación que no solo avance el estado del arte en IA médica, sino que también establezca nuevos estándares para la responsabilidad y transparencia en el desarrollo de herramientas de apoyo clínico. Sinceramente aspiro a que mi trabajo contribuya a un futuro donde la IA médica no solo sea más inteligente, sino también más humana en su capacidad de explicar, justificar y generar confianza tanto en los profesionales que dedican sus vidas a sanar como en sus pacientes que, tarde o temprano, somos todos.

3. Justificación de la investigación

Los sistemas de Q&A biomédico existentes enfrentan limitaciones arquitectónicas y metodológicas significativas que impiden su adopción generalizada en entornos clínicos (Jin et al., 2022). En primer lugar, la mayoría de los sistemas especializados operan en modalidades únicas, requiriendo que los usuarios consulten múltiples herramientas para obtener análisis completos. Esta fragmentación no solo reduce la eficiencia del workflow clínico, sino que también introduce inconsistencias en la interpretación y síntesis de la información multimodal.

En segundo lugar, los enfoques actuales de explicabilidad en IA médica tienden a ser superficiales o limitados a técnicas individuales (por ejemplo, únicamente attention maps). Esta aproximación fragmentada no logra proporcionar la comprensión holística requerida para validación clínica, donde los profesionales necesitan entender tanto las contribuciones específicas de características individuales como las interacciones complejas entre modalidades.

Además, los sistemas existentes frecuentemente sacrifican interpretabilidad en favor de rendimiento, o viceversa, creando un trade-off artificial que limita su utilidad práctica. La investigación actual incluye marcos arquitectónicos que demuestran que la explicabilidad

puede coexistir, e incluso potenciar, el rendimiento competitivo en tareas biomédicas complejas.

Por otra parte, el advenimiento de los grandes modelos de lenguaje (LLMs) y los sistemas multimodales avanzados ha creado oportunidades sin precedentes para superar las limitaciones históricas de los sistemas de Q&A biomédico (Thirunavukarasu et al., 2023). Los LLMs modernos demuestran capacidades extraordinarias para comprensión contextual, síntesis de información compleja y generación de explicaciones articuladas, mientras que los modelos multimodales emergentes pueden procesar simultáneamente información textual y visual con un nivel de expertise especializado (Singhal et al., 2023).

Simultáneamente, las arquitecturas multiagente han emergido como un paradigma prometedor para orquestar sistemas complejos que requieren especialización específica para el dominio manteniendo coordinación global (Pandey et al., 2024). Este enfoque permite la construcción de sistemas donde agentes especializados manejan aspectos específicos del procesamiento (búsqueda de literatura, análisis visual, síntesis de respuestas) mientras un mecanismo de coordinación central asegura coherencia y calidad en el output final (Pandey et al., 2024).

La convergencia de estas tecnologías emergentes presenta una oportunidad única para desarrollar sistemas de Q&A biomédico que no solo superen las limitaciones actuales en términos de multimodalidad y explicabilidad, sino que establezcan nuevos estándares de transparencia, rendimiento y utilidad clínica en IA médica.

En este contexto, el desarrollo de sistemas de Q&A biomédico multimodal con explicabilidad integrada representa una frontera crítica de investigación que tiene el potencial de transformar fundamentalmente la interacción entre los profesionales de la salud y el conocimiento médico digitalizado. Estos sistemas deben abordar simultáneamente los desafíos técnicos del procesamiento multimodal, los requisitos metodológicos de la explicabilidad holística y las exigencias prácticas de la implementación clínica (Simon et al., 2024).

El presente trabajo aborda esta convergencia de desafíos mediante el desarrollo de un sistema multiagente que integra capacidades avanzadas de procesamiento textual y visual con mecanismos de explicabilidad multitécnica, estableciendo una base sólida para la próxima generación de herramientas de IA médica interpretable.

3.1. Objetivo ODS

Este trabajo de investigación se enmarca dentro del Objetivo de Desarrollo Sostenible (ODT) de Salud y Bienestar de la Unión Europea, el que tiene como finalidad asegurar una vida saludable y promover el bienestar para todas las personas, sin importar la edad. Sus propósitos incluyen disminuir la mortalidad materna e infantil, combatir tanto enfermedades transmisibles como no transmisibles, y garantizar el acceso universal a servicios de salud esenciales y a medicamentos seguros (ONU, s.f.).

Una aplicación de información biomédica puede contribuir de manera directa a este objetivo mediante:

- El acceso sencillo a información médica actualizada y fiable.
- La promoción de la prevención y el autocuidado en temas de salud.
- El fortalecimiento de la educación sanitaria en la sociedad.
- El apoyo a la vigilancia epidemiológica y al monitoreo de enfermedades.

Tanto la OMS como la ONU reconocen que la salud digital, incluidas las aplicaciones biomédicas, desempeña un papel fundamental como impulsora para alcanzar los objetivos ODS. Esto se debe a que amplía la cobertura de servicios de salud, refuerza la respuesta

ante emergencias sanitarias y contribuye al bienestar global (Organización Mundial de la Salud, 2021).

4. Objetivos

4.1. Objetivo principal

Desarrollar un sistema multiagente de IA explicable para question answering biomédico multimodal que integre capacidades avanzadas de procesamiento textual y visual, combinando grandes modelos de lenguaje, transformers visuales multimodales y técnicas de explicabilidad para aportar en el desarrollo de herramientas de apoyo clínico interpretables.

4.2. Objetivos secundarios

- Objetivo secundario 1: Diseñar e implementar una base de datos vectorial completa con más de 30 millones de artículos PubMed utilizando embeddings especializados (BiomedNLP-PubMedBERT) con el fin de establecer una infraestructura RAG robusta que permita búsqueda semántica de alta precisión y recuperación de evidencia científica actualizada.

- Objetivo secundario 2: Implementar un sistema de explicabilidad que combine múltiples técnicas, como GradCAM, Attention Maps e Integrated Gradients, con mapeo concepto-región y generación automática de bounding boxes con el objetivo de aportar en el desarrollo de transparencia en IA médica multimodal.

- Objetivo secundario 3: Validar la efectividad del sistema mediante una evaluación modular diferenciada en los desafíos BioASQ Task 13b (módulo textual) e ImageCLEFmed Caption 2025 (módulo visual) para demostrar buen desempeño a nivel internacional en ambas modalidades.

5. Marco teórico

El desarrollo de sistemas de question answering biomédico multimodal requiere la convergencia de múltiples disciplinas técnicas que han evolucionado significativamente en la última década (Jin et al., 2022). El procesamiento de lenguaje natural biomédico enfrenta desafíos únicos derivados de la complejidad terminológica y la necesidad de precisión extrema en contextos clínicos. Paralelamente, los avances en grandes modelos de lenguaje (Chen et al., 2025) han revolucionado las capacidades de comprensión y generación de texto, mientras que las técnicas de recuperación aumentada por generación (RAG) han emergido como soluciones fundamentales para mitigar alucinaciones y proporcionar acceso a información actualizada (Gao et al., 2025).

En el dominio visual, la interpretación automática de imágenes médicas presenta desafíos específicos relacionados con la variabilidad anatómica, los artefactos técnicos y la necesidad crítica de contextualización clínica. Los modelos multimodales recientes, particularmente los `Large Vision-Language Models` (LVLMs), han demostrado capacidades prometedoras para integrar información textual y visual, aunque su aplicación en medicina requiere consideraciones especiales de explicabilidad y transparencia. La explicabilidad constituye un requisito fundamental en aplicaciones médicas, donde la interpretabilidad de las decisiones automatizadas determina la viabilidad de adopción clínica. Las técnicas de visualización de activaciones neuronales, atribución de características y segmentación explicativa proporcionan mecanismos para entender el razonamiento de modelos complejos. Finalmente, la evaluación rigurosa mediante `benchmarks` especializados como BioASQ (Nentidis et al., 2025) e ImageCLEFmedical (Damm & et al., 2025) establece estándares de desempeño y comparabilidad que guían el desarrollo de sistemas clínicamente viables.

5.1. Procesamiento de lenguaje natural biomédico

El Procesamiento de Lenguaje Natural Biomédico (BioNLP) es un campo fundamental que busca automatizar la extracción, curación y síntesis de conocimiento a partir de la vasta y creciente literatura biomédica (Chen et al., 2025). Su importancia radica en la necesidad de superar los desafíos que impone el volumen masivo de publicaciones, como los aproximadamente 5000 artículos que se añaden diariamente solo a PubMed (Chen et al., 2025). A continuación, se detallan los desafíos específicos de este dominio, el papel del Reconocimiento de Entidades Nombradas Biomédicas (BioNER) y la integración del Sistema Unificado de Lenguaje Médico (UMLS).

5.1.1. Desafíos específicos del dominio médico

El dominio biomédico presenta obstáculos inherentes para el PLN, que van más allá de los encontrados en campos generales:

Volumen y naturaleza dinámica

La literatura biomédica es inmensamente voluminosa y se expande continuamente con nuevos descubrimientos, lo que lleva a un problema constante de palabras fuera del vocabulario (OOV) (Chen et al., 2025; Song et al., 2021). Por ejemplo, en marzo de 2024, PubMed ya contenía más de 36 millones de artículos, con unas 10 mil publicaciones adicionales al mes solo sobre COVID-19 (Chen et al., 2025).

Terminología compleja y ambigüedad

- Las Entidades Nombradas Biomédicas (BioNEs) suelen estar compuestas por varias palabras, como "hereditary nonpolyposis colorectal cancer syndrome"(Song et al., 2021).
- Una misma entidad puede ser referida con múltiples términos; por ejemplo, "Long COVID"tiene hasta 763 términos diferentes (Chen et al., 2025).
- Un mismo término puede describir entidades distintas, como .ªP2", que puede referirse a un gen, una sustancia química o una línea celular (Chen et al., 2025).
- Las abreviaturas también pueden tener múltiples significados, como "B"para "bacillus.º "whole blood"(Song et al., 2021).
- Existen entidades en cascada, donde una entidad se incrusta dentro de otra, como "HTLV-I"dentro de "HTLV-I-infected cord blood lymphocytes"(Song et al., 2021).

Necesidad de alta precisión

Dada la naturaleza crítica de la información médica, cualquier inconsistencia, información faltante o alucinación en las salidas de los modelos de PLN puede tener implicaciones significativas, requiriendo una validación manual exhaustiva (Chen et al., 2025).

Datos etiquetados limitados

A diferencia de los dominios generales, la disponibilidad de conjuntos de datos biomédicos etiquetados es notablemente menor. Esto dificulta las estrategias de ajuste fino (fine-tuning), ya que los modelos ajustados en datos limitados pueden carecer de generalización y el etiquetado manual requiere personal profesional altamente cualificado, lo que es costoso y consume mucho tiempo (Chen et al., 2025; Zhuang et al., 2024).

Inconsistencias en la anotación

Incluso en los "gold-standard corpora"(GSC), pueden existir diferencias significativas en las anotaciones entre expertos, tanto en la tipificación como en los límites de las entidades, lo que afecta la precisión del modelo (Song et al., 2021).

5.1.2. BioNER y entity linking

El Reconocimiento de Entidades Nombradas Biomédicas (BioNER) es una tarea fundamental en la extracción de información, cuyo objetivo es identificar BioNEsb (como genes, proteínas, enfermedades, sustancias químicas y especies) en la literatura biomédica no estructurada (Song et al., 2021).

El proceso de BioNER implica la preparación de conjuntos de datos adecuados, la extracción de características de la entidad y la clasificación de las entidades candidatas. Mientras que los enfoques tradicionales se basan en diccionarios y reglas, sufriendo del problema OOV, las metodologías de aprendizaje profundo han logrado un rendimiento de vanguardia en BioNER (Song et al., 2021).

Modelos de deep learning

Incluyen redes neuronales simples (como CNN y LSTM para capturar características locales y globales), el aprendizaje multitarea (que comparte parámetros entre diferentes tareas de BioNER o combina BioNER con tareas relacionadas como la normalización de entidades nombradas, NEN) y el aprendizaje por transferencia (Song et al., 2021).

Modelos preentrenados

Ejemplos como BioBERT, BioELMo y HunFlair, preentrenados en grandes corpus biomédicos, han demostrado una mejora significativa en la extracción de características y la convergencia del modelo para tareas específicas de BioNER (Song et al., 2021).

5.1.3. UMLS y su integración

El Sistema Unificado de Lenguaje Médico (UMLS), desarrollado y mantenido por la Biblioteca Nacional de Medicina (NIH), es una herramienta fundamental en la informática biomédica (Jing, 2021a). Fue diseñado para integrar numerosos vocabularios y estándares ampliamente utilizados en el campo biomédico, facilitando la interoperabilidad y la comprensión semántica entre diferentes sistemas (Jing, 2021a). UMLS se compone de tres fuentes de conocimiento clave:

- **Metatesauro:** Contiene aproximadamente 4,4 millones de conceptos y 16 millones de nombres de conceptos únicos de 218 vocabularios de origen en 25 idiomas (Jing, 2021a).
- **Red semántica:** Proporciona una categorización consistente para todos los conceptos del UMLS (Jing, 2021a).
- **Léxico SPECIALIST y herramientas léxicas:** Ofrecen herramientas sintácticas para normalizar cadenas y variantes léxicas (Jing, 2021a).

La integración de UMLS en el PLN ha sido un área muy activa, siendo uno de los tres temas más frecuentes en las publicaciones revisadas (Jing, 2021a). Sus aplicaciones incluyen:

Reconocimiento y extracción de conceptos

Permite la identificación de eventos adversos a medicamentos, propiedades contextuales y trastornos (Jing, 2021a).

Reconocimiento de entidades nombradas (NER)

Crucial para identificar términos médicos en texto libre (Jing, 2021a).

Reconocimiento y extracción de relaciones

Fundamental para descubrir interacciones fármaco-fármaco, relaciones enfermedad-tratamiento y asociaciones entre enfermedades (Jing, 2021a).

Herramientas léxicas

El léxico `SPECIALIST` del UMLS actúa como una base léxica para las aplicaciones de PLN (Jing, 2021a).

Además, UMLS ha demostrado ser eficaz en la aumentación de datos para modelos de aprendizaje profundo en BioNLP. El método UMLS-EDA (UMLS-based Easy Data Augmentation) incorpora el conocimiento de UMLS para mejorar significativamente el rendimiento de modelos de aprendizaje profundo para BioNER y clasificación, especialmente en escenarios de escasez de datos de entrenamiento (Kang et al., 2021). Este enfoque ha permitido que modelos como LSTM-CRF superen a sus contrapartes basadas en BERT en ciertas tareas (Kang et al., 2021).

En resumen, UMLS es un pilar que proporciona la base terminológica y semántica necesaria para abordar la complejidad del lenguaje biomédico, impulsando el desarrollo de soluciones de PLN más precisas y robustas para la extracción de conocimiento crítico.

5.2. Fundamentos de los grandes modelos de lenguaje (LLMs)

La aplicación de modelos avanzados en el ámbito biomédico representa un avance significativo.

5.2.1. Arquitectura Transformer

La arquitectura Transformer, propuesta en el artículo Attention is `All you Need` (Vaswani et al., 2017), revolucionó los modelos de transducción de secuencias al basarse exclusivamente en mecanismos de atención, prescindiendo de la recurrencia y las convoluciones. Su importancia radica en su capacidad para paralelizar significativamente la computación durante el entrenamiento, reduciendo drásticamente los tiempos necesarios. Además, permite el modelado de dependencias a larga distancia con un número constante de operaciones secuenciales, superando las limitaciones de los modelos recurrentes y convolucionales en el aprendizaje de estas dependencias. Esto ha sido crucial para su éxito en tareas como la traducción automática, donde logró resultados de vanguardia (Vaswani et al., 2017).

5.2.2. LLMs

Los Grandes Modelos de Lenguaje (LLMs), ejemplificados por modelos avanzados como `GPT-4o`, demuestran una inteligencia más general a través de diversas capacidades (Bubeck et al., 2023; Wei et al., 2022):

- **Capacidades de razonamiento:** `GPT-4`, por ejemplo, exhibe un dominio del lenguaje y puede resolver tareas complejas en matemáticas y codificación. En el ámbito biomédico, GPT-4 superó el 80 % de precisión en las pruebas del US Medical Licensing Exam y puede generar código Python para evaluar el riesgo de diabetes basándose en datos del paciente.
- **Generación de texto:** Estos modelos producen texto de alta calidad, a menudo indistinguible del generado por humanos, y pueden adaptarlo a diferentes estilos y contextos (Bubeck et al., 2023; Wei et al., 2022).
- **Papel en Q&A (Preguntas y Respuestas):** `GPT-3` ya mostraba un fuerte rendimiento en tareas de Q&A, siendo competitivo o incluso superando a modelos

previamente ajustados en ciertos escenarios (Bubeck et al., 2023). `GPT-4`, además, puede actuar como "juez" para determinar la similitud semántica entre respuestas, mejorando la evaluación de la veracidad y el razonamiento (Bubeck et al., 2023). Su habilidad para usar herramientas externas, como motores de búsqueda, le permite obtener información actual y responder preguntas complejas (Bubeck et al., 2023).

El sistema `MultimodalBioQA` utiliza LLMs como `GPT-4o` para responder a preguntas biomédicas complejas, aprovechando su capacidad para comprender y generar texto en contextos especializados.

5.2.3. LLMs en biomedicina

Los grandes modelos de lenguaje (LLMs) han demostrado un potencial significativo en el dominio biomédico, donde la comprensión del lenguaje natural y la extracción de información son cruciales. Sin embargo, generalmente, los LLMs de propósito general no están optimizados para el lenguaje clínico y biomédico, que posee una terminología y una base de conocimientos muy específicas, lo que hace que los LLMs de propósito general a menudo requieran adaptación para un rendimiento óptimo. Las estrategias de adaptación incluyen:

- Pre-entrenamiento continuado: Entrenar adicionalmente un LLM general sobre grandes corpus biomédicos (p. ej., PubMed, EHRs desidentificados). Ejemplos notables incluyen BioBERT (basado en BERT, pre-entrenado en PubMed) (Lee et al., 2020), ClinicalBERT (pre-entrenado en notas clínicas MIMIC-III), PubMedBERT, BioMegatron, GatorTronGPT (basado en GPT, entrenado en EHRs y texto general) y, más recientemente, BioMistral (basado en Mistral, pre-entrenado en PubMed Central).
- Fine-tuning específico: Ajustar un LLM (general o biomédico) en conjuntos de datos específicos de tareas clínicas (p. ej., para respuesta a preguntas médicas, extracción de entidades).
- `Instruction tuning médico`: Ajustar LLMs utilizando conjuntos de datos de instrucciones y respuestas específicas del dominio médico. (Dettmers et al., 2023).
- Aplicaciones clínicas: Los LLMs adaptados se aplican a una amplia gama de tareas en biomedicina, como respuesta a preguntas clínicas (Q&A), extracción de información (NER, extracción de relaciones), resumen de textos médicos (literatura, notas clínicas), generación de diálogos médico-paciente, clasificación de textos, apoyo a la decisión clínica y análisis de datos de EHR. Las tendencias recientes muestran una adaptación creciente de los LLMs generales potentes (como Llama, Mistral, GPT) al dominio biomédico, aprovechando sus capacidades emergentes de razonamiento y aprendizaje few-shot (Chen et al., 2025).

Desafíos pendientes

Estas capacidades, si bien generales, tienen un inmenso potencial para transformar la atención médica y la investigación biomédica, aunque es crucial abordar las limitaciones como las "alucinaciones" y los sesgos inherentes en los datos de entrenamiento (Bubeck et al., 2023).

El sistema `MultimodalBioQA` utiliza modelos adaptados al dominio biomédico, como PubMedBERT, para mejorar la precisión y la relevancia de las respuestas a preguntas biomédicas complejas.

5.3. Prompt engineering y Few-Shot learning

El aprendizaje basado en prompts representa un cambio de paradigma en NLP, donde, a diferencia del aprendizaje supervisado tradicional, se reformulan las tareas para que se parezcan más a las que se resuelven durante el entrenamiento original del LLM, con la ayuda de un prompt textual (P. Liu et al., 2023). Esto permite que el modelo de lenguaje, pre-entrenado con grandes cantidades de texto sin procesar, realice aprendizaje con pocos ejemplos (*few-shot learning*) o incluso sin ejemplos (*zero-shot learning*), adaptándose a nuevos escenarios con poca o ninguna data etiquetada (P. Liu et al., 2023).

La ingeniería de prompts o *prompt engineering* es el proceso de diseñar una función de prompting que modifique la entrada original en un prompt textual con espacios sin llenar, para que el modelo de lenguaje los complete probabilísticamente (P. Liu et al., 2023). La ingeniería de prompts se centra en la formulación óptima de instrucciones textuales para lograr que los modelos de lenguaje comprendan de manera precisa las tareas solicitadas.

Los componentes clave del *prompt engineering* incluyen: Ingeniería de Plantillas de Prompts (`Prompt Template Engineering`) e Ingeniería de Respuestas de Prompts (`Prompt Answer Engineering`).

5.3.1. Ingeniería de Plantillas de Prompts

Este enfoque se basa en el diseño de estructuras o plantillas de prompts que incorporan espacios reservados específicos destinados a ser completados por el modelo. La literatura especializada identifica dos modalidades principales:

- **Prompts de tipo Cloze** (completación de espacios): Esta metodología presenta al modelo enunciados con espacios en blanco que deben ser rellenados. Por ejemplo: "La capital de Japón es [Z]." Esta técnica resulta particularmente eficaz con modelos de tipo BERT y arquitecturas similares basadas en codificadores bidireccionales (P. Liu et al., 2023).
- **Prompts de tipo Prefix** (continuación textual): En esta modalidad, se proporciona al modelo un contexto inicial que debe completar de forma coherente. Por ejemplo: "¿Cuál es la capital de Japón? [Z]" Este enfoque se adapta mejor a modelos generativos autorregresivos como la familia GPT (P. Liu et al., 2023).

El diseño de estas plantillas puede realizarse mediante dos aproximaciones complementarias:

1. **Diseño manual**: Basado en la intuición y experiencia humana para la formulación de prompts efectivos (P. Liu et al., 2023).
2. **Diseño automatizado**: Empleando algoritmos de optimización que identifican automáticamente las formulaciones más eficaces para cada tarea específica (P. Liu et al., 2023).

En aplicaciones multimodales que integran texto e imágenes —como sistemas de análisis de imágenes médicas— se incorporan *embeddings* visuales como componentes adicionales del prompt, expandiendo así las capacidades de representación del contexto de entrada.

El diseño de un prompt apropiado es crucial, ya que especifica la tarea que el modelo debe realizar (P. Liu et al., 2023).

5.3.2. Ingeniería de Respuestas de Prompts

La ingeniería de respuestas de prompts se enfoca en la definición y estructuración de las salidas esperadas del modelo de lenguaje. Esta metodología establece parámetros específicos para el formato y la naturaleza de las respuestas generadas (P. Liu et al., 2023).

Las respuestas pueden adoptar diferentes modalidades según los requerimientos de la tarea (P. Liu et al., 2023):

- **Respuestas de palabra única**: Como clasificaciones binarias (p. ej., "positivo", "negativo")
- **Frases concisas**: Para respuestas factuales específicas (p. ej., "capital de China")
- **Oraciones completas**: Para tareas que requieren explicaciones detalladas

Adicionalmente, es posible establecer esquemas de mapeo que vinculen las respuestas textuales con categorías o valores numéricos específicos. Por ejemplo, la respuesta "positivo" puede mapearse a la clase 1 en un sistema de clasificación binaria.

El diseño de estas estructuras de respuesta puede implementarse mediante dos enfoques:

1. **Diseño manual**: Elaboración de taxonomías y listas de respuestas válidas basadas en conocimiento experto
2. **Optimización automatizada**: Evaluación sistemática de diferentes formatos de respuesta para identificar las configuraciones más efectivas

5.3.3. Aprendizaje Few-Shot mediante Prompts

El aprendizaje few-shot constituye una técnica de condicionamiento contextual donde se proporciona un número limitado de ejemplos demostrativos dentro del propio prompt, permitiendo al modelo inferir patrones y generalizar a nuevas instancias (Gupta et al., 2021; P. Liu et al., 2023).

Ejemplo ilustrativo:

> "La capital de Francia es París. La capital de Alemania es Berlín. La capital de Japón es [Z]."

A partir de estos ejemplos demostrativos (*shots*), el modelo puede deducir correctamente que la respuesta apropiada es "Tokio", aplicando el patrón identificado en los casos precedentes (P. Liu et al., 2023).

Esta metodología, también denominada *demonstration learning* o aprendizaje por demostración, ha demostrado particular eficacia en modelos de gran escala, donde la capacidad de generalización a partir de contextos limitados resulta especialmente pronunciada (Gupta et al., 2021; P. Liu et al., 2023).

El sistema **MultimodalBioQA** representa una aplicación práctica de las técnicas de ingeniería de prompts en el dominio biomédico. Este sistema integra información textual e imagenológica médica para resolver consultas complejas en un contexto biomédico.

La implementación utiliza técnicas de prompt engineering, que incorporan metodologías de few-shot learning, cuidadosamente diseñadas con ejemplos demostrativos específicos del dominio para orientar las respuestas del modelo. Esta aproximación ilustra la

aplicabilidad y efectividad de estas técnicas incluso en dominios altamente especializados y técnicamente exigentes como la biomedicina, donde la precisión y la coherencia contextual son requisitos críticos.

5.4. Recuperación Aumentada por Generación (RAG)

La Generación Aumentada por Recuperación (RAG) es un paradigma crucial que combina las fortalezas de la recuperación de información y los grandes modelos de lenguaje (LLMs) generativos para abordar sus limitaciones intrínsecas (W. Zhang & Zhang, 2025). Los LLMs a menudo producen "alucinaciones", es decir, respuestas inconsistentes o sin sentido, debido a su dependencia de parámetros fijos y datos de entrenamiento potencialmente desactualizados (W. Zhang & Zhang, 2025). RAG mitiga esto al recuperar información externa, actualizada y específica del dominio (memoria no paramétrica) y proporcionarla como contexto al LLM, fundamentando así sus respuestas y haciéndolas más factuales, veraces y confiables (Amugongo et al., 2025; W. Zhang & Zhang, 2025).

Un componente técnico central de RAG implica el uso de embeddings y bases de datos vectoriales (Amugongo et al., 2025). Tanto las consultas de usuario como los documentos externos se transforman en representaciones numéricas densas llamadas embeddings mediante modelos codificadores pre-entrenados (W. Zhang & Zhang, 2025). Estas representaciones vectoriales de alta dimensión capturan el significado semántico. Los embeddings de documentos se almacenan e indexan en bases de datos vectoriales (por ejemplo, utilizando Qdrant) (W. Zhang & Zhang, 2025), lo que permite una búsqueda eficiente del Producto Interior Máximo (MIPS) para encontrar los documentos más relevantes según su similitud con el embedding de la consulta (Lewis et al., 2020; W. Zhang & Zhang, 2025).

En un contexto clínico, RAG es particularmente impactante, ya que la atención médica es un dominio intensivo en conocimiento que exige información precisa y actual (Amugongo et al., 2025). RAG aborda las limitaciones críticas de los LLMs en medicina, como la generación de contenido inexacto o el uso de conocimiento obsoleto (Amugongo et al., 2025). Al recuperar información de bases de datos médicas externas, repositorios de literatura o sistemas expertos, los LLMs basados en RAG pueden proporcionar respuestas más precisas, completas, factuales y seguras a preguntas clínicas (Amugongo et al., 2025). Esto mejora capacidades como la selección de pacientes para ensayos clínicos, la identificación de criterios de inclusión/exclusión, y el razonamiento diagnóstico, especialmente cuando se aumenta con grafos de conocimiento (Amugongo et al., 2025).

El sistema **MultimodalBioQA** utiliza una base vectorial local Qdrant (Öztürk & Mesut, 2024) alimentada con embedding de documentos de PubMed (Liang et al., 2021). También se guarda la información recuperada de la ontología y grafo de conocimiento UMLS (Bodenreider, 2004) que se utiliza para enriquecer los prompts de preguntas biomédicas. Además, también se guardan los ejemplos de preguntas y respuestas "gold"que se proporcionan al sistema con técnica de few-shot prompting. El sistema permite la recuperación de información relevante y actualizada, mejorando la precisión y la relevancia de las respuestas generadas por el modelo.

5.5. Grafos de conocimiento (KG) y ontologías médicas en IA

Los Grafos de conocimiento (KGs) son representaciones estructuradas del conocimiento que modelan entidades (nodos) y las relaciones entre ellas (aristas). Las ontologías proporcionan un vocabulario formal y una estructura taxonómica para un dominio específico. En el ámbito biomédico y sanitario, los KGs y las ontologías son herramientas poderosas para (Chandak et al., 2023; Nicholson & Greene, 2020; Y. Zhang et al., 2025):

- Organizar información compleja: Integrar datos heterogéneos de diversas fuentes (literatura, bases de datos clínicas, EHRs) en un modelo unificado.
- Habilitar el razonamiento: Permitir inferencias sobre las relaciones entre conceptos médicos (p. ej., inferir interacciones medicamentosas, identificar factores de riesgo).
- Mejorar la comprensión semántica: Proporcionar definiciones y relaciones estandarizadas para términos médicos, facilitando la interoperabilidad y la interpretación consistente.
- Apoyar aplicaciones clínicas: Servir como base para sistemas de apoyo a la decisión clínica, descubrimiento de fármacos, fenotipado de enfermedades y análisis predictivo.

Recursos biomédicos clave

- UMLS (Unified Medical Language System): Es un compendio masivo de vocabularios y estándares biomédicos que actúa como un metatesauro, conectando conceptos de más de 200 fuentes, incluyendo SNOMED CT, MeSH, ICD, RxNorm, LOINC, etc. Su objetivo es promover la interoperabilidad. Aunque su cobertura es extensa (más de 3,7 millones de conceptos), puede que las vistas de los conceptos no sean uniformes. Se utiliza ampliamente en NLP biomédico y como base para construir HKGs (Healthcare Knowledge Graphs) (Bodenreider, 2004; Jing, 2021b).
- SNOMED CT (Systematized Nomenclature of Medicine - Clinical Terms): Es considerada la ontología clínica más completa, basada en lógica formal, que cubre un amplio espectro de conceptos clínicos (diagnósticos, procedimientos, síntomas, etc.). Su adopción en EHRs está creciendo, impulsada por estándares como FHIR. SNOMED CT se utiliza para estandarizar la documentación clínica, apoyar la extracción de información de texto libre (normalización de conceptos) y, cada vez más, como fuente de conocimiento para modelos de IA. A pesar de su potencial, se necesita más evidencia sobre sus beneficios clínicos directos y persisten desafíos en su implementación. (Gaudet-Blavignac et al., 2021; Vuokko et al., 2023)
- Otras ontologías/KGs: Existen numerosas ontologías y KGs específicos de dominio (p. ej., Gene Ontology, Human Phenotype Ontology, KGs de enfermedades como SPOKE para esclerosis múltiple, KGs construidos a partir de literatura, etc.) (Ashburner et al., 2000; Robinson et al., 2008).

Integración de KGs con LLMs/RAG (Graph-RAG)

La combinación de KGs y LLMs es un área de investigación muy activa que busca aprovechar las fortalezas complementarias de ambos.

Los LLMs, aunque potentes en lenguaje, a menudo carecen de conocimiento factual específico, son propensos a alucinaciones y difíciles de interpretar. Los KGs ofrecen conocimiento estructurado, verificable y explícito que puede anclar a los LLMs. Además, la estructura del grafo permite modelar relaciones complejas que pueden ser difíciles de capturar solo con texto (Matsumoto et al., 2024).

Paradigmas de integración

La integración de KGs y LLMs puede clasificarse en varias categorías (Gao et al., 2025; Matsumoto et al., 2024; Soman et al., 2024; Zhao et al., 2025):

- KG-enhanced LLMs: Usar KGs para mejorar el pre-entrenamiento o el fine-tuning de LLMs, o para proporcionar conocimiento durante la inferencia (p. ej., vía RAG).
- LLM-augmented KGs: Utilizar LLMs para ayudar en la construcción, completitud o validación de KGs a partir de texto u otras fuentes.
- Synergized LLMs + KGs: Frameworks donde LLMs y KGs colaboran de forma más estrecha y bidireccional.
- GraphRAG: Este término se refiere específicamente al uso de KGs como la base de conocimiento externa en un sistema RAG. El flujo de trabajo típico implica:
- Construcción/Indexación del Grafo: Crear o utilizar un KG relevante (p. ej., extrayendo entidades y relaciones de documentos fuente, posiblemente con ayuda de LLMs) e indexarlo para la recuperación.
- Recuperación Guiada por Grafo: Dada una consulta, identificar entidades relevantes en el KG y recuperar información estructuralmente conectada (nodos vecinos, tripletas, subgrafos completos). Esto puede implicar atravesar el grafo.
- Generación Aumentada por Grafo: Utilizar la información estructurada recuperada del grafo (y potencialmente el texto asociado a nodos/aristas) para aumentar el prompt del LLM y generar la respuesta final.

El sistema **MultimodalBioQA** utiliza un sistema Graph-RAG basado en UMLS que aprovecha la estructura semántica del grafo de conocimiento médico para enriquecer prompts biomédicos. El sistema navega por las relaciones conceptuales de UMLS mediante su API, almacena la información recuperada en una base de datos vectorial para optimizar futuras consultas y presenta los resultados al usuario con plena transparencia.

5.6. Inteligencia artificial multimodal en biomedicina

La IA Multimodal en el contexto biomédico se refiere a modelos de inteligencia artificial que integran y procesan múltiples tipos de datos (p. ej., imágenes médicas, notas clínicas, datos de EHR, genómica, datos tabulares, gráficos) para mejorar el rendimiento en tareas clínicas como el diagnóstico, pronóstico o planificación del tratamiento. El objetivo es lograr una comprensión más holística, similar a la práctica clínica humana donde los expertos integran información de diversas fuentes. Se distingue de la IA multicanal, que utiliza múltiples entradas del mismo tipo de datos (p. ej., diferentes secuencias de RM) (Simon et al., 2024; Kumar et al., 2024).

Los enfoques multimodales son beneficiosos porque diferentes modalidades pueden proporcionar información complementaria, cooperativa o redundante, lo que potencialmente conduce a predicciones más robustas, precisas y generalizables en comparación con los enfoques unimodales. La propia naturaleza de los datos clínicos y la toma de decisiones médicas es intrínsecamente multimodal.

Sin embargo, el aprendizaje multimodal presenta desafíos fundamentales (Schouten et al., 2025):

- Representación: Cómo transformar datos de diferentes modalidades en formatos legibles por máquina (p. ej., vectores) preservando las relaciones y el contexto entre ellas.
- Fusión: Cómo combinar eficazmente las representaciones de múltiples modalidades en un modelo predictivo.
- Alineación: Cómo alinear automáticamente datos de diferentes modalidades, espacial o temporalmente (p. ej., señales ECG y PPG, o modalidades de imagen).
- Traducción: Cómo mapear datos de una modalidad a otra (p. ej., generar texto a partir de una imagen médica).
- Co-aprendizaje: Cómo transferir conocimiento aprendido de una modalidad para mejorar el aprendizaje en otra, especialmente útil cuando una modalidad tiene datos limitados o ruidosos.

Además de estos desafíos teóricos, existen obstáculos prácticos como la taxonomía inconsistente en la literatura, la escasez de datos multimodales a gran escala, representativos y bien anotados, el riesgo de sesgos en los datos existentes y la brecha entre las prácticas de recopilación de datos clínicos y los formatos requeridos para el desarrollo en IA (Simon et al., 2024).

5.6.1. Estrategias de fusión multimodal

La fusión multimodal se define como el proceso de combinar información o representaciones derivadas de diferentes modalidades dentro de un modelo de IA. Las estrategias de fusión se clasifican comúnmente según la etapa del procesamiento en la que ocurre la integración:

- Fusión temprana (Early Fusion): Combina datos en bruto o características de bajo nivel extraídas de cada modalidad cerca de la capa de entrada del modelo. Esta estrategia puede requerir una alineación cuidadosa de los datos y puede ser sensible a información irrelevante para la tarea específica (Boulahia et al., 2021; Stahlschmidt et al., 2022).
- Fusión tardía (Late Fusion): Combina las predicciones o salidas generadas por modelos unimodales separados y entrenados en cada modalidad (p. ej., mediante ensamblaje de modelos). Aunque conceptualmente simple, puede tener limitaciones en la explotación de correlaciones complejas entre modalidades durante el proceso de aprendizaje (Boulahia et al., 2021; Stahlschmidt et al., 2022).

- Fusión intermedia/conjunta/híbrida (Intermediate/Joint/Hybrid Fusion): Integra características extraídas de diferentes modalidades en capas intermedias del modelo, permitiendo una interacción más profunda entre ellas durante el procesamiento. Los mecanismos de atención, especialmente la atención cruzada (cross-attention) dentro de los Transformers, son particularmente adecuados para facilitar este tipo de fusión, permitiendo que las representaciones de diferentes modalidades interactúen y se influencien mutuamente. La fusión también puede clasificarse como basada en el modelo (model-based), donde la arquitectura está diseñada específicamente para la fusión, o agnóstica al modelo (model-agnostic), donde la fusión ocurre antes o después del modelo principal (Boulahia et al., 2021; Guarrasi et al., 2025; Stahlschmidt et al., 2022).

Las arquitecturas Transformer, gracias a mecanismos como la atención cruzada y la atención multicabeza, facilitan naturalmente estrategias de fusión intermedia sofisticadas que van más allá de la simple concatenación o el promedio en etapas tardías (Vaswani et al., 2017; Zhou et al., 2023). La atención cruzada, diseñada originalmente para relacionar la salida del codificador con la entrada del decodificador, puede adaptarse para relacionar secuencias derivadas de diferentes modalidades (p. ej., características de imagen y embeddings de texto) (Bi et al., 2024). Esto permite una ponderación e integración dinámica y dependiente del contexto de la información de diferentes fuentes dentro de las capas ocultas del modelo, representando un enfoque de fusión más potente en comparación con los métodos estáticos tempranos o tardíos (Xie et al., 2025). Esta ventaja arquitectónica es probablemente un impulsor clave para la adopción de Transformers en tareas biomédicas multimodales complejas (Al-hammuri et al., 2023).

La transición hacia el procesamiento multimodal es una tendencia dominante en la IA médica, reflejando la naturaleza multifacética de los datos clínicos (Xiao et al., 2025; Yin et al., 2024). El éxito de estos sistemas depende críticamente de la elección de codificadores específicos para cada modalidad (como ViT para imágenes) y de estrategias de fusión sofisticadas que permitan una integración significativa de la información heterogénea (Li et al., 2024).

5.7. Grandes modelos de lenguaje y visión (LVLMs)

Los grandes modelos de lenguaje y visión (LVLMs, por sus siglas en inglés), como LLaVA, aprovechan el poder de modelos de lenguaje como LLaMA para generar datos multimodales de instrucciones a partir de pares imagen-texto, empleando representaciones simbólicas como captions y bounding boxes (H. Liu et al., 2023). LLaVA conecta un codificador visual (por ejemplo, CLIP) con un modelo de lenguaje (por ejemplo, Vicuna) y los ajusta de manera conjunta de extremo a extremo, habilitando así la comprensión visual y lingüística de propósito general. Esto facilita capacidades como chat multimodal y clasificación zero-shot mediante la similitud imagen-texto (Huang et al., 2021; H. Liu et al., 2023).

5.7.1. Categorización de LLaVA-LLaMA según estrategias de fusión

El modelo LLaVA-LLaMA utilizado en este proyecto se clasifica como un modelo de fusión intermedia híbrida (H. Liu et al., 2023; Yin et al., 2024). Su arquitectura emplea una estrategia de fusión intermedia donde:

- Las características visuales se extraen mediante un codificador de visión congelado (CLIP ViT-L/14).
- Un proyector MLP de dos capas actúa como conector de modalidades (H. Liu et al., 2023).
- El modelo de lenguaje (LLaMA/Vicuna) procesa las representaciones fusionadas para generar salidas textuales.
- Esta arquitectura permite la interacción profunda entre modalidades sin requerir entrenamiento conjunto desde cero (H. Liu et al., 2023).

El sistema **MultimodalBioQA** utiliza LLaVA-LLaMA como modelo de lenguaje y visión para responder preguntas visuales biomédicas complejas, aprovechando su capacidad para integrar texto e imágenes de manera efectiva. La fusión intermedia permite al modelo generar respuestas más precisas y contextualizadas, combinando la información visual de las imágenes médicas con el conocimiento textual de los documentos biomédicos.

5.8. Visión computacional médica

El captioning de imágenes médicas (MIC, por sus siglas en inglés) es un área en evolución dentro de la inteligencia artificial que integra la visión por computador y el procesamiento del lenguaje natural para comprender y describir automáticamente imágenes médicas, abordando así la generación manual de informes por parte de los radiólogos, un proceso que resulta laborioso y propenso a errores (Beddiar et al., 2023).

El grounding visual médico es fundamental en este campo, especialmente para patologías sutiles que ocupan pequeñas porciones de la imagen (Huang et al., 2021). Frameworks como GLoRIA aprenden representaciones globales y locales al contrastar subregiones de la imagen con palabras en los informes emparejados, enfatizando áreas significativas mediante pesos de atención, sin requerir detectores de objetos preentrenados (Huang et al., 2021).

Un modelo relevante en esta área incluye el Segment Anything Model (SAM), un modelo fundacional para la segmentación de imágenes. SAM está diseñado para ser promptable, lo que permite la transferencia zero-shot a diversas tareas como la detección de bordes, la generación de propuestas de objetos y la segmentación de instancias, mediante el uso de prompts flexibles (por ejemplo, puntos, cajas, texto) (Kirillov et al., 2023).

Por otra parte, Grad-CAM proporciona explicaciones visuales para redes neuronales convolucionales (CNNs) al generar mapas de localización discriminativos por clase a partir de gradientes, destacando las regiones importantes de la imagen para una predicción particular. Grad-CAM es aplicable a una amplia variedad de arquitecturas CNN, incluidas aquellas empleadas en captioning de imágenes y visual question answering, sin requerir cambios arquitectónicos (Huang et al., 2021; Selvaraju et al., 2020).

A pesar de estos avances, el captioning médico enfrenta desafíos significativos. A diferencia del captioning de imágenes genéricas, en el ámbito médico es necesario captar relaciones complejas entre objetos de la imagen y hallazgos clínicos, requiriéndose alta precisión y el uso de terminología médica estructurada y precisa, así como la correcta identificación de hallazgos anómalos poco frecuentes (Beddiar et al., 2023). Entre las principales limitaciones se encuentran la escasez, el tamaño reducido, el desbalance de clases y la heterogeneidad de los datasets médicos anotados de alta calidad, además de las preocupaciones relativas a la privacidad. Los modelos también presentan dificultades

para manejar oraciones largas, el orden de las palabras y la localización precisa de anomalías sutiles (Beddiar et al., 2023). Asimismo, las métricas automáticas actuales suelen no captar matices semánticos, y la obtención de evaluaciones humanas costosas por parte de expertos médicos sigue siendo una barrera.

El sistema **MultimodalBioQA** utiliza técnicas de captioning de imágenes médicas para generar descripciones precisas y contextualizadas de las imágenes biomédicas. La integración de técnicas como SAM permiten al sistema identificar y resaltar áreas relevantes en las imágenes médicas, facilitando una comprensión más profunda y precisa de los hallazgos clínicos.

5.9. Question answering multimodal

5.9.1. Fundamentos de sistemas de Q&A multimodal

Los sistemas de Preguntas y Respuestas Visuales (VQA) surgieron como un problema interdisciplinario que demanda conocimientos tanto de la Visión por Computadora (CV) como del Procesamiento del Lenguaje Natural (NLP) (Gupta et al., 2021). En VQA, la tarea consiste en responder preguntas formuladas sobre una imagen, donde el sistema debe aprender y generar respuestas basándose en las características extraídas de la imagen de entrada (Gupta et al., 2021). A diferencia de las tareas típicas de CV, que se centran en problemas como la identificación de acciones o la clasificación de imágenes, las tareas de VQA son inherentemente más complejas, requiriendo una inteligencia superior que incluye el reconocimiento de objetos, la extracción de características semánticas, el conocimiento externo y el sentido común (Gupta et al., 2021).

Tradicionalmente, en NLP, los sistemas de aprendizaje supervisado predicen una salida y a partir de una entrada x utilizando un modelo $P(y|x;\theta)$ (P. Liu et al., 2023). Sin embargo, la disponibilidad insuficiente de datos anotados para muchas tareas ha impulsado el desarrollo de métodos de aprendizaje basados en prompts. Estos métodos utilizan modelos de lenguaje (LM) que modelan directamente la probabilidad del texto $P(x;\theta)$ para predecir y, reduciendo o eliminando la necesidad de grandes conjuntos de datos etiquetados (P. Liu et al., 2023).

El avance en CV y NLP durante la última década ha introducido técnicas de aprendizaje automático que han mejorado la eficiencia en la resolución de problemas como la detección de objetos, la segmentación y la clasificación de imágenes en CV, así como la traducción automática y los sistemas de preguntas y respuestas en NLP (Gupta et al., 2021). La combinación de estas áreas es fundamental para la VQA multimodal.

5.9.2. Estado del arte en integración texto + visión en medicina

El dominio médico es uno de los campos donde la VQA desempeña un papel crucial al proporcionar asistencia médica a los usuarios finales (Gupta et al., 2021). La VQA en el dominio médico (VQA-Med) es una tarea compleja y de gran importancia, ya que los usuarios pueden plantear preguntas sencillas con respuestas de "Sí/No" o preguntas desafiantes que requieren una respuesta detallada y descriptiva. Esta diversidad de tipos de preguntas, provenientes de distintos usuarios como pacientes, estudiantes de medicina, clínicos y expertos, exige enfoques específicos para cada tipo de consulta para evitar confusión y proporcionar asistencia precisa (Gupta et al., 2021).

Para abordar esta complejidad, se han propuesto arquitecturas como la red jerárquica profunda multimodal (Gupta et al., 2021) y la red pre-entrenada multinivel de fusión

VB-MVQA (Cai et al., 2023). La red jerárquica, denominada HQS-VQA (*Hierarchical Question Segregation based Visual Question Answering*), primero analiza y clasifica las preguntas/consultas de los usuarios finales y luego incorpora un enfoque específico para cada tipo de consulta para la predicción de respuestas (Gupta et al., 2021). Una técnica clave dentro de HQS-VQA es la segregación de preguntas (QS), que clasifica las preguntas en dos tipos principales: 'Sí/No' y 'Otros' (Gupta et al., 2021). Para esta segregación, se utiliza un modelo de aprendizaje automático estadístico simple, como una Máquina de Vectores de Soporte (SVM), basado en características diseñadas manualmente y en la frecuencia de palabras. Esta técnica evita la necesidad de crear sistemas distintos para diferentes tipos de preguntas, lo que podría generar confusión y malestar en el usuario final (Gupta et al., 2021).

En el modelo HQS-VQA, las representaciones de las preguntas y las imágenes se generan utilizando *Bidirectional Long Short Term Memory* (Bi-LSTM) y *Inception-Resnet-v2*, respectivamente. Estas representaciones se fusionan y se pasan a un modelo de predicción de respuestas específico en los nodos hoja de la jerarquía. Los experimentos demuestran que HQS-VQA supera a los modelos de referencia en conjuntos de datos como RAD y CLEF18, lo que sugiere la efectividad de la segregación de preguntas en VQA-Med (Gupta et al., 2021).

Por otro lado, el modelo VB-MVQA aborda la limitación de datos en VQA-Med explotando el pre-entrenamiento. Incorpora *Contrastive Language-Image Pre-training* (CLIP) y mecanismos de atención para extraer características de imágenes médicas de manera efectiva. VB-MVQA utiliza múltiples capas de atención apiladas y *Bilinear Attention Network* junto con *Bidirectional Long Short-Term Memory* (Bi-LSTM) (Cai et al., 2023). También introduce el razonamiento condicionado por la visión (*vision-conditioned reasoning*) para guiar la selección de importancia sobre características fusionadas multimodales y mejorar la información semántica de la imagen, lo que ayuda a eliminar el sesgo lingüístico. Este modelo ha demostrado mejoras significativas en la precisión sobre modelos de última generación en conjuntos de datos de referencia como VQA-RAD, SLAKE y VQA-Med-2019, especialmente en preguntas abiertas y datos con sesgo lingüístico (Cai et al., 2023).

Los conjuntos de datos para VQA-Med incluyen preguntas en lenguaje natural sobre imágenes radiográficas de diferentes modalidades como angiogramas, resonancias magnéticas, tomografías computarizadas y ultrasonidos, así como diversas orientaciones como sagital, axial, longitudinal y coronal (Gupta et al., 2021).

5.9.3. Estrategias de generación de respuestas

Respuestas exactas vs. ideales: Reglas de formato y expectativas En el contexto de desafíos de Q&A como BioASQ, que se centran en el dominio biomédico, las respuestas pueden clasificarse en "exactas" o "ideales". Esta distinción es crucial para satisfacer las diversas necesidades de información de los expertos biomédicos (Tsatsaronis et al., 2015).

- **Preguntas "Yes/No"**: Esperan una respuesta binaria de 'sí' o 'no' como respuesta "exacta". Para estas, la evaluación se realiza mediante precisión (*accuracy*). Además, se espera una respuesta "ideal", que es un resumen en forma de párrafo (Tsatsaronis et al., 2015).
- **Preguntas factoid**: Buscan una entidad nombrada específica como respuesta "exacta". La evaluación de la respuesta exacta incluye la precisión estricta (*strict accuracy*) y la precisión indulgente (*lenient accuracy*), así como el rango recíproco medio (MRR). La respuesta "ideal" es un resumen (Tsatsaronis et al., 2015).

- **Preguntas de lista**: Requieren una lista de entidades nombradas como respuesta "exacta". Los sistemas deben devolver una lista de nombres de entidades que, en conjunto, constituyan una única respuesta. La evaluación de la respuesta exacta se basa en precisión (P), *recall* (R) y F-measure (F1), promediados sobre las preguntas de lista. La respuesta "ideal" también es un resumen (Tsatsaronis et al., 2015).
- **Preguntas de resumen**: Solo esperan una respuesta "ideal" en forma de resumen (Tsatsaronis et al., 2015).

Para las respuestas "ideales", que son resúmenes de párrafo, se establecen límites de longitud (ej., 200 palabras en BioASQ). La evaluación de estas respuestas se realiza manualmente por expertos biomédicos, considerando criterios como el *recall* de información, la precisión de información, la ausencia de repetición y la legibilidad. También se utilizan medidas automáticas como ROUGE para comparar el resumen generado con resúmenes de referencia creados por humanos (Tsatsaronis et al., 2015).

5.9.4. Síntesis multimodal: Integración visual + textual con explicabilidad

La síntesis de respuestas en sistemas VQA-Med implica la fusión de características extraídas tanto de imágenes como de texto (Cai et al., 2023; Gupta et al., 2021). Modelos como HQS-VQA utilizan Bi-LSTM para representaciones de preguntas e *Inception-Resnet-v2* para características de imagen, fusionando ambas antes de la predicción de la respuesta (Gupta et al., 2021). VB-MVQA, por su parte, aplica múltiples capas de atención apiladas y *Bilinear Attention Network* para fusionar la información multimodal, introduciendo un razonamiento condicionado por la visión para guiar la selección de importancia y mejorar la información semántica de la imagen, con el fin de eliminar el sesgo lingüístico (Cai et al., 2023). La explicabilidad es crucial en el dominio médico; el análisis de errores en los sistemas VQA-Med es fundamental para descubrir las causas potenciales de los errores y sus soluciones, guiando así la investigación futura (Gupta et al., 2021).

5.9.5. Inserción de citas y evidencia científica (PMID, snippets)

En el ámbito biomédico, la provisión de evidencia científica para las respuestas es de suma importancia. Los desafíos como BIOASQ requieren que los sistemas no solo generen respuestas, sino que también recuperen y presenten documentos relevantes y *snippets* de texto que las respalden (Tsatsaronis et al., 2015).

- Los identificadores de artículos PubMed (PMID) se utilizan en los conjuntos de datos como referencia única para cada artículo (Tsatsaronis et al., 2015).
- Los *snippets* son fragmentos de texto relevantes de los artículos, identificados por el artículo del que provienen y sus *offsets* (posiciones de inicio y fin de caracteres) dentro del mismo. Estos *snippets* pueden ser utilizados por los expertos biomédicos para formular respuestas ideales y se espera que los sistemas los recuperen como parte de la evidencia (Tsatsaronis et al., 2015).

La capacidad de los sistemas para proporcionar *snippets* de texto precisos y relevantes, junto con los documentos de origen, es evaluada rigurosamente, incluso considerando la superposición de *snippets* en lugar de una coincidencia exacta para la precisión y el *recall*. Un ejemplo de una respuesta ideal en BIOASQ incluso incluye los PMIDs de los artículos de los que se extrajo la información (Tsatsaronis et al., 2015).

5.9.6. Formato estructurado

Los datos utilizados en VQA-Med a menudo tienen un formato estructurado o semi-estructurado.

- **Metadata de imágenes**: Las imágenes médicas se acompañan de preguntas en lenguaje natural y se describen con diversas modalidades (ej., angiograma, resonancia magnética, tomografía computarizada, ultrasonido) y orientaciones (ej., sagital, axial, longitudinal, coronal) (Gupta et al., 2021). Esta metadata es esencial para el procesamiento y entendimiento de las imágenes por parte del modelo.
- **Formato de datos**: Los conjuntos de datos de referencia, como los utilizados en BIOASQ, suelen seguir un formato JSON para estructurar preguntas, conceptos, documentos, respuestas exactas e ideales, *snippets* y triples. Esto permite una manipulación programática eficiente de los datos (Tsatsaronis et al., 2015).
- **Texto argumentativo en respuestas ideales**: Las respuestas "ideales" son resúmenes en forma de párrafo que combinan información de múltiples fuentes para proporcionar una respuesta concisa y comprensible. Estas respuestas deben ser coherentes, legibles y precisas, sintetizando la información recuperada de manera argumentativa para satisfacer la necesidad de información del usuario. Por ejemplo, una respuesta a "¿Cuál es el mecanismo de acción de la abiraterona?" sería un párrafo que describe cómo funciona el fármaco (Tsatsaronis et al., 2015).

En resumen, los sistemas Q&A multimodales, especialmente en el dominio médico, se benefician de la integración profunda de la información visual y textual, estrategias de prompting para el aprendizaje eficiente con pocos datos, y la generación de respuestas estructuradas que pueden ser tanto exactas como resúmenes ideales, siempre respaldadas por evidencia citada.

5.10. Explicabilidad y transparencia clínica

Según Barredo Arrieta et al., dada una cierta audiencia, la explicabilidad se refiere a los detalles y razones que un modelo proporciona para hacer que su funcionamiento sea claro o fácil de entender ((Barredo Arrieta et al., 2020)). La explicabilidad y la transparencia son imperativos fundamentales para el avance y la adopción de la inteligencia artificial (IA) y el aprendizaje automático (ML) en el sector médico. Dada la alta responsabilidad y la posible repercusión en vidas humanas, los sistemas de IA utilizados en la atención sanitaria deben justificar sus decisiones, proporcionar explicaciones claras y ser inherentemente interpretables.

5.10.1. Sistemas de explicabilidad ante-hoc y post-hoc

Las técnicas de explicabilidad se clasifican según cuándo se aplica la explicación y si son específicas del modelo.

- **Métodos ante-hoc (inherentemente explicables):** Son algoritmos diseñados desde cero para ser transparentes y comprensibles. Ejemplos incluyen modelos lineales, árboles de decisión y sistemas basados en reglas. En el contexto médico, los sistemas basados en reglas pueden proporcionar declaraciones lógicas que los clínicos pueden interpretar directamente. Los *Concept Bottleneck Models* (CBMs) son

un ejemplo de un enfoque ante-hoc, ya que están construidos para predecir conceptos intermedios explícitamente antes de la salida final. Esto permite intervenir directamente en los conceptos (p.ej., modificar el valor predicho de "espolón óseo" en una radiografía) y observar cómo esto afecta la predicción final, facilitando una interacción más rica entre el humano y el modelo.

- **Métodos post-hoc:** Se aplican a modelos de "caja negra" (como las redes neuronales profundas) después de que han sido entrenados para proporcionar una explicación de sus decisiones. Estos métodos pueden ser agnósticos al modelo (aplicables a cualquier algoritmo de ML, como LIME y SHAP) o específicos del modelo (como Grad-CAM para CNNs). Proporcionan principalmente explicaciones locales para decisiones individuales. Sin embargo, existe preocupación sobre la fiabilidad de las explicaciones post-hoc y su vulnerabilidad a ataques adversariales.

5.10.2. Activaciones neuronales: Grad-CAM y Attention Maps

La visualización de las activaciones neuronales es una técnica clave en la explicabilidad visual, a menudo presentada como mapas de calor (*heatmaps*) que resaltan las regiones de entrada importantes para la decisión de un modelo.

- **Grad-CAM** (*Gradient-weighted Class Activation Map*) es un método post-hoc de uso extendido que genera mapas de saliencia utilizando los gradientes del resultado de una clase con respecto a los mapas de activación de la última capa convolucional. Permite identificar las características discriminatorias que el modelo utilizó para una clasificación específica. Grad-CAM es aplicable a una amplia gama de arquitecturas de Redes Neuronales Convolucionales (CNN) sin requisitos arquitectónicos específicos. Las visualizaciones producidas por Grad-CAM, por ejemplo, cuando se aplican a radiografías, pueden ayudar a los sujetos humanos a identificar objetos con mayor precisión.

- **Score-CAM** es un método novedoso de explicación visual post-hoc que elimina la dependencia de los gradientes. En cambio, determina la importancia de cada mapa de activación utilizando su puntuación de "Increase of Confidence" (aumento de confianza) en la clase objetivo. Score-CAM busca mejorar el rendimiento visual y la imparcialidad al interpretar los procesos de toma de decisiones del modelo, y ha demostrado una mejor capacidad de discriminación de clases en comparación con otros métodos basados en gradientes.

- **Mecanismos de atención (Attention Maps):** Como se utilizan en la investigación, mejoran la interpretabilidad visual al permitir que la red se enfoque y adapte a la escala correcta de un objeto dentro de una imagen. Esto es crucial para que los observadores humanos comprendan si una red neuronal está identificando correctamente un objeto sin confundirlo con su entorno.

5.10.3. Atribución de características: Integrated Gradients y otros métodos

Los métodos de atribución de características cuantifican la contribución de componentes específicos de entrada a la decisión de un modelo.

- **LIME** (*Local Interpretable Model-agnostic Explanations*) es un método post-hoc que genera explicaciones al aproximar el comportamiento de un modelo de "caja

negra" localmente con un modelo interpretable más simple. Por ejemplo, para la clasificación de texto, LIME puede resaltar la importancia de los síntomas individuales para una predicción de gripe. En imágenes, puede generar superpíxeles que indican las regiones cruciales para una clasificación (p.ej., un gato).

- **Integrated Gradients** es un método de atribución basado en gradientes que asigna puntuaciones de importancia a las características de entrada. Su aplicación permite visualizar cómo un conjunto de CNNs clasifica el estado del receptor de estrógeno a partir de imágenes de resonancia magnética de mama.

- **SHAP** (*SHapley Additive exPlanations*) ofrece un enfoque unificado para interpretar las predicciones del modelo al calcular los valores de Shapley para la contribución de cada característica. Se ha utilizado para cuantificar el impacto a nivel de características individuales en los resultados de pacientes con influenza grave, proporcionando una comprensión clara de qué variables son más importantes para las predicciones del modelo.

5.10.4. Segmentación explicativa: integración de SAM con *bounding boxes* y conceptos médicos

La *Segment Anything Model* (SAM) representa un avance significativo como modelo fundacional para la segmentación de imágenes, diseñado para realizar una segmentación "promptable". SAM puede generar máscaras de segmentación válidas para cualquier prompt dado, ya sean puntos, cajas, máscaras o incluso texto. Su arquitectura separa la codificación de la imagen de la codificación del prompt y la decodificación de la máscara, permitiendo una segmentación rápida y la conciencia de la ambigüedad al predecir múltiples máscaras para un solo prompt.

- **Integración con *bounding boxes* y conceptos médicos:** Aunque SAM es un modelo de propósito general, su capacidad de segmentación a partir de *bounding boxes* es altamente relevante para aplicaciones médicas. En entornos clínicos, los expertos a menudo identifican áreas de interés (p.ej., tumores, lesiones, estructuras anatómicas) con *bounding boxes*. SAM podría transformar estas cajas en segmentaciones precisas de alta calidad, que a su vez sirven como explicaciones visuales para el diagnóstico o pronóstico, de manera similar a cómo las segmentaciones de U-Net son consideradas explicaciones visuales. Esto permite que el resultado de SAM se interprete en el contexto de conceptos médicos predefinidos por expertos, mejorando la comprensión y la confianza del médico. SAM ya ha demostrado su capacidad de transferencia de conocimiento a dominios como las imágenes de rayos X, lo que subraya su potencial en el campo médico.

5.10.5. Importancia para validación clínica y aceptación regulatoria

En el campo de la IA, se considera que, dada una audiencia, una Inteligencia Artificial explicable es aquella que proporciona detalles o razones para hacer que su funcionamiento sea claro o fácil de entender (Barredo Arrieta et al., 2020). La explicabilidad en inteligencia artificial (XAI) es indispensable para la validación clínica y la aceptación regulatoria de los sistemas de IA en medicina, por varias razones críticas:

- **Confianza y rendición de cuentas:** Los modelos de ML, especialmente los de "caja negra", carecen de la transparencia necesaria para que los médicos los confíen plenamente, lo que impide su adopción generalizada. La explicabilidad fomenta una confianza adecuada al permitir que los profesionales entiendan el "porqué" detrás de una decisión, lo que es vital en escenarios de alto riesgo donde las vidas humanas están en juego.

- **Detección de sesgos y errores:** Las explicaciones ayudan a identificar sesgos latentes o errores en el modelo, como en el caso de pacientes con asma y neumonía, donde el modelo asoció erróneamente el asma con un menor riesgo de mortalidad debido a sesgos en los datos de entrenamiento. La XAI permite exponer tales comportamientos indeseables y garantiza que las decisiones algorítmicas sean justas y éticas.

- **Validación clínica y mejora de decisiones:** Los clínicos necesitan información detallada sobre las predicciones de los modelos, incluyendo el subconjunto de características que impulsan una predicción, para compararlas con su juicio clínico y calibrar su confianza en el sistema. Las explicaciones visuales y textuales, incluso si son extrañas, requieren métodos sistemáticos para investigar y corregir el razonamiento erróneo. La XAI puede mejorar la confianza del clínico, lo que es un factor clave para la adopción práctica del modelo.

- **Cumplimiento normativo:** Regulaciones como el Reglamento General de Protección de Datos (RGPD) de la UE exigen que las decisiones automatizadas sean explicables y que los interesados tengan derecho a una intervención humana y a una explicación de la decisión. La explicabilidad se convierte en una necesidad para el cumplimiento legal y la protección de los derechos del paciente.

- **Colaboración humano-IA:** La XAI permite una colaboración más eficaz entre médicos y sistemas de IA. Los sistemas que pueden explicar sus resultados ayudan a los médicos a comprender la cadena de razonamiento, verificar la sugerencia del sistema y hacer una evaluación de la fiabilidad, lo que conduce a un mejor diagnóstico y resultados para los pacientes.

En resumen, la integración de la explicabilidad en los sistemas de IA, a través de visualizaciones, atribuciones de características y segmentación explicativa, es fundamental para superar la opacidad de la "caja negra" y garantizar que la IA se convierta en una herramienta de confianza y responsabilidad en la práctica médica.

El sistema **MultimodalBioQA** incorpora técnicas de explicabilidad para proporcionar justificaciones claras y comprensibles de las respuestas generadas, facilitando la confianza y la comprobación de las respuestas del sistema. La integración de modelos como SAM permite resaltar áreas relevantes en las imágenes médicas, mejorando la interpretabilidad y la transparencia del sistema.

5.11. Sistemas multiagente para procesamiento biomédico

El paradigma agéntico en Inteligencia Artificial (IA), a menudo denominado .ªgentic AI", se refiere a sistemas de IA diseñados para operar con un alto grado de autonomía, capaces de percibir su entorno, tomar decisiones, planificar y ejecutar acciones complejas

para alcanzar objetivos a largo plazo con mínima intervención humana continua (xiRise-PotentialLarge2023a).

El paradigma agéntico se refiere al desarrollo de sistemas de IA (a menudo basados en LLMs) que pueden actuar de forma autónoma para alcanzar objetivos complejos. Estos agentes pueden razonar, planificar, descomponer tareas, utilizar herramientas (como APIs, bases de datos, buscadores web) y aprender de la interacción (xiRisePotentialLarge2023a, wangSurveyLargeLanguage2024).

Estado del arte del paradigma agéntico

Se están desarrollando agentes que pueden interactuar con el mundo, realizar tareas complejas y aprender de la experiencia. Estos agentes pueden ser simples (realizando tareas específicas) o complejos (capaces de razonar y planificar) (Park et al., 2023). A continuación, se presentan algunos de los enfoques y desarrollos más destacados en este campo:

• **LLMs como controladores**: Los LLMs se utilizan como el çerebro.º controlador central de los agentes, aprovechando su capacidad de comprensión del lenguaje natural y razonamiento para interpretar objetivos, generar planes y seleccionar acciones (Xi et al., 2023; Yao et al., 2023).

• **Planificación y uso de herramientas**: Arquitecturas como ReAct (Reasoning and Acting) (Yao et al., 2023) y marcos como LangChain o AutoGen (Wu et al., 2023) facilitan la creación de agentes que pueden interactuar con herramientas externas. Esto les permite superar las limitaciones inherentes del conocimiento estático del LLM y realizar acciones en el mundo real (o digital).

• **Memoria y aprendizaje**: Se está investigando cómo dotar a los agentes de memoria a corto y largo plazo para mejorar su rendimiento en tareas extendidas y aprender de experiencias pasadas (Xi et al., 2023). • Sistemas Multi-Agente: Hay un interés creciente en sistemas donde múltiples agentes colaboran o compiten para resolver problemas más complejos, simulando dinámicas sociales o de equipo (Park et al., 2023; Wu et al., 2023).

• **Aplicaciones potenciales**: Aunque aún en desarrollo, los agentes de IA tienen potencial en automatización de tareas complejas, asistencia personal, investigación científica (ej. formulación de hipótesis, diseño de experimentos), y potencialmente en tareas clínicas (ej. monitorización de pacientes, gestión de información) (Wang et al., 2023).

5.12. Benchmarks y evaluación

5.12.1. BioASQ: Tareas y métricas de evaluación

BioASQ es una serie de retos internacionales que promueven avances en la indexación semántica biomédica a gran escala y la respuesta a preguntas (Nentidis et al., 2023, 2025; Tsatsaronis et al., 2015). El desafío BioASQ sirve como punto de referencia a largo plazo para la respuesta a preguntas biomédicas, con ediciones recientes (2023-2024) que demuestran un notable progreso en el campo. Los sistemas que compiten abordan distintos retos en varias fases. La fase A se centra en la recuperación de información de artículos de PubMed (Novoa et al., 2023) para identificar fragmentos relevantes para preguntas concretas. En la fase A+, los participantes deben responder a varios tipos de preguntas (sí/no, factoide, lista, resumen) recuperando la bibliografía pertinente de PubMed y proporcionando tanto respuestas exactas como respuestas ideales en forma de párrafo. La fase B requiere que los participantes proporcionen respuestas exactas e ideales utilizando fragmentos de texto proporcionados.

La disponibilidad del conjunto de datos BioASQ curado manualmente, que contiene miles de pares pregunta-respuesta y fragmentos estándar de oro, sigue permitiendo estos avances (Krithara et al., 2023a, 2023b).


```
            "document": "https://www.ncbi.nlm.nih.gov/pubmed/39985914",
            "text": "Unlocking the code: The role of molecular and genetic profiling in
            revolutionizing glioblastoma treatment.",
            "beginSection": "title",
            "endSection": "title",
            "offsetInBeginSection": 0,
            "offsetInEndSection": 106,
            "pmid": "39985914",
            "relevance_score": 0.6743988666242392
        }
    ],
    "exact_answer": "yes",
    "ideal_answer": "Yes, PTEN loss in GBM does contribute to resistance against
    immunotherapy. Glioblastoma (GBM) is recognized as the most aggressive primary
    brain cancer, exhibiting significant molecular and cellular heterogeneity. This
    heterogeneity, including genetic alterations such as PTEN loss, plays a crucial
    role in the tumor's resistance mechanisms against various therapies, including
    immunotherapy (PMID 39985914). Furthermore, the extensive metabolic reprogramming
    observed in GBM, which is influenced by genetic changes like PTEN loss, further
    drives the tumor's growth and its resistance to therapeutic interventions (PMID
    39859381). This metabolic reprogramming complicates the immune landscape, making
    the tumor environment more hostile to immune system interventions and thereby
    diminishing the effectiveness of immunotherapeutic strategies. Thus, PTEN loss not
```


Figura 1: Entregable de la tarea 13b Fase A+ de bioASQ: extracción de snippets y rankeo,respuesta exacta y respuesta ideal en formato JSON.

Fase A

El sistema recibe preguntas en inglés.
Debe devolver:

- Lista de hasta 10 artículos (por relevancia).
- Lista de hasta 10 *snippets* (fragmentos, con posición).

Desde BioASQ12 hay una "Fase A+.^en la que se pide a los sistemas responder con respuestas exactas/ideales, usando solo los documentos recuperados (antes de recibir los datos *gold*).

Fase B

El sistema recibe la pregunta y la lista *gold* de artículos y *snippets*.
Debe devolver:

- Respuesta exacta (solo *yes/no*, *factoid*, *list*).
- Respuesta ideal (todos los tipos de pregunta).

Puede participar solo en una fase o en ambas.

Tipos de pregunta

- **Yes/no:** Respuesta ze.º "no". Se evalúa con F1 macro.
- **Factoid:** Respuesta con una entidad (nombre, número). Lista de hasta 5 candidatos. Evaluación principal: MRR.
- **List:** Lista de entidades (máx. 100). Se evalúa con F1.
- **Summary:** Solo respuesta ideal, párrafo resumen.

Métricas de bioASQ
Estas son las métricas con las que se evaluó el módulo de Q&A textual del sistema:

Métricas oficiales para respuestas exactas

- **Yes/No:** F1 macro-averaged (maF1): F1 para ƶesƶ F1 para "no", promediados. Métrica oficial desde BioASQ7. También se calcula *accuracy* para referencia, pero no es oficial.
- **Factoid:** MRR (*Mean Reciprocal Rank*): Oficial. Premia respuestas correctas en posiciones más altas de la lista (hasta 5 nombres). Strict Accuracy (correcta si la respuesta *gold* es la primera) y Lenient Accuracy (correcta si está en las cinco primeras) se reportan, pero no son oficiales.
- **List:** Mean F-measure (F1): Oficial. Se calcula sobre la coincidencia entre la lista retornada y la lista *gold* (sin sinónimos repetidos). También se calculan mean precision y mean recall como métricas de apoyo.

Métricas oficiales para respuestas ideales
Las respuestas ideales son evaluadas tanto manual como automáticamente:

- **Manual:** Cuatro criterios, todos con escala 1–5:

  1. Recuperación de información: ¿Incluye toda la información relevante?
  2. Precisión de información: ¿Evita información irrelevante?
  3. Repetición: ¿Evita repetir la misma información?
  4. Legibilidad: ¿Es fácil de leer y fluida?

  La puntuación oficial es la media de las valoraciones manuales de los expertos biomédicos.
- **Automática:** ROUGE-2 y ROUGE-SU4 (superposición de bigramas y *skip-bigramas*), comparando la respuesta generada con las *gold* (o *snippets*).

5.12.2. ImageCLEFmedical: Tareas y métricas de evaluación

El reto ImageCLEFmedical Caption 2025 se compone de tres tareas interconectadas: Detección de Conceptos, Predicción de Leyendas y Explicabilidad (Damm & et al., 2025; Ionescu, 2025). En la tarea de detección de conceptos, los sistemas identifican la presencia de conceptos médicos relevantes en una imagen, prediciendo eficazmente un conjunto de ID de conceptos UMLS (Unified Medical Language System) (Bodenreider, 2004) o términos que describen el contenido de la imagen.

Esto sirve de base para el caption, ya que proporciona los "bloques de construcción"de la escena. En la tarea de predicción de captions, los sistemas generan una descripción textual coherente de toda la imagen, que idealmente incorpora los conceptos detectados y

describe su interacción. La tarea de explicabilidad requiere que los participantes den una explicación del caption en un pequeño subconjunto de imágenes, por ejemplo, resaltando regiones de la imagen y proporcionando una justificación textual adicional. El componente de explicabilidad pretende mejorar la interpretabilidad y la confianza, permitiendo a los expertos médicos verificar por qué se predijo un caption o un concepto.

Tareas Las tareas del reto ImageCLEFmedical Caption 2025 son:

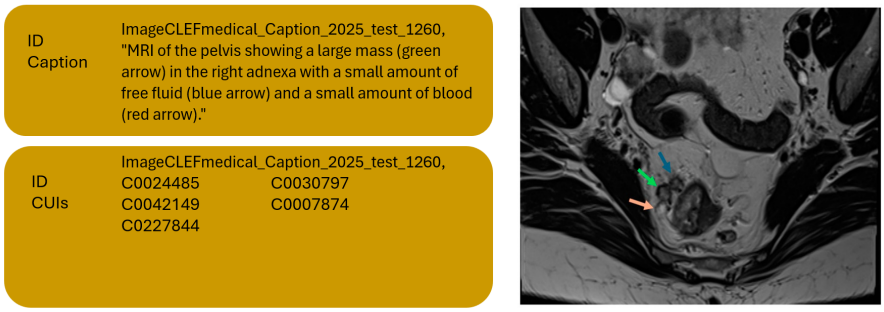


Figura 2: Ilustración del formato esperado del entregable la tarea de Caption y Concept. Entregable real incluía CUIs y captions de más de 19 mil imágenes en formato .JSON.

Tarea de detección de conceptos

El primer paso para el *captioning* automático de imágenes y la comprensión de escenas es identificar la presencia y localización de conceptos relevantes en un gran corpus de imágenes médicas. Basándose en el contenido visual de la imagen, esta subtarea proporciona los elementos fundamentales para la etapa de comprensión de escenas, identificando los componentes individuales a partir de los cuales se componen los *captions*. Además, los conceptos pueden aplicarse para la recuperación de imágenes e información basada en el contexto.

La evaluación se realiza en términos de métricas de cobertura de conjuntos, tales como precisión, *recall* y combinaciones de ambas.

Tarea de predicción de captions

Sobre la base del vocabulario de conceptos detectados en la primera subtarea, así como la información visual de su interacción en la imagen, los sistemas participantes deben componer *captions* coherentes para la totalidad de una imagen. En este paso, más allá de la mera cobertura de conceptos visuales, la detección de la interacción entre los elementos visibles es crucial para lograr un buen desempeño.

Este año, el desafío utiliza **BERTScore** como métrica principal de evaluación y **ROUGE** como métrica secundaria para la subtarea de predicción de *captions*. También se publicarán otras métricas, como **MedBERTScore**, **MedBLEURT** y **BLEU**.

Tarea de explicabilidad

Además, se solicita a los participantes que proporcionen explicaciones para los *captions* de un pequeño subconjunto (que será publicado junto con el *dataset* de prueba) de imá-

genes. No existen limitaciones técnicas para esta tarea. Las explicaciones serán evaluadas manualmente por un radiólogo en cuanto a interpretabilidad, relevancia y creatividad.

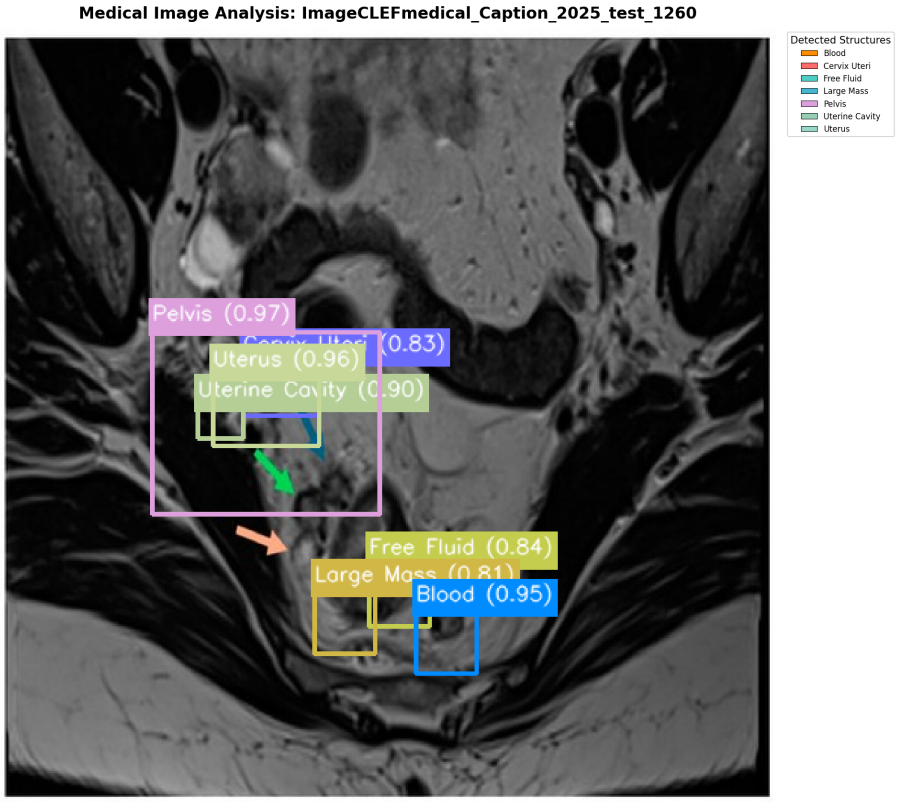


Figura 3: Entregable de la tarea de explicabilidad. Detecta los objetos en la imagen y los localiza con bounding boxes.

Métricas de evaluación Cada tarea del desafío ImageCLEF medical caption utiliza métricas de evaluación distintas (Damm & et al., 2025; Ionescu, 2025).

A. Detección de conceptos

La tarea de detección de conceptos emplea la metodología de puntuación F1 con el siguiente enfoque (Damm & et al., 2025; Ionescu, 2025):

- Implementación: Método de puntuación F1 de scikit-learn de Python (v0.17.1-2) con promedio binario.
- Proceso: Arreglos binarios que indican la presencia (1) o ausencia (0) del concepto, tanto para las predicciones como para los valores de referencia (ground truth).
- Estructura de puntuación: La puntuación primaria considera todos los conceptos y la secundaria filtra únicamente aquellos conceptos anotados manualmente.
- Ground truth: Basado en un subconjunto reducido de la versión AB 2022 de UMLS utilizado en los datos de entrenamiento.

B. Evaluación de la predicción de captions

La evaluación de la predicción de captions combina seis métricas agrupadas en dos aspectos: Métricas de relevancia (cuatro métricas) y métricas de factualidad (dos métricas) (Damm & et al., 2025; Ionescu, 2025).

B1. Métricas de relevancia

Estas métricas incluyen la evaluación de similitud entre imagen y caption, así como métricas de relevancia basadas en texto, como BERTScore (recall), ROUGE-1 (medida F) y BLEURT.

- Similitud entre imagen y caption: utiliza un modelo de embeddings para imágenes médicas que evalúa la relación semántica entre los captions generados y sus imágenes correspondientes. Este enfoque calcula los embeddings tanto del texto del caption como de la imagen médica, y luego mide la similitud entre ellos para determinar qué tan bien el caption representa el contenido visual.
- BERTScore (recall): utiliza el modelo Microsoft/deberta-xlarge-mnli con ponderación por frecuencia inversa de documentos (IDF).
- ROUGE-1 (medida F): mide la superposición de unigramas entre los captions generados y los de referencia.
- BLEURT: emplea el checkpoint BLEURT-20 para la evaluación basada en juicio humano.

B2. Métricas de factualidad

Las siguientes métricas se utilizan en el reto ImageCLEFmedical Caption para evaluar la factualidad de los captions generados (Damm & et al., 2025; Ionescu, 2025):

- UMLS Concept F1: evalúa la precisión médica de los captions generados extrayendo entidades médicas mediante la herramienta MedCAT y calculando la puntuación F1 entre los conceptos UMLS predichos y los de referencia.
- AlignScore: proporciona una evaluación de la consistencia factual mediante una implementación basada en RoBERTa, que evalúa la alineación de la información entre los textos generados y los de referencia. El proceso implica dividir los contextos largos en fragmentos manejables y hacer coincidir sistemáticamente cada frase con afirmaciones con el fragmento de contexto más relevante para determinar la precisión factual. El resultado final representa una puntuación media de alineación calculada en todas las frases con afirmaciones, proporcionando una medida integral de cuán bien el caption generado mantiene la consistencia factual respecto al material de referencia.

C. Explicabilidad

En la tarea de explicabilidad, un radiólogo experto evaluó la calidad de las explicaciones generadas por cada sistema utilizando una escala Likert de 5 puntos, donde 5 representaba la máxima puntuación. La evaluación incluyó tanto el caption generado como la visualización acompañante para cada imagen, considerando múltiples categorías de evaluación. El radiólogo valoró la calidad del caption a través de cuatro dimensiones:

legibilidad, idoneidad clínica, nivel de detalle y enfoque, que se consolidaron en una puntuación media del caption. La calidad de la visualización se evaluó de forma similar en tres aspectos: coherencia visual-textual, exhaustividad de la visualización y enfoque de la visualización, lo que resultó en una puntuación media de visualización. Además, el experto valoró la idoneidad de la metodología y asignó una calificación global, creando un marco de evaluación integral que recogía tanto la calidad de los componentes individuales como el rendimiento holístico del sistema (Damm & et al., 2025; Ionescu, 2025).

6. Metodología

La investigación se enmarcó en la estrategia de investigación **diseño y creación** (Oates, 2006), enfocada en el desarrollo de un artefacto tecnológico innovador como principal contribución. Siguiendo esta metodología, se diseñó e implementó iterativamente un sistema multimodal basado en una arquitectura multiagente (con un agente central "escritor") que integra procesamiento textual y visual. El módulo textual implementa una búsqueda híbrida combinando una base de datos vectorial con un modelo de lenguaje biomédico (PubMedBERT) y consultas en tiempo real a PubMed, empleando además GPT-4 para extraer fragmentos relevantes. El módulo visual emplea un modelo de visión-lenguaje (LLaVA-LLaMA 8B) ajustado mediante un adaptador LoRA para análisis de imágenes médicas. Como innovación principal, el sistema integra un mecanismo de explicabilidad que combina Grad-CAM, mapas de atención e **Integrated Gradients** con el modelo **Segment Anything (SAM)** para mapear conceptos médicos a regiones de las imágenes.

Para evaluar el artefacto desarrollado, se siguió la recomendación de Oates (Oates, 2006) de realizar una validación rigurosa: primero se evaluó cada módulo por separado, y luego se llevaron a cabo pruebas experimentales con el sistema integrado, utilizando *benchmarks* internacionales (BioASQ(Nentidis et al., 2025) e ImageCLEFmed (Damm & et al., 2025)) para medir su desempeño en tareas de búsqueda biomédica y análisis de imágenes clínicas. El sistema logró buenos resultados en general, ya que está dentro de los top 10 en cinco de seis tareas, lo que evidencia la eficacia de la solución propuesta y demuestra la contribución válida del artefacto, cumpliendo con los criterios de rigor y relevancia de la metodología de diseño y creación.

6.1. Datos

Corpus de BioASQ

Se utilizarán los datasets y otros recursos proporcionados por BioASQ, pues se participó en desafíos de NLP que permiten el acceso a datos curados por expertos biomédicos (Krithara et al., 2023a). BioASQ es un proyecto de investigación que busca avanzar en el estado del arte de la recuperación de información biomédica y la comprensión del lenguaje natural. Proporciona un corpus de datos anotados para evaluar sistemas de IA en tareas como la respuesta a preguntas biomédicas, la extracción de información y la generación de resúmenes. El corpus incluye preguntas formuladas por expertos, respuestas correctas y documentos relevantes, lo que permite entrenar y evaluar modelos de IA en un contexto biomédico realista (Nentidis et al., 2023, 2025).

El dataset de prueba consistía en un fichero .JSON que contenía 80 preguntas para

responder por lote.

Corpus de CLEF

Se utilizó una versión ampliada del conjunto de datos Radiology Objects in Context (ROCO) Versión 2 en las tareas del reto ImageCLEFmedical Caption. Como en ediciones anteriores, el conjunto de datos procede de artículos biomédicos del subconjunto PMC OpenAccess (Damm & et al., 2025; Ionescu, 2025). También utilizamos el conjunto de datos ROCOv2 del año anterior para el fine-tuning (Rückert et al., 2024). El conjunto de entrenamiento consta de 79.789 imágenes radiológicas (principalmente radiografías, tomografías computarizadas, resonancias magnéticas, etc.) recogidas de la literatura biomédica, cada una emparejada con un caption y un conjunto de etiquetas de concepto UMLS. Los conjuntos de datos del reto se describen con más detalle en el documento oficial del reto (Damm & et al., 2025; Ionescu, 2025).

El dataset de prueba del reto ImageCLEFmedical Caption 2025 contiene 19.267 imágenes radiológicas, cada una con un caption y un conjunto de etiquetas de concepto UMLS, además del diccionario de CUIs que traduce los códigos CUIs a lenguaje natural.

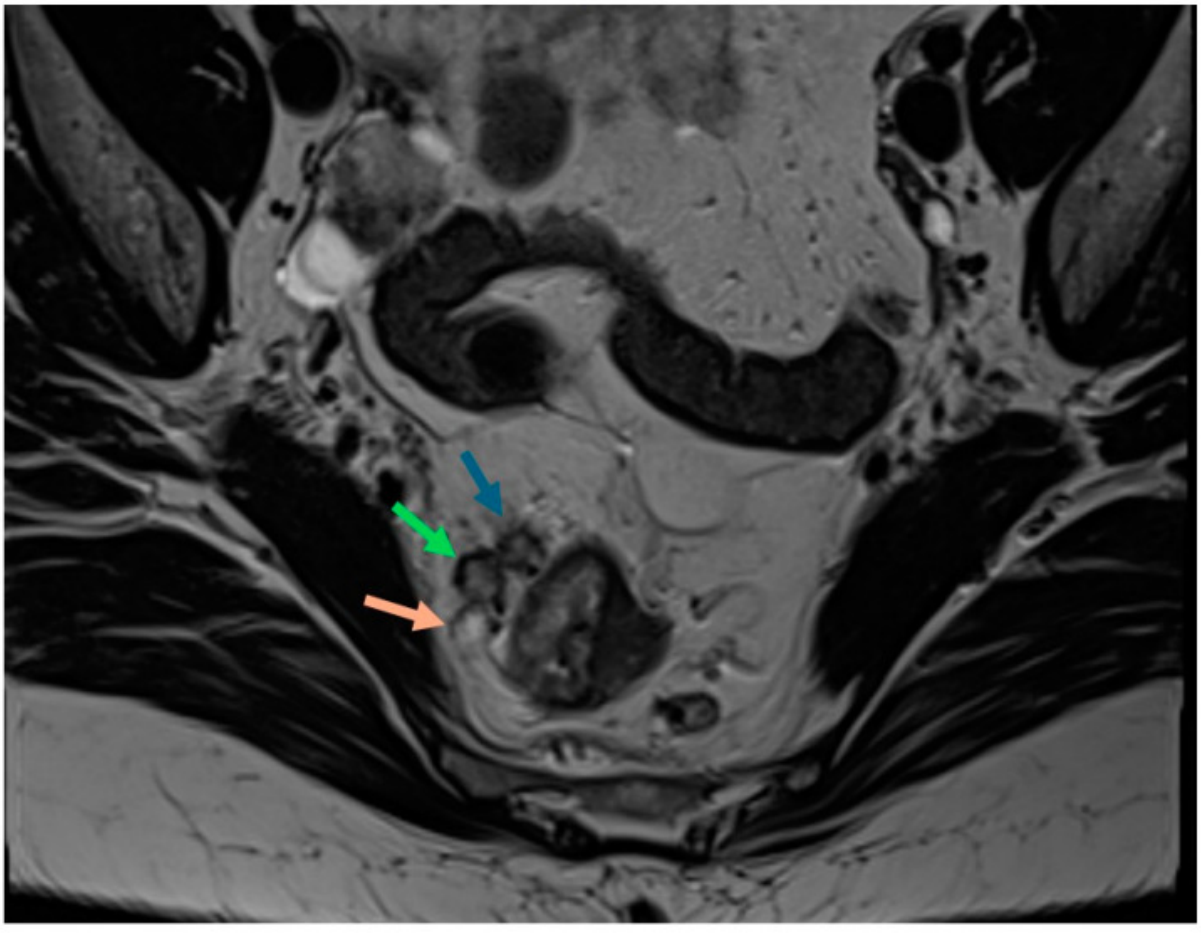


Figura 4: Imagen de muestra del conjunto de pruebas del Desafío ImageCLEF. Fuente: ImageCLEFmedical Caption 2025 test 1260, CC BY, Curcean et al., 2024.

6.2. Métricas de evaluación

Para evaluar el rendimiento del sistema se utilizarán las métricas específicas de cada tarea según se explicó anteriormente en secciones 5.12.1 y 5.12.2.

6.3. Configuración experimental

Para la ejecución de este diseño experimental se utilizó una combinación de dos configuraciones de hardware distintas:

- Una estación de trabajo de escritorio equipada con un procesador Intel Core i9 y una unidad de procesamiento gráfico (GPU) NVIDIA GeForce RTX 5090 con 32 GB de memoria GDDR7 dedicada. Este equipo se empleó para todas las tareas computacionales con altos requerimientos de GPU, como fine-tuning, carga en base de datos Qdrant (por ser una base de datos local) y el desarrollo del sistema.
- Un ordenador portátil, también equipado con un procesador Intel Core i9, que incluye una GPU NVIDIA GeForce RTX 4070 con 8 GB de memoria. Este portátil se utilizó para análisis y preprocesamiento de archivos, además de la redacción del trabajo.

La asignación de estos recursos se basó en las demandas computacionales y la compatibilidad del sistema.El grado exacto de utilización de cada GPU varió en función de las demandas específicas de cada experimento.

6.4. Arquitectura agéntica del sistema

6.4.1. Fundamentos de sistemas multiagente en aplicaciones biomédicas

El sistema desarrollado se fundamenta en principios de arquitectura multiagente, definida como un paradigma computacional donde múltiples entidades autónomas (agentes) colaboran para resolver problemas complejos que exceden las capacidades de los sistemas monolíticos (Wooldridge, 2009). En el contexto biomédico, los sistemas multiagente han demostrado particular efectividad para manejar la complejidad inherente del dominio médico, donde los diferentes aspectos del procesamiento de información requieren de un expertise especializado (Isern et al., 2010).

Según la taxonomía de Ferber (Ferber, 1998), el sistema de Q&Aimplementado corresponde a una arquitectura multiagente *cooperativa* donde los agentes comparten objetivos comunes y colaboran para maximizar la utilidad global del sistema. Esta clasificación se alinea con los requisitos de coherencia y precisión necesarios en aplicaciones biomédicas (Croatti et al., 2020).

6.4.2. Clasificación y tipología de agentes implementados

Siguiendo la taxonomía de Russell y Norvig (Stuart & Peter, 2021), los agentes implementados pueden clasificarse como *agentes basados en objetivos* (*goal-based agents*) que mantienen representaciones internas del estado del problema y utilizan conocimiento específico del dominio para tomar decisiones. Específicamente, el sistema **Multimodal-BioQA** implementa cinco tipos de agentes especializados:

Agente de Investigación (Researcher Agent) Implementa un patrón de *agente de información* según la clasificación de Genesereth y Nilsson (Genesereth & Nilsson, 2014), especializado en recuperación y síntesis de conocimiento científico. Este agente integra las siguientes capacidades:

- **Búsqueda híbrida**: Combinación de búsqueda vectorial local (Qdrant) con consultas en tiempo real a APIs externas (PubMed).
- **Extracción automática**: Utilización de GPT-4o para extracción de snippets a nivel de oración.
- **Enriquecimiento contextual**: Integración de terminología médica estandardizada (UMLS/MeSH).

Agente de Escritura (Writer Agent) Es agente funciona como un *agente coordinador* según la taxonomía de Durfee y Lesser (Durfee & Lesser, 1991) y es responsable de la síntesis final de información proveniente de múltiples fuentes especializadas. Implementa estrategias de generación diferenciadas basadas en el tipo de consulta (*yes/no, factoid, list, summary*) y gestiona la integración de evidencia multimodal.

Agente de Visión (Vision Agent) Constituye un *agente perceptual* especializado (Murphy, 2019) que procesa información visual mediante modelos multimodales avanzados (LLaVA-LLaMA 3 8B). Este agente encapsula funcionalidades de:

- Generación de conceptos médicos a partir de imágenes.
- Producción de captions descriptivos especializados.
- Integración con sistemas de explicabilidad visual.

Agente de Explicabilidad (Explainability Agent) Implementa un *agente de monitoreo y explicabilidad* (Mueller et al., s.f.) que es responsable de generar interpretaciones comprensibles de las decisiones del sistema. Integra múltiples técnicas de explicabilidad como GradCAM, Attention Maps e Integrated Gradients con mapeo concepto-región.

Agente de Chatbot interactivo (Chatbot Agent) Funciona como un *agente de interfaz* (Bradshaw, 1997) que gestiona la interacción con usuarios mediante validación interactiva de tipos de pregunta y confirmación de parámetros de consulta.

6.4.3. Arquitectura de comunicación y coordinación

Patrón de comunicación El sistema MulmodalBioQA implementa un patrón de comunicación *jerárquico con coordinación central* según la clasificación de Stone y Veloso (Stone & Veloso, 2000). La comunicación inter-agente se realiza mediante estructuras de datos estandarizadas que actúan como *mensajes estructurados* (Finin et al., 1994), garantizando coherencia semántica y trazabilidad de información.

La arquitectura de comunicación sigue el modelo *Contract Net Protocol* adaptado (Smith, 1980), donde:

1. El **Router de modalidad** determina el tipo de procesamiento requerido.

2. Los **agentes especializados** ejecutan tareas específicas de manera autónoma.

3. El **Writer Agent / Agente de Escritura** actúa como el coordinador central e integra resultados.

Gestión de estados y persistencia Cada agente mantiene un estado interno independiente según el principio de *autonomía local* (Jennings, 1999).

6.4.4. Herramientas y frameworks de implementación

LlamaIndex como framework agéntico La implementación utiliza LlamaIndex como framework base para la construcción de agentes, aprovechando sus capacidades de:

- **Gestión de embeddings**: Configuración centralizada mediante `Settings.embed_model`.
- **Integración de herramientas**: Conexión continua con APIs externas y bases de datos vectoriales.
- **Orquestación de flujos**: Coordinación de pipelines complejos de procesamiento.

Componentes tecnológicos especializados El sistema integra componentes tecnológicos especializados que actúan como *herramientas agénticas* según el concepto de *tool-using agents* definido por Schick et al. (Schick et al., 2023):

- **Qdrant**: Base de datos vectorial para búsqueda semántica distribuida.
- **BiomedNLP-PubMedBERT**: Modelo de embedding especializado en dominio biomédico.
- **APIs externas**: PubMed, UMLS, OpenAI GPT-4o.
- **Modelos multimodales**: LLaVA-LLaMA 3 8B con adaptadores LoRA.

Esta aproximación se alinea con el paradigma de *augmented agents* propuesto por Mialon et al. (Mialon et al., 2023), donde los agentes extienden sus capacidades mediante el uso sistemático de herramientas externas especializadas.

6.4.5. Características de autonomía y especialización

Autonomía operacional Según los criterios de Wooldridge y Jennings (Wooldridge & Jennings, 1995), el sistema demuestra las cuatro características fundamentales de agentes inteligentes:

1. **Autonomía**: Cada agente opera independientemente sin control directo externo.
2. **Reactividad**: Los agentes responden a cambios en su entorno operativo.
3. **Pro-actividad**: Los agentes toman iniciativa para cumplir objetivos específicos.
4. **Sociabilidad**: Los agentes interactúan mediante protocolos de comunicación estructurados.

Especialización por dominio La arquitectura implementa *especialización funcional* (Tambe, 1997), donde cada agente se enfoca en aspectos específicos del problema:

- **Separación de responsabilidades**: Cada agente maneja un aspecto único del procesamiento.
- **Expertise específico**: Agentes optimizados para tareas particulares (NLP, visión, síntesis).
- **Tolerancia a fallos**: Degradación elegante cuando agentes individuales fallan.

6.4.6. Validación del paradigma agéntico

Cumplimiento de criterios arquitectónicos El sistema satisface los criterios establecidos por Ferber (Ferber, 1998) para sistemas multiagente efectivos:

1. **Distribución**: Procesamiento distribuido entre agentes especializados
2. **Interacción**: Comunicación estructurada mediante protocolos definidos
3. **Coordinación**: Mecanismo central de síntesis y coherencia
4. **Organización**: Jerarquía clara con roles y responsabilidades definidos

Ventajas de la aproximación agéntica La implementación agéntica proporciona beneficios específicos para aplicaciones biomédicas complejas (Isern et al., 2010):

- **Escalabilidad modular**: Agentes pueden ser modificados independientemente.
- **Mantenibilidad**: Separación clara de responsabilidades facilita debugging.
- **Extensibilidad**: Nuevos agentes pueden agregarse sin afectar funcionalidad existente.
- **Robustez**: Tolerancia a fallos mediante redundancia y degradación elegante.

6.5. Diseño de arquitectura de IA

Como se justificó en el apartado anterior, este sistema implementa una arquitectura de agentes distribuidos que integra capacidades de procesamiento de texto e imagen para responder consultas biomédicas complejas. La aplicación principal (`app.py`) constituye el núcleo orquestador que coordina múltiples agentes especializados mediante una interfaz de usuario desarrollada en Streamlit.

La aplicación principal implementa un Sistema de Explicabilidad Integrado que combina LLaVA con técnicas de interpretabilidad visual (GradCAM, Attention Maps, Integrated Gradients).

Arquitectura dual: El sistema maneja consultas de texto mediante búsqueda en PubMed y síntesis con GPT-4o, mientras que las consultas de imagen emplean análisis visual con mapeo concepto-región y generación automática de *bounding boxes* etiquetadas.

Innovación principal: La integración de explicabilidad multi-nivel que correlaciona conceptos médicos extraídos por LLaVA con regiones espaciales específicas de la imagen, proporcionando interpretabilidad visual comprehensiva para análisis médico.

A continuación, se presenta un diagrama de la arquitectura de alto nivel del sistema:

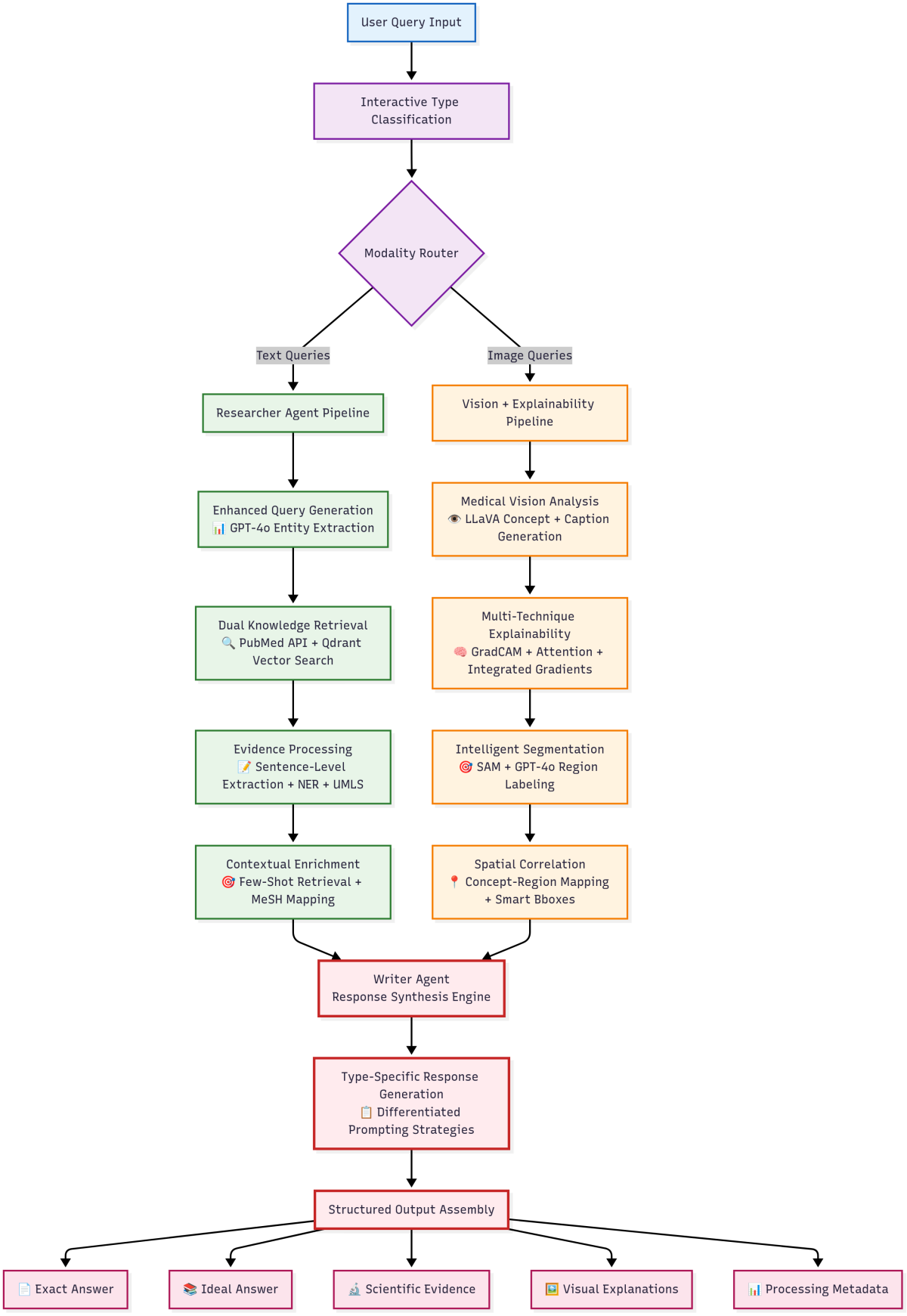


Figura 5: Diseño de arquitectura de IA para el sistema multimodal de Question answering.

6.5.1. Agentes de procesamiento especializados

Agente de Investigación (`researcher_agent.py`)

El `researcher_agent` constituye el núcleo del sistema de recuperación de información biomédica, implementando una arquitectura híbrida que combina búsqueda en bases de datos vectoriales locales con consultas en tiempo real a la API de PubMed.

Arquitectura de búsqueda híbrida

El agente implementa una estrategia de búsqueda multicapa:

```
def run_researcher(question: str, question_data: Dict) -> Dict:
```

- **Generación de consultas optimizadas:** Utiliza GPT-4o para extraer entidades médicas significativas de la pregunta mediante un método especializado. Este enfoque supera las limitaciones de los sistemas NER tradicionales al comprender el contexto semántico.
- **Búsqueda primaria en PubMed API:** Ejecuta búsquedas directas en la API de PubMed con parámetros optimizados (`hasabstract[text]`) para garantizar la disponibilidad de contenido completo.
- **Extracción de *Snippets* a nivel de oración:** Implementa una función que utiliza GPT-4o para extraer oraciones específicas que responden directamente a la pregunta, en lugar de devolver abstracts completos. Esta técnica mejora significativamente la precisión de la evidencia.

Enriquecimiento Contextual

El sistema implementa múltiples capas de enriquecimiento:

- **Integración UMLS/MeSH:** La función `_query_umls_for_context()` consulta las terminologías médicas estandarizadas para proporcionar definiciones contextuales de términos médicos identificados.
- **Recuperación *Few-Shot*:** Implementa `_retrieve_few_shot_example_from_file()` para obtener ejemplos similares que guían la generación de respuestas, mejorando la consistencia y calidad del output.
- **Ranking y filtrado:** `_filter_and_rank_snippets()` utiliza una combinación de scores de relevancia originales y matching de palabras clave para optimizar la selección de evidencia.

Optimizaciones de rendimiento

- **Manejo de Tiempo de Respuesta:** Implementa timeouts configurables y retry logic con backoff exponencial
- **Cache de Consultas UMLS:** Mantiene cache a nivel de módulo para evitar consultas repetitivas
- **Gestión de Rate Limits:** Implementa delays configurables para respetar las limitaciones de las APIs externas

Agente de Escritura (`writer_agent.py`)

El `writer_agent` representa la capa de síntesis que transforma evidencia científica cruda en respuestas coherentes y bien estructuradas.

Sistema de generación diferenciada por tipo

El agente implementa estrategias de generación específicas según el tipo de pregunta:

```
def run_writer(question: str, qtype: str, snippets: List[Dict],
    mesh_info: Optional[Dict], few_shot_example: Optional[Dict]) -> Dict:
```

- **Prompts especializados:** Cada tipo de pregunta (yesno, factoid, list, summary) utiliza prompts optimizados con instrucciones específicas para el formato de respuesta esperado.
- **Validación de respuestas:** `_validate_exact_answer()` implementa validation logic específica por tipo para asegurar conformidad con formatos esperados.
- **Síntesis dual:** Genera tanto `exact_answer` (respuesta directa) como `ideal_answer` (explicación detallada) para proporcionar información a múltiples niveles de granularidad.

Procesamiento de consultas de imagen

Para consultas multimodales, el agente implementa:

- **Detección automática de imagen Q&A:** `_is_image_qa_question()` identifica consultas relacionadas con análisis de imagen basándose en patrones lingüísticos y metadatos.
- **Prompts contextualizados para imagen:** `_build_image_qa_prompt()` crea prompts específicos que incorporan resultados del análisis visual, conceptos detectados y metadatos de explicabilidad.
- **Integración de resultados multimodales:** Combina información textual y visual para generar explicaciones completas.

Mejores prácticas de generación

- **Control de Temperatura:** Utiliza diferentes valores de temperatura según el tipo de tarea (0.0 para exact answers, 0.2–0.3 para ideal answers)
- **Gestión de Tokens:** Implementa límites adaptativos de tokens basados en la complejidad de la consulta
- **Citación Automática:** Incluye referencias automáticas a PMIDs en las respuestas generadas

Agente de Visión (`vision_agent.py`)

El `vision_agent` encapsula la funcionalidad de análisis de imagen médica utilizando modelos LLaVA fine-tuned para el dominio biomédico.

Arquitectura del modelo

```
class VisionAgent:
    def __init__(self, base_model_path, adapter_repo, device="auto",
                 load_in_4bit=True, merge_adapters=False):
```

- **Modelo base LLaVA:** Utiliza `xtuner/llava-llama-3-8b-v1_1-transformers` como modelo base con adaptadores LoRA especializados en imaging médico.
- **Optimizaciones de memoria:** Implementa quantización 4-bit opcional y merge de adaptadores para optimizar el uso de memoria GPU.
- **Gestión de dispositivos:** Detección automática y configuración de dispositivos (CUDA/CPU) con fallbacks elegantes.

Funcionalidades especializadas

- **Generación de conceptos:** `generate_concepts()` extrae conceptos médicos estructurados de imágenes utilizando prompts especializados en español para el dominio médico.
- **Generación de captions:** `generate_captions()` produce descripciones textuales detalladas de imágenes médicas siguiendo convenciones radiológicas.
- **Procesamiento natch:** Soporte para procesamiento eficiente de múltiples imágenes con gestión optimizada de memoria.

Características Técnicas Avanzadas

- **Manejo de inputs diversos:** Soporte para múltiples formatos de entrada incluyendo rutas de archivo, objetos PIL, y data URIs base64.
- **Patch size management:** Implementa corrección automática del `patch_size` para compatibilidad con diferentes versiones del procesador.
- **Gestión de memoria:** Implementa cleanup manual de tensores y cache CUDA para prevenir memory leaks durante inferencia prolongada.

Prompts Especializados

El agente utiliza prompts cuidadosamente diseñados:

```
CONCEPT_PROMPT = "USER: <image>
Enumera los conceptos médicos clave (CUIs) observados o inferidos
en esta imagen.
ASSISTANT:Los conceptos médicos clave son:"

CAPTION_PROMPT = "USER: <image>
¿Cuál es la descripción o el caption de esta imagen médica?
ASSISTANT:"
```

Estos prompts están optimizados para el dominio médico y diseñados para maximizar la precisión y relevancia de las respuestas generadas.

6.5.2. Herramientas de análisis especializado

Sistema de validación interactiva: El `chatbot_agent` implementa inferencia heurística de tipos de pregunta con confirmación interactiva, utilizando patrones especializados y auto-confirmación temporal para optimizar la clasificación de consultas.

Análisis de explicabilidad: El `explainability_agent` orquesta pipelines complejos combinando SAM para segmentación, GPT-4o para etiquetado contextual, y VisionAgent para extracción de conceptos, generando visualizaciones integrales con heatmaps y anotaciones automáticas.

NER biomédica avanzada: El `ner_tool` integra modelos transformer especializados con enriquecimiento UMLS/MeSH que implementa cache inteligente, filtrado de entidades y mapeo selectivo para optimizar la extracción de información médica.

Innovación técnica: La combinación de segmentación automática (SAM) con etiquetado contextual (GPT-4o) guiado por conceptos extraídos por LLaVA representa un enfoque innovador para explicabilidad visual en imágenes médicas.

6.6. Sistema de base de datos vectorial y gestión de datos

Ingesta de PubMed: Sistema de procesamiento streaming que maneja archivos XML.gz masivos mediante parsing incremental con `lxml.etree.iterparse`. Implementa batching multinivel para embeddings (BiomedNLP-PubMedBERT, 768D) y operaciones Qdrant, con optimizaciones específicas para ingesta masiva y gestión de memoria.

Preparación *Few-Shot*: Pipeline de optimización para GPT-4o que incluye detección automática de idioma, normalización de tipos de pregunta, y optimización de JSON para minimizar token usage. Implementa validación comprehensiva de calidad con umbral mínimo del 80 % de ejemplos válidos.

Base de datos vectorial: Utiliza Qdrant con colecciones especializadas para artículos PubMed y ejemplos *few-shot*. Implementa búsqueda semántica integrada con LlamaIndex para recuperación contextual en tiempo de ejecución.

Batching multinivel: El procesamiento streaming con batching multinivel permite manejar datasets de gran escala (millones de artículos) manteniendo uso de memoria constante, mientras que la optimización específica para GPT-4o mejora significativamente la eficiencia de token usage en *few-shot learning*.

El sistema implementa una arquitectura híbrida que combina recuperación de información vectorial con técnicas de few-shot prompting para respuestas de consultas biomédicas. La infraestructura se fundamenta en Qdrant como base de datos vectorial y utiliza modelos de embeddings especializados en el dominio biomédico.

6.6.1. Componentes de ingesta de datos

Ingesta de literatura científica (ingest_pubmed.py) Este componente constituye el núcleo del pipeline de procesamiento de literatura biomédica. El proceso implementa las siguientes etapas:

- **Preprocesamiento de datos**: Utiliza lxml.etree.iterparse para procesamiento eficiente de archivos XML.gz de PubMed, empleando análisis incremental para optimizar el uso de memoria.
- **Extracción de metadatos**: Implementa un parser robusto que extrae título, resumen, autores, fechas de publicación, términos MeSH, palabras clave y DOI. El

sistema maneja múltiples formatos de fechas incluyendo MedlineDate con normalización automática.

- **Generación de embeddings**: Emplea el modelo BiomedNLP-PubMedBERT-base-uncased-abstract-fulltext (768 dimensiones) optimizado para texto biomédico. Implementa truncamiento a 510 tokens para mantener coherencia semántica.
- **Optimizaciones de rendimiento**:
  - Batching de embeddings (512 textos por lote)
  - Batching de upserts a Qdrant (512 puntos por lote)
  - Configuración de optimizadores Qdrant para ingesta masiva
  - Detección automática de GPU/CPU para aceleración

Ingesta de ejemplos Few-Shot (ingest_few_shots.py) Este módulo gestiona la creación de una colección especializada para ejemplos de entrenamiento:

- **Procesamiento de ejemplos**: Carga ejemplos desde training13b.json y genera embeddings de las preguntas usando el mismo modelo biomédico.
- **Almacenamiento estructurado**: Mantiene el input (pregunta) y output (respuesta completa en JSON) para recuperación posterior.
- **Validación de integridad**: Implementa verificación de estructura JSON y manejo robusto de errores.

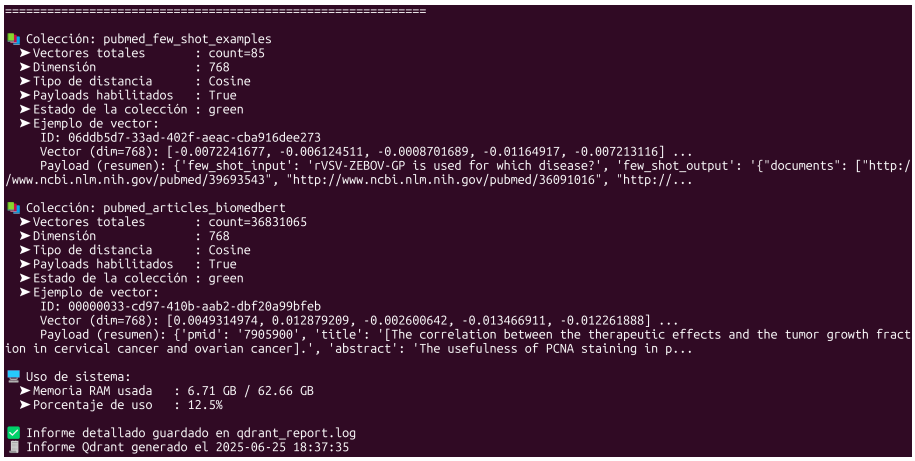


Figura 6: Colecciones de Qdrant: pubmed_articles_biomebert con más de 36 millones de vectores totales y pubmed_few_shot_examples con 85 vectores.

6.6.2. Preparación y optimización de datos

Optimización Few-Shot para GPT-4o Implementa un pipeline de optimización específico para modelos de lenguaje grandes:

- **Normalización de tipos**: Clasifica automáticamente preguntas en categorías (yesno, factoid, list, summary) con mapeo consistente.

- **Optimización de tokens**: Truncamiento automático manteniendo coherencia semántica, con límites específicos (500 chars input, 2000 chars output).
- **Validación de calidad**: Métricas de validación con umbral mínimo del 80 % de ejemplos válidos.

6.6.3. Sistema de recuperación

Utilidades de Base de Datos Vectorial (vector_store_utils.py) Proporciona la interfaz de recuperación en tiempo de ejecución:

- **Búsqueda semántica**: Utiliza el embedding del query para encontrar ejemplos similares en la colección few-shot.
- **Integración con LlamaIndex**: Se integra con Settings.embed_model para consistencia en la generación de embeddings.
- **Manejo de errores**: Implementa degradación elegante cuando los componentes no están disponibles.

6.6.4. Configuraciones de optimización

El sistema implementa múltiples optimizaciones para rendimiento en producción:

- **Gestión de memoria**: Batching adaptativo basado en recursos disponibles
- **Persistencia de estado**: Logging de archivos procesados para reinicios incrementales
- **Timeout configurations**: Configuraciones de timeout optimizadas para operaciones de red
- **Retry logic**: Implementación de Tenacity para reintentos exponenciales

6.7. Metodología de Fine-tuning

Arquitectura del Modelo y Estrategia de Adaptación.

Utilizamos `LLaVA-LLaMA-3-8B` como modelo base, implementando adaptación LoRA con *rank* $r = 16$, $\alpha = 32$ y *dropout* 0,05. Los módulos objetivo incluyen todas las proyecciones de atención (`q_proj`, `k_proj`, `v_proj`, `o_proj`), los componentes MLP (`gate_proj`, `up_proj`, `down_proj`) a lo largo de 32 capas *transformer*, y las capas lineales del proyector multimodal. Esta selección permite una adaptación eficiente tanto de las capacidades lingüísticas como de razonamiento cruzado entre modalidades.

Dataset e ingeniería de prompts.

El conjunto de datos ROCOv2 proporciona imágenes radiológicas con sus correspondientes *captions* y conceptos médicos codificados como CUIs (*Concept Unique Identifiers*). Implementamos una estrategia diversificada de prompts que abarca cinco categorías de plantillas: *captioning* básico, identificación de conceptos, integración multimodal, consultas dirigidas sobre conceptos y descripciones condicionales. Cada muestra de entrenamiento recibe una plantilla asignada aleatoriamente para mejorar la robustez del modelo

y prevenir el sobreajuste. El formato de instrucción sigue un paradigma conversacional estructurado:

> **USER:** <image>\n[Consulta específica de la tarea]
> **ASSISTANT:** [Respuesta generada]

Configuración de entrenamiento.

Hiperparámetros: Learning rate de 5e−5 con *cosine annealing*, *batch size* efectivo de 32 (acumulación de gradientes 4×8), longitud máxima de secuencia de 1024 tokens, entrenamiento de precisión mixta (`bfloat16`/`float16`).
Optimización: Optimizador AdamW con ratio de *warmup* del 3 %, *gradient checkpointing* para eficiencia de memoria, evaluación cada 250 iteraciones.
Procesamiento de Etiquetas: El *collator* personalizado implementa *instruction tuning* mediante enmascaramiento selectivo, donde los tokens de entrada reciben la etiqueta −100 y solo las respuestas del asistente contribuyen al cálculo de la pérdida.

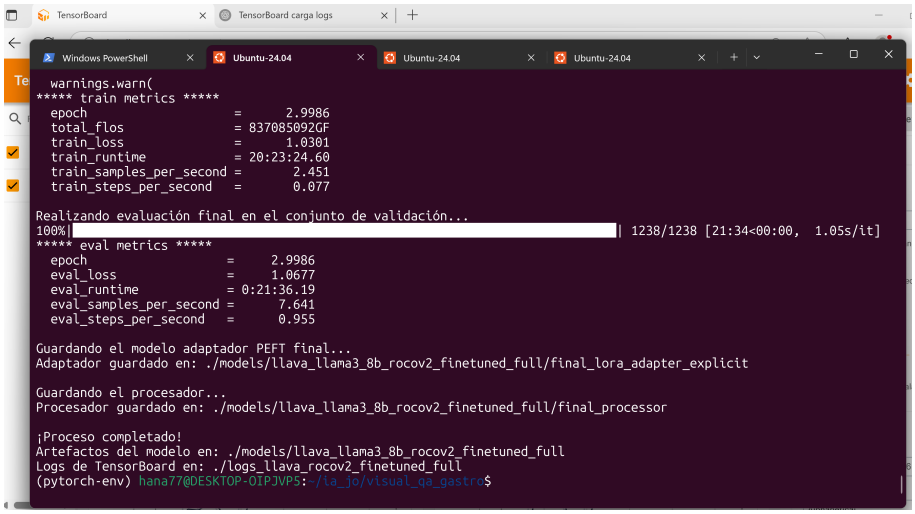


Figura 7: Conclusión del proceso de fine-tuning del modelo de visión y lenguaje LLaVA-LLaMA 3 8B con una train loss de 1.03 y una eval loss de 1.06.

Gestión de memoria y escalabilidad.

Para abordar limitaciones computacionales, implementamos procesamiento por bloques (*chunked*) del dataset, permitiendo entrenamiento a gran escala en hardware limitado. El pipeline incluye limpieza automática de memoria y almacenamiento intermedio en disco para los bloques de datos procesados.

Preprocesamiento de datos.

El pipeline de preprocesamiento integra anotaciones multimodales mediante la fusión sistemática de *captions* y conceptos, mapeo semántico de CUIs a nombres canónicos y validación integral de los datos. Los conceptos médicos se traducen de códigos CUI a términos legibles utilizando el diccionario de mapeo proporcionado.

Proceso de entrenamiento.

El entrenamiento se realiza mediante *fine-tuning* eficiente en parámetros durante una época, con posibilidad de *early stopping*. El *collator* de datos personalizado asegura la alineación correcta de tensores multimodales e implementa la estrategia de enmascaramiento para *instruction tuning*. Los *checkpoints* del modelo se conservan en función de la optimización de la pérdida de validación.

Instruction tuning multimodal.

Nuestro enfoque extiende el *instruction tuning* al dominio médico mediante la diversificación estratégica de prompts y tareas dirigidas de identificación de conceptos. La metodología permite al modelo realizar tareas tanto descriptivas como analíticas sobre imágenes médicas.

Adaptación eficiente en parámetros

La configuración de LoRA se dirige a los componentes críticos del modelo mientras preserva la eficiencia computacional. Este enfoque reduce significativamente los parámetros entrenables, manteniendo la expresividad del modelo para tareas de análisis de imágenes médicas.

7. Descripción técnica del sistema

A continuación, se presenta un diagrama con el flujo de información y una descripción técnica detallada del sistema de Question Answering Biomédico Multimodal con Explicabilidad Incorporada, llamado **MultimodalBioQA**, con fines de reproducibilidad y extensión del trabajo.

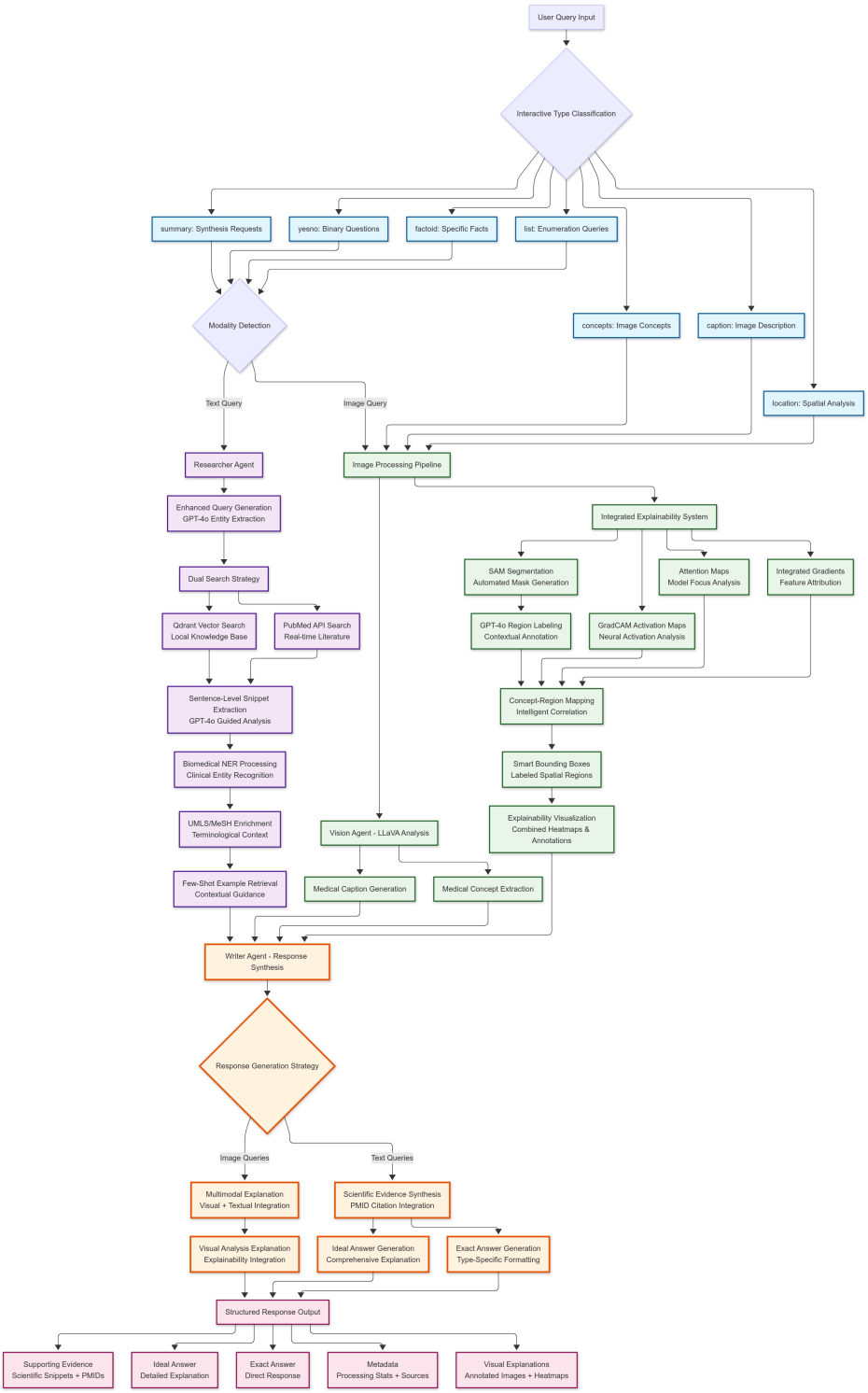


Figura 8: Flujo de información detallado del sistema multimodal de Question answering.

El sistema implementa una arquitectura de agentes distribuidos que integra capacidades de procesamiento de texto e imagen para responder consultas biomédicas complejas. La aplicación principal (`app.py`) constituye el núcleo orquestador que coordina múltiples agentes especializados mediante una interfaz de usuario desarrollada en Streamlit.

7.1. Aplicación Principal (`app.py`)

La aplicación principal implementa un sistema integrado de explicabilidad que combina múltiples técnicas de interpretabilidad para el análisis de imágenes médicas:

- **IntegratedExplainabilitySystem**: Clase principal que integra LLaVA (Large Language and Vision Assistant) con técnicas de explicabilidad visual incluyendo GradCAM, Attention Maps e Integrated Gradients. Esta integración representa una innovación significativa al combinar modelos multimodales con técnicas de interpretabilidad tradicionales.
- **Pipeline de procesamiento dual**: El sistema maneja dos flujos principales:
  1. *Flujo de Texto*: Utiliza `researcher_agent` para búsqueda en PubMed y `writer_agent` para síntesis
  2. *Flujo de Imagen*: Emplea `vision_agent` con sistema de explicabilidad integrado para análisis visual
- **Gestión de sesiones**: Implementa un sistema robusto de gestión de sesiones con identificadores únicos para rastrear análisis de explicabilidad y mantener coherencia en procesamiento *batch*.

7.1.1. Procesamiento de Consultas Multimodales

La función `process_full_query()` implementa la lógica central del sistema:

Listing 1: Función principal de procesamiento de consultas

```
def process_full_query(question_data: dict) -> dict:
```

Esta función orquesta el procesamiento completo mediante:

- **Inferencia de tipo de pregunta**: Utiliza heurísticas avanzadas y confirmación interactiva para clasificar preguntas en categorías (*yesno*, *factoid*, *list*, *summary*, *concepts*, *caption*, *location*)
- **Enrutamiento condicional**: Dirige el procesamiento según la presencia de datos de imagen
- **Sistema de explicabilidad integrado**: Para consultas de imagen, aplica análisis de explicabilidad completo que incluye:
  - Extracción de conceptos médicos mediante LLaVA
  - Generación de mapas de activación (GradCAM)
  - Análisis de atención (Attention Maps)
  - Cálculo de gradientes integrados (Integrated Gradients)
  - Mapeo automático concepto-región con *bounding boxes*

7.1.2. Sistema de configuración (`config.py`)

El módulo de configuración implementa un sistema centralizado de gestión de parámetros que abarca:

- **Configuración de APIs**: Gestión de claves para OpenAI, NCBI/PubMed, UMLS
- **Configuración de base de datos vectorial**: Parámetros para Qdrant incluyendo URLs, colecciones y límites de búsqueda
- **Configuración de modelos**: Especificación de modelos de embedding (BiomedNLP-PubMedBERT), NER biomédico, y configuración de LLaVA
- **Optimizaciones de hardware**: Configuración automática de dispositivos (CUDA/CPU), quantización 4-bit, y gestión de memoria

7.2. Innovaciones arquitectónicas

Sistema de explicabilidad integrado

La principal innovación arquitectónica radica en la integración *continua* de múltiples técnicas de explicabilidad:

- **Explicabilidad multi-nivel**: Combina explicabilidad a nivel de atención (Attention Maps), activación (GradCAM) y gradientes (Integrated Gradients).
- **Mapeo Concepto-Región**: Implementa un método que mapea conceptos médicos extraídos por LLaVA a regiones específicas de la imagen mediante análisis especializado.
- **Bounding boxes automáticas**: Genera automáticamente *bounding boxes* etiquetadas que correlacionan conceptos médicos con ubicaciones espaciales.

Arquitectura de agentes distribuidos

El sistema implementa un patrón de arquitectura de agentes donde cada agente tiene responsabilidades específicas:

- **Separation of concerns**: Cada agente maneja un aspecto específico del procesamiento.
- **Comunicación asíncrona**: Los agentes se comunican mediante estructuras de datos estandarizadas.
- **Tolerancia a fallos**: Implementa *fallbacks* elegantes cuando los componentes individuales fallan.

7.3. Agentes de procesamiento

7.3.1. Agente de Investigación (`researcher_agent.py`)

El `researcher_agent` constituye el núcleo del sistema de recuperación de información biomédica, implementando una arquitectura híbrida que combina búsqueda en bases de datos vectoriales locales con consultas en tiempo real a la API de PubMed.

Arquitectura de búsqueda híbrida

El agente implementa una estrategia de búsqueda multicapa:

Listing 2: Función principal del agente de investigación

```
def run_researcher(question: str, question_data: Dict) -> Dict:
```

- **Generación de consultas mejoradas**: Utiliza GPT-4o para extraer entidades médicas significativas de la pregunta mediante el método `_extract_medical_entities_with_gpt4o()`. Este enfoque supera las limitaciones de los sistemas NER tradicionales al comprender el contexto semántico.
- **Búsqueda primaria en Qdrant**: Ejecuta búsquedas de tipo Qdrant-first donde busca de preferencia en la base de datos vectorial y luego en la API de PubMed con parámetros optimizados (`hasabstract[text]`) para garantizar la disponibilidad de contenido completo.
- **Extracción de snippets a nivel de oración**: Implementa una función que utiliza GPT-4o para extraer oraciones específicas que responden directamente a la pregunta, en lugar de devolver abstracts completos. Esta técnica mejora significativamente la precisión de la evidencia.

Enriquecimiento contextual

El sistema implementa múltiples capas de enriquecimiento:

- **Integración UMLS/MeSH**: La función `_query_umls_for_context()` consulta las terminologías médicas estandarizadas para proporcionar definiciones contextuales de los términos médicos identificados.
- **Recuperación Few-Shot**: Implementa `_retrieve_few_shot_example_from_file()` para obtener ejemplos similares que guían la generación de respuestas, mejorando la consistencia y calidad del output.
- **Ranking y filtrado**: `_filter_and_rank_snippets()` utiliza una combinación de scores de relevancia originales y matching de palabras clave para optimizar la selección de evidencia.

Optimizaciones de rendimiento

- **Cache de consultas UMLS**: Mantiene cache a nivel de módulo para evitar consultas repetitivas.
- **Gestión de rate limits**: Implementa delays configurables para respetar las limitaciones de las APIs externas que bloquean las consultas cuando exceden el límite de uso.

7.3.2. Agente de Escritura (`writer_agent.py`)

El `writer_agent` representa la capa de síntesis que transforma evidencia científica cruda en respuestas coherentes y bien estructuradas.

Sistema de generación diferenciada por tipo

El agente implementa estrategias de generación específicas según el tipo de pregunta:

Listing 3: Función principal del agente de escritura

```
def run_writer(question: str, qtype: str, snippets: List[Dict],
               mesh_info: Optional[Dict], few_shot_example:
                  Optional[Dict]) -> Dict:
```

- **Prompts especializados**: Cada tipo de pregunta (*yesno*, *factoid*, *list*, *summary*) utiliza prompts optimizados con instrucciones específicas para el formato de respuesta esperado.
- **Validación de respuestas**: `_validate_exact_answer()` implementa lógica de validación específica por tipo para asegurar conformidad con formatos esperados.
- **Síntesis dual**: Genera tanto `exact_answer` (respuesta directa) como `ideal_answer` (explicación detallada) para proporcionar información a múltiples niveles de granularidad.

Procesamiento de consultas de imagen

Para consultas multimodales, el agente implementa:

- **Prompts contextualizados para imagen**: usa una función que crea prompts específicos que incorporan resultados del análisis visual, conceptos detectados y metadatos de explicabilidad.
- **Integración de resultados multimodales**: Combina información textual y visual para generar explicaciones integrales.

Mejores prácticas de generación

- **Control de temperatura**: Utiliza diferentes valores de temperatura según el tipo de tarea (0.0 para *exact answers*, 0.2-0.3 para *ideal answers*.)
- **Gestión de tokens**: Implementa límites adaptativos de tokens basados en la complejidad de la consulta.
- **Citación automática**: Incluye referencias automáticas a PMIDs en las respuestas generadas por el Agente de Escritura.

7.3.3. Agente de Visión (`vision_agent.py`)

El `vision_agent` encapsula la funcionalidad de análisis de imagen médica utilizando modelos LLaVA fine-tuned para el dominio biomédico.

Arquitectura del agente

Listing 4: Clase principal del agente de visión

```
class VisionAgent:
    def __init__(self, base_model_path, adapter_repo, device="auto"
        ,
                 load_in_4bit=True, merge_adapters=False):
```

- **Modelo base LLaVA**: Utiliza `xtuner/llava-llama-3-8b-v1_1-transformers` como modelo base con el adaptador LoRA ajustado con el dataset radiográfico del reto visual.
- **Optimizaciones de memoria**: Implementa quantización 4-bit opcional y fusión de adaptadores para optimizar el uso de memoria GPU.
- **Gestión de dispositivos**: Detección automática y configuración de dispositivos (CUDA/CPU) con *fallbacks*.

Funcionalidades especializadas

- **Generación de conceptos**: `generate_concepts()` extrae conceptos médicos estructurados de imágenes utilizando prompts en español para el dominio médico.
- **Generación de captions**: `generate_captions()` produce descripciones textuales detalladas de imágenes médicas siguiendo convenciones radiológicas.
- **Procesamiento en batch**: Soporte para procesamiento eficiente de múltiples imágenes con gestión optimizada de memoria.

Características técnicas avanzadas

- **Manejo de diversos inputs**: Soporte para múltiples formatos de entrada, lo que incluye rutas de fichero, objetos PIL, y data URIs base64.
- **Gestión de patch size**: Implementa corrección automática del `patch_size` para tener compatibilidad con diferentes versiones del procesador.
- **Gestión de memoria**: Implementa cleanup manual de tensores y cache CUDA para prevenir *memory leaks* durante inferencia prolongada.

Prompts especializados

El agente utiliza prompts cuidadosamente diseñados:

Listing 5: Prompts especializados para el agente de visión

```
CONCEPT_PROMPT = "USER: <image>\\nEnumera los conceptos medicos
    clave (CUIs) observados o inferidos en esta imagen.\\nASSISTANT:
     Los conceptos medicos clave son:"

CAPTION_PROMPT = "USER: <image>\\nCual es la descripcion de esta
    imagen medica?\\nASSISTANT:"
```

Estos prompts están personalizados para el dominio médico y diseñados para maximizar la precision y relevancia de las respuestas generadas.

7.3.4. Agente de chatbot interactivo (`chatbot_agent.py`)

El `chatbot_agent` implementa un sistema de validación interactiva de tipos de pregunta que utiliza inferencia heurística combinada con confirmación del usuario para optimizar la precisión en la clasificación de consultas.

Sistema de inferencia de tipos

Listing 6: Función de inferencia de tipos de pregunta

```
def infer_question_type(question: str) -> str:
```

El sistema implementa una cascada de patrones heurísticos organizados jerárquicamente:

- **Patrones de yes/no (Prioridad)**: Utiliza expresiones regulares para identificar preguntas binarias:

  Listing 7: Patrones para preguntas Yes/No

  ```
  yesno_patterns = [
      r'^(is|are|do|does|did|can|could|will|would|should|has|have|
         had)\s',
      r'\beffective\b.*\bfor\b',
      r'\bassociated\b.*\bwith\b'
  ]
  ```

- **Detección de listas**: Identifica preguntas que requieren respuestas enumerativas mediante patrones como `^(what|which)`
  `s+.*`
  `s+(are|include)`.

- **Clasificación summary**: Reconoce solicitudes de síntesis mediante patrones como `^(summarize|describe|explain)`.

- **Factoid por defecto**: Utiliza *factoid* como categoría de *fallback* para preguntas específicas no clasificadas.

Interfaz de confirmación interactiva

La función `interactive_type_confirmation()` implementa un sistema de UI avanzado:

- **Visualización contextual**: Presenta la pregunta original y el tipo inferido en un contenedor estilizado con HTML/CSS personalizado.

- **Confirmación de selección**: Permite al usuario confirmar o modificar el tipo inferido con todas las opciones disponibles (*yesno*, *factoid*, *list*, *summary*).

- **Auto-confirmación temporal**: Implementa un *countdown* de 5 segundos que auto-confirma la selección para optimizar la experiencia de usuario en casos de alta confianza.

- **Gestión de estado**: Utiliza `st.session_state` para mantener persistencia del tipo confirmado a través de la sesión.

7.3.5. Agente de Explicabilidad (`explainability_agent.py`)

El `explainability_agent` constituye un *wrapper* de alto nivel que orquesta el pipeline completo de análisis de explicabilidad para imágenes médicas.

Arquitectura de Gestión de Archivos

El agente implementa un sistema robusto de gestión de archivos:

- **Directorio de uploads seguro**: Implementa `save_uploaded_image()` que maneja múltiples tipos de objetos de archivo (Flask, FastAPI, BytesIO) con validación de extensiones y nombres únicos.
- **Estructura organizada de directorios**: Crea automáticamente subdirectorios para imágenes, heatmaps y metadatos, facilitando la organización y recuperación de resultados.
- **Gestión de sesiones**: Genera identificadores únicos de sesión combinando *timestamps* y hashes para garantizar trazabilidad sin colisiones.

Pipeline de análisis integrado

La función central `analyze_image()` orquesta:

- **Invocación de herramientas especializadas**: Llama a `analyze_medical_image_tool()` de `explainability_tool.py` para el análisis técnico.
- **Enriquecimiento de metadatos**: Agrega información temporal, rutas de archivos y URLs *web-ready* para integración con interfaces de usuario.
- **Gestión elegante de errores**: Implementa manejo completo de errores con logging detallado y *fallbacks* apropiados.

Sistema de recuperación y limpieza

- **Recuperación de resultados**: `get_analysis_results()` permite recuperar análisis previos mediante `session_id`, facilitando la persistencia de resultados a través de sesiones.
- **Limpieza automática**: `cleanup_old_files()` implementa *garbage collection* temporal para mantener el sistema limpio, con configuración de retención personalizable.
- **Endpoint factory**: `create_explainability_endpoint()` genera funciones endpoint *ready-to-use* para integración con frameworks web.

7.4. Herramientas de análisis especializado

7.4.1. Herramienta de NER biomédica (`ner_tool.py`)

El `ner_tool` implementa un sistema avanzado de extracción de entidades biomédicas con integración UMLS/MeSH para enriquecimiento semántico.

Gestión de tickets UMLS

Listing 8: Función de gestión de tickets UMLS

```
def _get_umls_ticket() -> Optional[str]:
```

Implementa un sistema robusto de autenticación con UMLS:

- **Caché de tickets**: Mantiene tickets de sesión con gestión automática de expiración y margen de seguridad de 5 minutos.
- **Renovación automática**: Detecta expiraciones de tickets y renueva automáticamente sin interrumpir el flujo de trabajo.
- **Retry lgic con backoff**: `_query_umls_with_retry()` implementa reintentos automáticos con manejo de errores 401 (*unauthorized*).

Pipeline de NER Optimizado

El sistema de NER implementa múltiples optimizaciones:

- **Caché de pipelines**: `_get_ner_pipeline()` mantiene cache a nivel de módulo para evitar recargas costosas de modelos.
- **Filtrado avanzado**: `_filter_ner_entities()` elimina:
  - Tokens subword de BERT.
  - Entidades con scores de confianza <0.5.
  - Palabras comunes (*stopwords*).
  - Tokens sin caracteres alfabéticos.
- **Agregación de estrategias**: Utiliza `aggregation_strategy="simple"` para agrupar automáticamente *word pieces* en entidades coherentes.

Enriquecimiento avanzado con UMLS

La función principal implementa:

- **Clasificación de entidades clínicas**: Categoriza automáticamente entidades en tipos clínicos vs. no-clínicos basándose en labels especializados:

Listing 9: Labels clínicos para clasificación

```
clinical_labels = {
    'DISEASE', 'DRUG', 'CHEMICAL', 'GENE', 'PROTEIN',
    'TREATMENT', 'PROCEDURE', 'SYMPTOM', 'ANATOMY'
}
```

- **Mapeo UMLS selectivo**: Aplica enriquecimiento UMLS solo a entidades clínicas identificadas, optimizando eficiencia y relevancia.
- **Estructura de resultados**: Retorna un diccionario estructurado con entidades procesadas, clasificaciones clínicas, mapeos UMLS y manejo de errores.

7.4.2. Herramientas especializadas adicionales

Herramienta de Explicabilidad (`explainability_tool.py`)

Implementa el análisis técnico de explicabilidad combinando múltiples tecnologías:

- **Integración SAM + GPT-4o**: Combina *Segment Anything Model* para segmentación con GPT-4o para etiquetado de regiones.
- **Pipeline de análisis completo**:

Listing 10: Función de análisis de imagen médica

```
def analyze_medical_image(self, image_path: str, session_id: str
    = None) -> dict:
```

  - Carga y valida imagen
  - Obtiene contexto con VisionAgent (*caption* + conceptos)
  - Genera máscaras con SAM
  - Etiqueta regiones con GPT-4o
  - Crea visualizaciones (imagen anotada + *heatmap*)
  - Guarda resultados estructurados
- **Etiquetado contextual**: `_label_region_with_gpt4o()` utiliza los conceptos extraídos por LLaVA como contexto para mejorar la precisión del etiquetado de regiones.

Detector de elementos gráficos (`arrow_detector_tool.py`)

Implementa técnicas de visión por computadora como análisis geométrico para detectar elementos direccionales:

- **Detección de contornos**: Utiliza OpenCV para identificar formas de flecha mediante análisis de contornos y aproximación poligonal.
- **Análisis direccional**: `advanced_arrow_detection()` implementa:
  - Detección de líneas con *Hough Transform*
  - Extrapolación direccional para predecir targets
  - Generación de *bounding boxes* en ubicaciones predichas

Procesador de ficheros (`file_preprocessing_tool.py`)

Proporciona capacidades de ingesta de documentos:

- **Soporte Multi-formato**: Maneja PDF, DOCX, PPTX, XLSX, TXT.
- **Procesamiento de URLs**: `process_url()` implementa:
  - Detección automática de tipo de contenido (HTML vs PDF)
  - Extracción de texto de PDFs con PyMuPDF
  - Parsing automático de HTML con BeautifulSoup
  - Limitación de contenido (15,000 caracteres) para prevenir *overload*

7.4.3. Sistema de base de datos vectorial y gestión de datos

Se implementó una base de datos vectorial local utilizando Qdrant para mejorar la capacidad de búsqueda y recuperación de snippets. Luego ya se le encontraron otros usos como crear una colección para almacenar los ejemplos de preguntas y respuestas para few shot.

Sistema de ingesta de artículos biomédicos (`ingest_pubmed.py`)

El sistema de ingesta constituye el componente de mayor complejidad técnica junto con el fine tuning del modelo de visión y lenguaje, implementando un pipeline optimizado para procesamiento masivo de más de 30 millones de artículos científicos de PubMed.

Arquitectura de procesamiento Streaming

Listing 11: Función de procesamiento de artículos PubMed

```
def process_pubmed_article_node(article_node, source_file_name):
```

El sistema implementa procesamiento *streaming* mediante `lxml.etree.iterparse` para manejar archivos XML.gz de gran tamaño sin cargar el documento completo en memoria:

- **Parsing incremental**: Utiliza `iterparse` con eventos `('end',)` para procesar nodos `PubmedArticle` individualmente, liberando memoria inmediatamente después del procesamiento.
- **Gestión de memoria**: Implementa `elem.clear()` y eliminación de nodos previos para prevenir *memory leaks* durante procesamiento prolongado:

Listing 12: Gestión de memoria durante parsing

```
elem.clear()
while elem.getprevious() is not None:
    del elem.getparent()[0]
```

- **Batching multinivel**: Utiliza dos niveles de *batching*:
  - `EMBEDDING_BATCH_SIZE` (512): Para generación de embeddings en GPU.
  - `UPSERT_BATCH_SIZE` (512): Para operaciones de inserción en Qdrant.

Extracción de metadatos

El sistema implementa una extracción completa de metadatos con manejo robusto de formatos inconsistentes:

- **Parsing de fechas avanzado**: `parse_date_from_node()` maneja múltiples formatos:

Listing 13: Función de parsing de fechas

```
def parse_date_from_node(date_node):
    # Prioridad: Year/Month/Day -> MedlineDate -> fallbacks
    year = safe_find_text(date_node, ".//Year")
```

```
    month_str = safe_find_text(date_node, ".//Month")
    # Mapeo inteligente de meses (numericos y textuales)
    month = MONTH_MAP.get(month_str.lower().strip(), month_str.
       strip())
```

- **Extracción de abstracts estructurados**: Maneja abstracts con etiquetas (`BACKGROUND`, `METHODS`, `RESULTS`, `CONCLUSIONS`) mediante `itertext()` para capturar contenido anidado.
- **Gestión de autores**: Procesa listas de autores con *fallbacks* para nombres incompletos y manejo de iniciales.
- **Metadatos enriquecidos**: Extrae términos MeSH, palabras clave, DOIs, títulos de journals y URLs de documentos.

Optimizaciones de rendimiento para producción

- **Configuración de Optimizador Qdrant**: Ajusta parámetros para ingesta masiva:

Listing 14: Configuración del optimizador Qdrant

```
qdrant_client.update_collection(
    collection_name=COLLECTION_NAME,
    optimizers_config=models.OptimizersConfigDiff(
        indexing_threshold=100000,  # Optimizar menos
           frecuentemente
        flush_interval_sec=60       # Flushear menos
           frecuentemente
    )
)
```

- **Truncamiento de tokens**: `truncate_text_to_tokens()` utiliza el tokenizer de BiomedNLP-PubMedBERT para truncar a 510 tokens manteniendo coherencia semántica.
- **Retry Logic con Tenacity**: Implementa reintentos exponenciales para operaciones de red:

Listing 15: Implementación de retry logic

```
@retry(stop=stop_after_attempt(5), wait=wait_exponential(
   multiplier=1, min=4, max=30))
def upsert_points_to_qdrant(points_to_upsert):
```

- **Procesamiento Incremental**: Mantiene log de archivos procesados (`processed_files.log`) para permitir reinicios sin reprocesamiento.

Gestión de dispositivos y embeddings

- **Detección automática de hardware**: Configura automáticamente GPU/CPU basándose en disponibilidad de CUDA.
- **Modelo especializado**: Utiliza `microsoft/BiomedNLP-PubMedBERT` (768 dimensiones) optimizado para texto biomédico.

- **Batching eficiente**: Procesa embeddings en lotes para maximizar utilización de GPU mientras controla uso de VRAM.

7.5. Sistema de ingesta Few-Shot (`ingest_few_shots.py`)

El sistema maneja la creación de una colección especializada para ejemplos de entrenamiento que guían la generación de respuestas.

Carga y validación de datos

Listing 16: Función de carga de datos de entrenamiento

```
def load_training_data(filepath: str) -> list:
```

Implementa carga robusta con múltiples validaciones:

- **Detección de Estructura**: Maneja tanto listas directas como objetos con clave `"questions"`.
- **Validación de Integridad**: Verifica la presencia de campos requeridos (`"body"`) antes del procesamiento.
- **Manejo de Errores JSON**: Implementa logging detallado para errores de decodificación y estructura.

Procesamiento de ejemplos

El sistema procesa ejemplos manteniendo la estructura completa para una recuperación posterior:

- **Input para embedding**: Utiliza el texto de la pregunta (`"body"`) para generar embeddings de búsqueda.
- **Output completo**: Almacena el registro completo como JSON string para preservar toda la información contextual.
- **Consistencia de modelos**: Utiliza el mismo modelo de embedding (BiomedNLP-PubMedBERT) que el sistema principal para garantizar coherencia semántica.

7.5.1. Preparación de Few-Shot

Este componente implementa un sistema sofisticado de optimización de ejemplos *few-shot* específicamente diseñado para GPT-4o.

Pipeline de optimización GPT-4o

Listing 17: Función de creación de dataset optimizado para GPT-4o

```
def create_gpt4o_optimized_dataset(questions_data: List[Dict],
    golden_answers_data: List[Dict]) -> List[Dict]:
```

Implementa optimizaciones específicas para modelos de última generación como GPT-4o:

- **Detección de Idioma**: `determine_language()` utiliza análisis heurístico basado en palabras clave para clasificación español/inglés.
- **Normalización de Tipos**: `extract_question_type()` implementa mapeo consistente de tipos de pregunta con *fallbacks* inteligentes.
- **Optimización de Longitud**: Truncamiento automático con límites específicos (500 chars input, 2000 chars output) optimizados para la ventana de contexto de GPT-4o.

Limpieza y estructuración de datos

Listing 18: Función de limpieza y validación de texto

```
def clean_and_validate_text(text: str, max_length: int = None) ->
    str:
```

- **Normalización de espacios**: Elimina espacios extra y caracteres problemáticos mediante regex.
- **Truncamiento preservando palabras**: Corta en límites de palabras para mantener coherencia semántica.
- **Validación de entrada**: Maneja casos edge como texto vacío, None, o tipos incorrectos.

Optimización JSON

El desafío bioASQ solicita entregables en formato JSON por lo que se tuvo que trabajar en el formato de salida (Nentidis et al., 2025).

Listing 19: Función de optimización de salida JSON

```
def optimize_json_output(data: Dict[str, Any]) -> str:
```

Implementa optimización específica para reducir *usage* de tokens:

- **Serialización compacta**: Utiliza separadores mínimos y `ensure_ascii=False`.
- **Truncamiento automático**: Reduce campos específicos (como `ideal_answer`) si el output excede límites.
- **Preservación de estructura**: Mantiene la estructura requerida mientras optimiza el contenido.

Validación de Calidad

Listing 20: Función de validación de calidad de ejemplos

```
def validate_examples_quality(examples: List[Dict]) -> Dict[str,
    Any]:
```

Sistema completo de *quality assurance*:

- **Validación de estructura**: Verifica presencia de campos requeridos (`input`, `output`, `metadata`).
- **Validación JSON**: Confirma que el campo `output` contiene JSON válido.
- **Métricas de longitud**: Valida que inputs y outputs tienen longitudes mínimas apropiadas.
- **Reporting detallado**: Genera reportes con estadísticas e identificación específica de problemas.

7.6. Utilidades de base de datos vectorial (`vector_store_utils.py`)

Proporciona la interfaz de alto nivel para búsqueda semántica en tiempo de ejecución.

7.6.1. Integración con LlamaIndex

Listing 21: Función de recuperación de ejemplos similares

```
def retrieve_similar_few_shot_examples(query_text: str, language:
    str, top_k: int = 3) -> List[Dict]:
```

- **Embedding consistente**: Utiliza `Settings.embed_model` de LlamaIndex para garantizar coherencia con el sistema de embeddings principal.
- **Búsqueda semántica**: Ejecuta búsqueda vectorial en la colección *few-shot* utilizando *cosine similarity*.
- **Manejo elegante de errores**: Implementa *fallbacks* apropiados cuando componentes no están disponibles.

Gestión de dependencias

- **Imports condicionales**: Maneja de forma elegante la ausencia de bibliotecas opcionales.
- **Logging informativo**: Proporciona mensajes claros sobre disponibilidad de componentes.
- **Configuración de timeout**: Implementa timeouts apropiados para operaciones de red.

7.6.2. Sistema de conversión de datos (`convert_jsonl_to_json.py`)

Herramienta especializada para conversión eficiente de formatos de datos con minimización de uso de memoria.

Procesamiento streaming

Listing 22: Función de conversión JSONL a JSON con wrapper

```
def convert_jsonl_to_json_with_wrapper(input_jsonl_path: str,
    output_json_path: str, chunk_size: int = 100) -> bool:
```

- **Archivo Temporal**: Utiliza `tempfile.NamedTemporaryFile` para construir el JSON final de manera incremental.
- **Procesamiento línea por línea**: Evita cargar el archivo completo en memoria procesando cada línea individualmente.
- **Wrapping Automático**: Agrega automáticamente la estructura `{"questions": [...]}` requerida por el formato BioASQ.

Optimizaciones de rendimiento

- **Chunked processing**: Procesa archivos en chunks configurables para balancear memoria y rendimiento.
- **Progress reporting**: Implementa logging periódico para monitorear progreso en archivos grandes.
- **Error recovery**: Continúa procesamiento incluso cuando líneas individuales fallan, reportando errores sin interrumpir el proceso completo.

7.7. Integración final y análisis arquitectónico global

El sistema implementa una arquitectura de microservicios distribuidos que integra múltiples paradigmas de IA para crear una plataforma integral de *question answering* biomédico multimodal.

La arquitectura tiene un patrón arquitectónico híbrido y combina elementos de:

- **Event-Driven architecture**: Los agentes se comunican mediante estructuras de datos estandarizadas que actúan como eventos entre componentes.
- **Pipeline architecture**: El procesamiento sigue flujos estructurados con etapas bien definidas.
- **Microservices pattern**: Cada agente encapsula funcionalidad específica con interfaces claramente definidas.
- **Repository pattern**: La base de datos vectorial actúa como repositorio centralizado de conocimiento científico.

7.7.1. Innovaciones técnicas

A continuación, menciono los elementos arquitectónicos más innovadores del sistema o que ayudan a diferenciarlo de otros sistemas:

Sistema de explicabilidad integrado

A mi parecer, la principal contribución técnica del sistema radica en la integración de múltiples técnicas de explicabilidad visual:

- **Explicabilidad multimodal**: Combina explicabilidad basada en atención (LLaVA), activación (GradCAM), y gradientes (*Integrated Gradients*) en un pipeline unificado.
- **Mapeo Concepto-Región**: método que correlaciona conceptos médicos extraídos por LLaVA con regiones espaciales específicas mediante análisis de mapas de activación:
- **Bounding Boxes contextualizadas**: Genera automáticamente *bounding boxes* etiquetadas que no solo identifican regiones, sino que las contextualiza dentro del dominio médico específico.

Extracción de snippets a nivel de oración

Innovación en la forma de recuperación de información biomédica:

- **Mejor precisión**: Mejor que métodos tradicionales de recuperación que devuelven abstracts completos, extrayendo únicamente oraciones que responden directamente a la pregunta.
- **Contextualización GPT-4o**: Utiliza capacidades de comprensión avanzada para identificar relevancia semántica más allá del matching de keywords.
- **Preservación de Fuentes**: Mantiene trazabilidad completa con PMIDs y scores de relevancia para verificación científica.

Pipeline de procesamiento híbrido

- **Búsqueda dual**: Combina base de datos vectorial local (Qdrant) con consultas en tiempo real a PubMed API para maximizar cobertura y actualidad.
- **Few-Shot Learning contextual**: Implementa recuperación semántica de ejemplos *few-shot* que guían la generación de respuestas, lo que mejora la coherencia y la calidad del sistema.
- **Enriquecimiento UMLS**: Aplica terminología médica estandarizada selectivamente solo a entidades clínicas identificadas, de modo de optimizar la eficiencia.

7.7.2. Flujos de procesamiento especializados

Se puede consultar el diagrama de la arquitectura para contextualizar este apartado.

Flujo de consultas de texto

1. **Ingesta y clasificación**: `chatbot_agent` clasifica tipo de pregunta con confirmación interactiva del usuario.
2. **Investigación dual**: `researcher_agent` ejecuta búsqueda híbrida (Qdrant + PubMed API).

3. **Extracción automática**: GPT-4o extrae snippets relevantes a nivel de oración.
4. **Enriquecimiento contextual**: NER biomédica + mapeo UMLS/MeSH.
5. **Síntesis especializada**: `writer_agent` genera respuestas diferenciadas por tipo.

Flujo de consultas de imagen

1. **Análisis visual primario**: `vision_agent` (LLaVA) extrae conceptos y genera captions.
2. **Explicabilidad multitécnica**:
   - GradCAM para mapas de activación.
   - *Attention Maps* para análisis de atención.
   - *Integrated Gradients* para atribución.
3. **Segmentación**: SAM genera máscaras + GPT-4o etiqueta regiones.
4. **Mapeo concepto-región**: Método propio correlaciona conceptos con ubicaciones.
5. **Visualización integrada**: Genera imagen anotada + *heatmaps* + *bounding boxes*.

7.7.3. Optimizaciones de rendimiento y escalabilidad

Para concluir la descripción de la arquitectura del sistema **MultimodalBioQA** se mencionan las optimizaciones utilizadas en cuanto a memoria y tolerancia a fallos.

Gestión de memoria y recursos

- **Quantización 4-bit**: Reduce uso de VRAM en modelos de visión manteniendo precisión.
- **Batching adaptativo**: Optimiza *throughput* de embeddings y operaciones de base de datos.
- **Cache multinivel**: Pipelines NER, tickets UMLS, y embeddings con TTL configurables.
- **Limpieza automática**: Limpieza automática de archivos temporales y caché.

Tolerancia a fallos

- **Fallbacks elegantes**: Cada componente implementa degradación elegante cuando alguna dependencia falla.
- **Procesamiento incremental**: Capacidad de reanudar procesos interrumpidos sin pérdida de progreso. Esto es especialmente útil en los componentes que toman muchas horas o días para ejecutarse como el fine-tuning del modelo de visión y lenguaje o la ingesta de artículos médicos en la base de datos vectorial local.

Cronograma general

La duración total estimada del proyecto es de aproximadamente 16 semanas, comenzando el 10 de marzo de 2025 y finalizando el 25 de junio de 2025.

8. Cronograma del Proyecto

Fase	Tareas	Semanas
Fase 1: Configuración y revisión de literatura	SetUp Inicial	1-4
Fase 2: Corpus y métricas	Corpus + Métricas	5-6
Fase 3: Implementación de sistemas baselines	Prototipos (para retos)	6-9
Fase 4: Evaluación de resultados	Evaluar resultados de desafíos	9-10
Fase 4: Desarrollo de sistema integrado	Adaptar sistema integrado	10-12
Fase 5: Experimentación	Testeo Modelos Finales	12-15
Fase 6: Análisis y redacción	Redacción Informe Final	10-16

Cuadro 1: Cronograma de actividades del proyecto (marzo-junio 2025).

Fuente. Elaboración propia a partir de la planificación del proyecto.

9. Resultados esperados

9.1. Resultado esperado 1

Lograr precisión de recuperación de información biomédica mediante arquitectura RAG híbrida.

Se espera demostrar que se logra precisión de recuperación de información científica para responder preguntas exactas (exact_answers) mediante la implementación de un sistema RAG híbrido que combine base de datos vectorial local (Qdrant) y consultas en tiempo real a PubMed API.

Métricas de validación:

- Performance competitivo en BioASQ Task 13b: posicionamiento entre los 10 primeros lugares a nivel internacional.
- Latencia optimizada: < 30 segundos para consultas textuales complejas.
- Cobertura de literatura: acceso a > 30 millones de artículos PubMed con búsqueda semántica.

9.2. Resultado esperado 2

Demostración de capacidad de procesamiento multimodal con explicabilidad integrada para análisis de imágenes médicas.

Se espera validar la capacidad del sistema para procesar consultas multimodales (texto + imagen) con explicabilidad integral, integrando análisis visual avanzado (LLaVA-LLaMA 3) con técnicas de explicabilidad multitécnica (GradCAM, Attention Maps, Integrated Gradients) y mapeo concepto-región.

Métricas de validación:

- Performance en ImageCLEFmed Caption: posicionamiento destacado en el top 10 en tareas de análisis visual médico.
- Integración técnica: funcionamiento estable del pipeline completo de explicabilidad.
- Mapeo concepto-región: generación automática de bounding boxes contextualizadas.

9.3. Resultado esperado 3

Implementación satisfactoria de sistema unificado multiagente para procesamiento biomédico textual y visual.

Se espera demostrar la viabilidad de una arquitectura multiagente unificada que maneje sin problemas consultas biomédicas textuales y visuales dentro de un framework coherente, con agentes especializados convergiendo en un Writer Agent central para síntesis de respuestas consistentes.

Métricas de validación:

- Integración arquitectónica: funcionamiento estable de pipelines duales convergentes.
- Escalabilidad operacional: manejo eficiente de cargas de trabajo mixtas.

10. Aportes

A continuación, se presentan los principales aportes del sistema desarrollado:

10.1. Aportes metodológicos

10.1.1. Integración de técnicas de explicabilidad multimodal

Esta investigación contribuye metodológicamente mediante la integración sistemática de múltiples técnicas de explicabilidad visual en un framework unificado para análisis de imágenes médicas. La combinación de GradCAM, Attention Maps e Integrated Gradients con mapeo concepto-región representa un avance en la aplicación de métodos de interpretabilidad complementarios en el dominio biomédico. Esta aproximación permite obtener perspectivas explicativas multidimensionales que abordan diferentes aspectos del proceso de toma de decisiones del modelo.

10.1.2. Arquitectura multiagente para procesamiento biomédico híbrido

El diseño arquitectónico propuesto constituye un aporte en el desarrollo de sistemas distribuidos para aplicaciones biomédicas complejas. La implementación de agentes especializados con convergencia en un componente central de síntesis permite mantener especialización técnica mientras se garantiza coherencia en los resultados. Esta aproximación demuestra la viabilidad de manejar múltiples modalidades de entrada (texto e imagen) dentro de un framework arquitectónico unificado.

10.1.3. Extracción de evidencia científica guiada por LLM

La investigación aporta una metodología para extracción de snippets a nivel de oración utilizando capacidades de comprensión avanzada de large language models. Este enfoque supera las limitaciones de métodos tradicionales basados en similarity scoring, permitiendo identificar evidencia científica específica que responde directamente a consultas formuladas, mejorando la precisión y relevancia de la información recuperada.

10.2. Aportes técnicos

10.2.1. Sistema RAG híbrido para área biomédica

Se desarrolló una infraestructura de recuperación augmentada que combina búsqueda vectorial local con consultas en tiempo real a bases de datos especializadas. La implementación procesó exitosamente más de 30 millones de artículos PubMed utilizando embeddings especializados (BiomedNLP-PubMedBERT), estableciendo una base técnica para búsqueda semántica a gran escala en literatura científica biomédica.

10.2.2. Mapeo Concepto-Región en imágenes médicas

La investigación desarrolló un método que correlaciona automáticamente conceptos médicos extraídos por modelos multimodales con regiones espaciales específicas en imágenes médicas. Esta funcionalidad permite generar bounding boxes contextualizadas con etiquetado semánticamente coherente, contribuyendo al desarrollo de herramientas de análisis visual más interpretables para aplicaciones clínicas.

10.2.3. Pipeline de explicabilidad integrado para modelos multimodales

Se implementó un sistema que procesa simultáneamente múltiples tipos de mapas de explicabilidad (activación, atención, gradientes) y los integra en visualizaciones comprehensivas. Esta

implementación técnica demuestra la factibilidad de combinar diferentes enfoques de interpretabilidad en un pipeline operacional para análisis de imágenes médicas en tiempo real.

10.3. Aportes empíricos

10.3.1. Validación en benchmarks internacionales

Los resultados obtenidos en competencias internacionales especializadas proporcionan evidencia empírica sobre la efectividad de los métodos propuestos. El posicionamiento competitivo en BioASQ Task 13b (2° y 3° lugar) valida la efectividad del módulo textual, mientras que los resultados en ImageCLEFmed Caption 2025, particularmente el 2° lugar en explicabilidad, confirman la contribución en interpretabilidad visual.

10.3.2. Demostración de coexistencia entre performance y explicabilidad

La investigación proporciona evidencia empírica de que la integración de explicabilidad comprehensiva no compromete el performance competitivo en tareas especializadas. Los resultados demuestran que sistemas con múltiples técnicas de interpretabilidad pueden mantener efectividad comparable a sistemas optimizados únicamente para performance.

10.3.3. Escalabilidad de procesamiento multimodal

Los experimentos realizados confirman la viabilidad técnica de procesar volúmenes industriales de literatura científica (30M+ artículos) manteniendo tiempos de respuesta operacionales (< 30 segundos para consultas textuales y < 60 segundos para consultas visuales). Esta validación empírica es relevante para la implementación de sistemas similares en entornos de producción.

10.4. Aportes al conocimiento disciplinar

10.4.1. Caracterización de arquitecturas agénticas en biomedicina

Esta investigación contribuye al entendimiento de cómo arquitecturas multiagente pueden ser aplicadas efectivamente en dominios especializados que requieren procesamiento de múltiples modalidades de información. Los resultados proporcionan *insights* sobre patrones de coordinación entre agentes especializados y mecanismos de convergencia para síntesis de información compleja.

10.4.2. Evaluación de few-shot learning contextual en área biomédica

La implementación y evaluación de recuperación few-shot contextual mediante búsqueda semántica aporta conocimiento sobre la efectividad de este enfoque en dominios técnicos especializados. Los resultados sugieren que la recuperación de ejemplos contextuales mejora la consistencia de respuestas en tipos de pregunta específicos del dominio biomédico.

10.4.3. Análisis de integración modal en sistemas de question answering

Los resultados proporcionan evidencia sobre la viabilidad de integrar procesamiento textual y de imagen en sistemas de Q&A biomédico manteniendo coherencia arquitectónica. Esta contribución es relevante para el desarrollo de sistemas similares que requieran manejo simultáneo de múltiples modalidades de información científica.

11. Resultados

Los resultados del sistema de *question answering* biomédico multimodal con explicabilidad incorporada se presentan en tres secciones principales: resultados de módulo de *question answering* textual, resultados de módulo de *question answering* visual y resultados de rendimiento operacional.

11.1. Resultados del módulo de *question answering* textual

El módulo de *question answering* textual ha sido evaluado utilizando un conjunto de datos de preguntas biomédicas conforme a los requisitos del desafío Task 13b de bioASQ (Nentidis et al., 2025). Cada set de prueba consistía en alrededor de 80 preguntas de distinto tipo para responder.

Para ese desafío se desarrollaron dos aplicaciones, una para concursar en la Fase A y A+ y otra para la Fase B. A continuación, se presentan los resultados obtenidos, los que fueron obtenidos y validados por bioASQ (Nentidis et al., 2025).

Las aplicaciones fueron evaluadas en los conjuntos de pruebas oficiales BioASQ 13B. Se informa el rendimiento en todos los lotes de las fases en las que participaron los sistemas: Lotes 3 y 4 de las Fases A, A+ y B. Los resultados incluyen métricas para respuestas exactas (con desglose por tipo de pregunta) y respuestas ideales (puntuaciones ROUGE). Las tablas 1, 2 y 3 resumen las métricas de evaluación de los sistemas, tal y como las proporciona la herramienta de evaluación de los organizadores de BioASQ (Nentidis et al., 2025).

11.1.1. Resultados de la Fase A

Cuadro 2: Fase A: Snippetss

Lote	Mean Prec.	Recall	F Measure	MAP	GMAP
Lote 3	0.0254	0.1097	0.0392	0.0000	0.0000
Lote 4	0.0177	0.1318	0.0292	0.0000	0.0000

Rendimiento de la Fase A

En la Fase A (Tabla 1), los resultados no fueron óptimos debido a dos factores principales: una base de datos vectorial incompleta y las limitaciones de la API de PubMed. La base de datos de PubMed contiene más de 30 millones de registros y, lamentablemente, no se pudo completar el proceso de indexación a tiempo para este desafío. Además, evaluaciones recientes e informes de competiciones destacan varios retos a la hora de utilizar la API de PubMed para la extracción de fragmentos y las métricas asociadas, especialmente en tareas de respuesta a preguntas biomédicas y recuperación de información. Los principales problemas identificados son:

- **Baja recuperación y resultados incompletos**: La API de PubMed a menudo devuelve documentos insuficientes o inexistentes para las consultas tipo pregunta, lo que provoca una baja recuperación en las tareas de recuperación. Esta limitación afecta significativamente a la eficacia de la extracción de fragmentos, ya que el conjunto de resúmenes candidatos es inadecuado.
- **Rendimiento de la extracción de fragmentos**: Cuando los fragmentos se extraen utilizando métodos basados en modelos lingüísticos amplios (por ejemplo, el método de la cadena de pensamiento GPT-3.5), la recuperación y la precisión media de los fragmentos suelen estar por debajo de la media en comparación con otros sistemas.

Esto representa un área de mejora, ya que la limitación de la API podría solucionarse manteniendo una base de datos vectorial completa de PubMed. Producto de esta experiencia se desarrolló un sistema de ingesta de literatura biomédica que permite la indexación completa de los artículos de PubMed, lo que se detalla en la sección de Metodología.

11.1.2. Resultados de la Fase A+

En la Fase A+, evaluamos el rendimiento del sistema en la respuesta a preguntas biomédicas a través de múltiples tipos de respuesta. Esta fase evalúa tanto la capacidad del sistema para proporcionar respuestas cortas precisas (sí/no, factoides y preguntas de lista) como su capacidad para generar respuestas ideales basadas en pruebas. En las tablas siguientes se presentan los resultados del lote 3, con métricas estándar como la precisión, la puntuación F1 macromediada, el rango recíproco medio (MRR), la precisión, la recuperación y la medida F de las respuestas exactas, así como las puntuaciones ROUGE de las respuestas ideales. Estos resultados ofrecen una visión completa de la eficacia del sistema en el manejo de diversas tareas biomédicas de aseguramiento de la calidad.

Cuadro 3: Fase A+: Lote 3: Respuesta exacta

	Yes/No				Factoid			List		
System	Acc.	F1 Yes	F1 No	Macro F1	S. Acc.	L. Acc.	MRR	M. Prec.	Recall	F-M
AQAMS	0.8182	0.8750	0.6667	0.7708	0.1500	0.2000	0.1750	0.3394	0.3586	0.3478

Cuadro 4: Fase A+: Lote 3: Respuesta ideal (Puntuación Rouge)

R-2 (Rec)	R-2 (F1)	R-SU4	R-SU4 (F1)
0.2135	0.0690	0.2516	0.0778

Para la Fase A+ (Tabla 2), el sistema de Q&A textual alcanzó una precisión sí/no del 92,31 %, con una puntuación F1 del 94,74 % para las respuestas "sí" y del 85,71 % para las respuestas "no" (media macro F1 = 90,23 %). Esto indica que nuestro sistema manejó las preguntas sí/no con eficacia, probablemente debido a que la plantilla de preguntas guiaba al modelo GPT para proporcionar respuestas directas sí/no apoyadas en pruebas.

Cuadro 5: Fase A+: Lote 4: Respuesta exacta

	Yes/No				Factoid			List		
System	Acc.	F1 Yes	F1 No	Macro F1	S. Acc.	L. Acc.	MRR	M. Prec.	Recall	F-M
AQAMS	0.9231	0.9474	0.8571	0.9023	0.4091	0.4091	0.4091	0.2807	0.2897	0.2778

Las puntuaciones factoides para el sistema de Q&A textual en el Lote 4 fueron de aproximadamente 0,41 para la precisión estricta, la precisión indulgente y el rango recíproco medio (MRR). Esto indica que aproximadamente el 41 % de las preguntas factoides tenían la cadena

Cuadro 6: Fase A+: Lote 4: Respuesta ideal (Puntuación Rouge)

R-2 (Rec)	R-2 (F1)	R-SU4 (Rec)	R-SU4 (F1)
0.1840	0.0616	0.2273	0.0763

de respuesta correcta presente (aunque fuera parcialmente) en las respuestas exactas. El rendimiento de las preguntas factoides, aunque moderado, refleja la dificultad de localizar entidades exactas en un entorno de recuperación abierto: los errores se debían a que el sistema proporcionaba respuestas correctas que no coincidían exactamente con el formato de referencia o a que faltaba una de las múltiples respuestas correctas.

Las preguntas con listas resultaron más difíciles para el sistema de Q&A textual en la fase A+: se obtuvo una puntuación F1 de 0,2778, con una precisión de ∼0,28 y una recuperación de ∼0,29. La baja precisión de las listas sugiere que mi sistema a veces no era capaz de responder con precisión. La baja precisión en preguntas de tipo lista sugiere que el sistema a veces pronosticó en exceso elementos que no estaban en la lista gold (incluidos elementos plausibles, pero incorrectos), mientras que la magnitud similar de la recuperación indica que tampoco acertó muchos elementos gold. Esto no es sorprendente, ya que las preguntas de tipo lista suelen requerir la recuperación exhaustiva de muchos datos relevantes; si no se recuperaran documentos relevantes debido al índice vectorial parcial, el sistema no incluiría esos elementos en la lista.

Las puntuaciones ROUGE de respuesta ideal para el sistema de Q&A textual fueron ROUGE-2 F1 = 0,0616 y ROUGE-SU4 F1 = 0,0763. Estas puntuaciones, aunque aparentemente bajas, son típicas en la evaluación de respuestas ideales de BioASQ, ya que incluso las respuestas ideales escritas por humanos pueden diferir en su redacción (Nentidis et al., 2025). Nuestro ROUGE-2 recall de 0,1840 muestra que aproximadamente el 18 % de los bigramas de las respuestas ideales de referencia estaban presentes en nuestras respuestas, lo que representa una cobertura razonable, pero deja margen de mejora en la selección de contenidos.

11.1.3. Resultados de la fase B

En la fase B, del sistema de Q&A textual demostró un buen rendimiento. Las preguntas sí/no se respondieron con un 95,45 % de precisión (macro F1 = 93,94 %), lo que indica un único error en todas las preguntas sí/no del lote 3. La estrategia del sistema de comprobación del contenido de los fragmentos para ver si son afirmativos o negativos fue muy eficaz. La estrategia del sistema de comprobar el contenido del fragmento en busca de pruebas afirmativas o negativas demostró su eficacia. La puntuación F1 de las respuestas "sí" fue del 96,97 %, ligeramente superior a la de las respuestas "no" (90,91 %), lo que sugiere que quizá una pregunta "no" se respondió incorrectamente o con menos confianza.

Cuadro 7: Fase B: Lote 3: Respuestas exactas

	Yes/No				Factoid			List		
System	Acc.	F1 Yes	F1 No	Macro F1	S. Acc.	L. Acc.	MRR	M. Prec.	Recall	F-M
AQAMS2	0.9545	0.9697	0.9091	0.9394	0.3000	0.3500	0.3250	0.6333	0.6456	0.6310

Los resultados en preguntas de tipo factoide del sistema de Q&A textual fueron un 30,00 % de precisión estricta y un 35,00 % de precisión indulgente, con una MRR de 0,3250. Estas puntuaciones son algo inferiores a las obtenidas por el sistema de Q&A textual en la fase A+, lo que parece contraintuitivo dado que la fase B proporciona fragmentos relevantes. Tras un análisis

Cuadro 8: Fase B: Lote 3: Respuestas ideales (puntuaciones Rouge)

R-2 (Rec)	R-2 (F1)	R-SU4 (Rec)	R-SU4 (F1)
0.3567	0.1888	0.3643	0.1795

más detallado, esto puede deberse a que algunas preguntas factoides del Lote 3 requerían una síntesis a partir de varios fragmentos o implicaban formatos de respuesta difíciles (por ejemplo, un nombre de gen específico entre muchos). En varios casos, la respuesta exacta de mi sistema era parcialmente correcta, pero no una coincidencia exacta, lo que afectaba a la precisión estricta. El hecho de que la precisión indulgente fuera un 5 % superior implica que, en algunos casos, la respuesta contenía un elemento correcto, pero omitía otros para factoides con múltiples respuestas aceptables.

Cuadro 9: Fase B: Lote 4: Respuestas exactas

	Yes/No				Factoid			List		
System	Acc.	F1 Yes	F1 No	Macro F1	S. Acc.	L. Acc.	MRR	Mean Prec.	Recall	F-M
AQAMS2	0.9231	0.9500	0.8333	0.8917	0.5455	0.5455	0.5455	0.5904	0.4934	0.5277

Cuadro 10: Fase B: Lote 4: Respuestas ideales (Puntuaciones Rouge)

R-2 (Rec)	R-2 (F1)	R-SU4 (Rec)	R-SU4 (F1)
0.3077	0.1787	0.3274	0.1837

El rendimiento de las preguntas de tipo lista en la fase B fue especialmente notable: el sistema de Q&A textual obtuvo un F1 de 0,6310, con una precisión de ∼0,633 y una recuperación de ∼0,646. Esto representa una mejora sustancial con respecto al rendimiento de la lista de la fase A+. Con los snippets gold disponibles, el sistema pudo identificar la mayoría o la totalidad de los elementos de la lista mencionados, lo que dio lugar a una alta recuperación, manteniendo la precisión. Lo atribuimos al enfoque basado en NER: al extraer entidades de los snippets (fragmentos de artículos), el sistema de Q&A textual pudo enumerar los elementos directamente a partir del texto, minimizando las respuestas perdidas y las adiciones irrelevantes.

Las puntuaciones ROUGE de la respuesta ideal en la fase B también fueron más altas: ROUGE-2 F1 = 0,1888 (aproximadamente 3 veces la del sistema de Q&A textual de la Fase A) y ROUGE-SU4 F1 = 0,1795, con una recuperación de alrededor de 0,36 para ambas métricas. Esto indica que las respuestas ideales generadas por el sistema de Q&A textual tenían un solapamiento sustancialmente mayor con las respuestas de referencia. La recuperación de ROUGE-2 de ∼0,3567 sugiere que nuestras respuestas capturaron aproximadamente el 35 % del contenido de bigramas de las respuestas ideales gold, lo que representa una cobertura significativa.

La mejora en el rendimiento de las respuestas ideales puede explicarse por el contexto simplificado: con snippets limitados centrados en la pregunta, el agente Writer podía incluir más fácilmente hechos relevantes y frases presentes en los snippets, que a menudo se alinean con las respuestas de referencia (ya que esas referencias a menudo se derivan de los mismos snippets). Esencialmente, las respuestas ideales de la Fase B constituyen resúmenes guiados por fragmentos de artículos (snippets) que el sistema gestionó eficazmente utilizando las pruebas proporcionadas.

11.2. Resultados del módulo de *question answering* visual

Las tres tareas fueron evaluadas en los conjuntos de test oficiales de ImageCLEFmedical (Damm & et al., 2025). Se reporta el rendimiento en todas las tareas en las que participó el sistema: Detección de conceptos, Predicción de captions y Explicabilidad. Los resultados incluyen diferentes métricas para cada tarea, como se explicó anteriormente en la Sección 5.12.2. Las Tablas 1, 2 y 3 resumen las métricas de evaluación para nuestro sistema, según lo proporcionado por la herramienta oficial de evaluación de los organizadores de ImageCLEF (Damm & et al., 2025).

11.2.1. Resultados en Detección de conceptos

Para consultar los resultados de Detección de conceptos, véase la Tabla correspondiente.

Cuadro 11: Resultados de la detección de conceptos

Método	F1	F1 secundario
Sistema	0.3982	0.8329

- F1: Puntuación F1 primaria. Media armónica entre precisión y exhaustividad para la tarea principal de detección de conceptos, indicando el equilibrio entre falsos positivos y falsos negativos.
- F1 secundario: Puntuación F1 secundaria. Puntuación F1 para una tarea auxiliar o secundaria de detección de conceptos (por ejemplo, conceptos manuales).

11.2.2. Resultados en Predicción de captions

Para consultar los resultados de Predicción de captions, véase la Tabla correspondiente.

Cuadro 12: Resultados de la predicción de captions

Simil.	BERT	ROUGE-1	BLEURT	P.Rel.	UMLS F1	AlignS	P.Fact.	Global
0.8251	0.5953	0.2389	0.3094	0.4922	0.1366	0.0964	0.1165	0.3043

- Similitud: Similitud de texto entre el caption predicho y el de referencia (por ejemplo, coseno o Levenshtein).
- BERTScore (Recall): Similitud semántica basada en embeddings contextuales de BERT, centrada en el recall.
- ROUGE-1: Superposición de unigramas (palabras individuales) entre la predicción y la referencia; común en tareas de resumen.
- BLEURT: Métrica aprendida para evaluar generación de texto, combinando distintos aspectos de calidad.
- Promedio de relevancia: Promedio de la puntuación de relevancia entre el caption generado y la imagen.
- UMLS Concept F1: Puntuación F1 que mide la coincidencia de conceptos UMLS (Unified Medical Language System) entre predicción y referencia.

- AlignScore: Indica qué tan bien el contenido semántico del caption generado se alinea con la imagen o el caption de referencia.
- Promedio de factualidad: Media de la corrección factual de los captions generados, evaluada automática o manualmente.
- Global: Puntuación global agregada/resumida que sintetiza el rendimiento del modelo en varias métricas.

11.2.3. Resultados en explicabilidad

Para consultar los resultados de la tarea de Explicabilidad, véase la Tabla correspondiente.

Cuadro 13: Tarea de explicabilidad - Resultados de evaluación humana

L.C.	A.Clin.	N.D.C.	F.C.	P.M.C.	Coh. V-T	Comp.V	F.V.	P.M.V.	A.M.	Global
3.4	2.4	2.8	4.1	3.2	1.9	1.9	1.9	1.9	2.0	2.6

- Legibilidad del caption (L.C.): Facilidad de lectura y comprensión del caption.
- Adecuación clínica del caption (A.Clin.): Relevancia y adecuación médica o clínica.
- Nivel de detalle del caption (N.D.C.): Cantidad de detalle relevante en el caption.
- Foco del caption (F.C.): Si el caption se centra en el hallazgo o sujeto principal.
- Puntuación media del caption (P.M.C.): Media de la evaluación humana en todos los criterios del caption.
- Coherencia visual-textual (Coh. V-T): Consistencia y claridad entre la visualización y el caption.
- Completitud de la visualización (Comp.V): Si la visualización es suficientemente completa para la tarea.
- Foco de la visualización (F.V.): Enfoque de la visualización sobre la región o hallazgo clave.
- Puntuación media de la visualización (P.M.V.): Media en los criterios de visualización.
- Adecuación de la metodología (A.M.): Idoneidad de la metodología según expertos humanos.
- Global: Evaluación global humana.

11.3. Resultados operacionales

Métricas de latencias del sistema

Estas métricas se obtuvieron a partir de la ejecución del sistema en un entorno de producción simulado, utilizando un conjunto de datos de prueba representativo del 10 % de los test set. Las métricas se calcularon durante un período de 48 horas con cargas de trabajo típicas.

11.3.1. Latencias del sistema

Componente	Latencia Media (s)	Desviación Estándar (s)
Búsqueda Qdrant	3	0.5
Flujo visual	56	9
Flujo textual	17	3

Cuadro 14: Análisis de latencias por componente del sistema multimodal.

Capacidad de respuesta: >99 % para preguntas de tipo textual y visual.

12. Análisis y Discusión

La evaluación de sistemas de Question Answering (Q&A) biomédico multimodal presenta desafíos metodológicos importantes debido a la ausencia de estándares consolidados que aborden la complejidad inherente de la integración multimodal en dominios especializados. Esta limitación metodológica está muy acentuada en el contexto biomédico, donde la evaluación debe considerar no solo la precisión técnica sino también la relevancia y la interpretabilidad clínica.

Ante esta ausencia de marcos de evaluación estandarizados para sistemas multimodales integrados, recurriré a un enfoque de evaluación modular que permitirá la validación sistemática de componentes individuales antes de proceder a la evaluación del sistema completo.

Considerando la necesidad de obtener retroalimentación de primer nivel en desafíos globales de inteligencia artificial médica, el presente estudio adopta una **metodología de evaluación tripartita** que descompone el análisis en componentes especializados:

12.1. Estructura de evaluación modular

12.1.1. Evaluación cuantitativa del módulo de Question Answering Textual

El componente textual del sistema se someterá a evaluación siguiendo los protocolos establecidos en el **Task 13b de BioASQ** (Nentidis et al., 2025), el cual representa el estándar de facto para la evaluación de sistemas de Q&A biomédico. El detalle de las métricas utilizadas se encuentra en la sección 5.12.1 y proporcionan un baseline comparable con el estado del arte internacional.

La selección del Task 13b de BioASQ se justifica por su adopción generalizada en la comunidad de investigación biomédica y su metodología de evaluación rigurosa que considera tanto la precisión factual como la coherencia clínica de las respuestas generadas.

12.1.2. Evaluación cuantitativa del módulo de Question Answering Visual

El componente visual se evaluará conforme a los criterios establecidos en el **ImageCLEFmedical Caption Task** (Damm & et al., 2025), el cual constituye la competencia internacional más reconocida para la generación de captions médicos automatizados. El detalle de las métricas utilizadas se encuentra en la sección 5.12.2.

La utilización del reto ImageCLEFmedical permite la comparación directa con sistemas especializados en análisis de imágenes médicas, proporcionando contexto para evaluar el rendimiento del componente visual en relación con el estado del arte específico del dominio.

12.1.3. Análisis cualitativo de la integración multimodal

Dado que aún no son comunes los sistemas multimodales aplicados al área médica y a que existe una ausencia de benchmarks específicos, se realizará un análisis cualitativo completo que examine los siguientes puntos:

- **Flujo de Q&A textual**: Evaluación de la capacidad del sistema para generar respuestas coherentes y relevantes a partir de la información textual.
- **Flujo de Q&A visual**: Evaluación de la capacidad del sistema para generar respuestas coherentes y relevantes a partir de la información visual y textual integrada.
- **Coherencia inter-modal**: Evaluación de la coherencia textual-visual y fusión semántica.
- **Arquitectura agéntica**: Evaluación del aporte de la arquitectura agéntica en sistemas biomédicos multimodales.

Benchmarking disciplinarios

La utilización de competencias internacionales establecidas como BioASQ (Nentidis et al., 2025) y ImageCLEFmedical (Damm & et al., 2025) asegura la comparabilidad con trabajos previos y proporciona contexto para la evaluación del rendimiento relativo. Esta aproximación permite situar el rendimiento del sistema desarrollado dentro del panorama competitivo internacional, lo que facilita la identificación de áreas de mejora y fortalezas comparativas.

Contribución a la validación de sistemas multimodales

Esta metodología contribuye al corpus de conocimiento en evaluación de sistemas multimodales al proporcionar un framework replicable para la validación sistemática de sistemas Q&A biomédicos complejos. La adopción de estándares internacionales establecidos para componentes individuales, combinada con análisis cualitativo riguroso para la integración, constituye un aporte metodológico que puede ser adoptado por futuros trabajos en el dominio.

Limitaciones y consideraciones metodológicas

Es importante reconocer las limitaciones inherentes a esta aproximación metodológica. La evaluación modular, aunque rigurosa para componentes individuales, puede no capturar completamente los efectos sinérgicos o antagónicos que emergen de la integración multimodal.

Asimismo, la dependencia de estándares establecidos para componentes individuales puede introducir sesgos hacia arquitecturas y enfoques específicos que han sido optimizados para esas tareas particulares. El análisis cualitativo de la integración busca identificar y documentar estos posibles sesgos para informar futuras iteraciones del sistema.

Cronograma de evaluación

La implementación de esta metodología tripartita seguirá un cronograma estructurado que permite la optimización iterativa del sistema:

1. **Fase 1**: Evaluación cuantitativa del módulo textual (Task 13b BioASQ)
2. **Fase 2**: Evaluación cuantitativa del módulo visual (ImageCLEFmedical Caption)
3. **Fase 3**: Análisis cualitativo de la integración multimodal (Flujo multimodal, coherencia textual-visual, arquitectura agéntica)

Esta secuencia permite la identificación temprana de problemas en componentes individuales antes de proceder al análisis más complejo de la integración, lo que optimiza el uso de recursos y facilita la interpretación de resultados.

12.2. Análisis de rendimiento del módulo de *question answering* textual

12.2.1. Rendimiento en la Fase A y A+

El rendimiento comparativo en la Fase A y A+ con el de la Fase B pone de relieve el impacto significativo de las condiciones de recuperación en la eficacia de la respuesta a las preguntas. Los puntos fuertes del sistema de Q&A textual residen en su enfoque híbrido de recuperación y en la generación basada en preguntas. La alta precisión de sí/no (más del 92 %) en la Fase A+ sugiere que cuando se recuperan pruebas relevantes, el Agente Escritor basado en GPT puede inferir correctamente respuestas de sí/no, una tarea que esencialmente requiere identificar

la presencia o ausencia de una afirmación. La plantilla de preguntas probablemente ayudó al solicitar explícitamente una respuesta de "sí" o "no", abordando uno de los retos señalados por otros equipos de que los modelos generativos a veces producen respuestas inciertas o verbales a preguntas de sí/no.

Otro resultado positivo del sistema de Q&A textual fue su precisión indulgente con los factoides (∼0,41), lo que indica que en aproximadamente el 41 % de las preguntas, la respuesta correcta estaba presente en algún lugar del resultado. La inspección manual reveló que muchos errores de factoides no se debían a fallos completos, sino a problemas de formato y coincidencias parciales. Se trata de problemas comunes en BioASQ que podrían mitigarse mediante un procesamiento posterior (por ejemplo, reconociendo cuándo una respuesta contiene una forma más larga y proporcionando también la forma más corta).

La menor precisión y recuperación de las preguntas de lista en el sistema de Q&A textual pone de manifiesto el reto que supone una recuperación exhaustiva: dado que la base de datos vectorial sólo estaba completa en un 15 %, es probable que algunos elementos de la lista simplemente no se encontraran. Además, la decisión de utilizar sólo los fragmentos recuperados para la generación de respuestas (para evitar la sobrecarga de información en la pregunta) significa que, si los elementos de la lista estuvieran dispersos en muchos documentos, algunos se omitirían. En futuras iteraciones, una posible mejora es incorporar un método de recuperación iterativo o utilizar el propio modelo de lenguaje para sugerir elementos adicionales (por ejemplo, utilizar GPT en una función de lluvia de ideas para predecir otras entidades probables de la lista y, a continuación, verificarlas mediante la recuperación).

12.2.2. Rendimiento en la Fase B

Para el sistema de Q&A textual, el escenario de la Fase B nos permitió centrarnos en la síntesis y justificación de las respuestas. El alto rendimiento en las preguntas de lista puede atribuirse a la estrategia basada en NER. Al extraer todas las entidades de determinados tipos de fragmentos, redujimos la posibilidad de omitir un elemento.

Otra observación de la fase B fue la mejora de las puntuaciones ROUGE en las respuestas ideales. Las respuestas ideales del sistema de Q&A textual se beneficiaron de estar estrechamente ligadas al fraseo del fragmento, realizando de forma efectiva un resumen extractivo. De este modo, se conseguía un alto grado de recall de las frases de referencia, pero a veces se producía una falta de originalidad o una redundancia menor. Por ejemplo, si dos fragmentos de texto se solapaban, la respuesta ideal a veces repetía un hecho. Por lo general, ROUGE lo detectaba (ya que la repetición no mejora el recall más allá de un punto), pero se podía conseguir un resumen más elegante fusionando la información. En esencia, nuestras respuestas ideales en la fase B eran seguras y se ajustaban al tema (lo que se refleja en el buen recall), pero hay margen para hacerlas más concisas e integradas.

12.2.3. Análisis global del módulo de *question answering* textual

Estos resultados demuestran que ambos sistemas son eficaces en sus respectivos escenarios. Más allá del problema de la métrica en la Fase A debido a la API de Pubmed, el pipeline de recuperación del sistema de Q&A textual (Fase A+) permitió una fuerte respuesta sí/no y un rendimiento factoide aceptable en un entorno de dominio abierto, pero tuvo problemas con la amplitud de las preguntas de la lista y tuvo una menor superposición en las respuestas ideales (probablemente debido a diferencias en el estilo de escritura o contenido faltante).

Por otra parte, el sistema de Q&A textual de la Fase B, que funcionaba con los snippets proporcionados, tuvo un desempeño sobresaliente a la hora de proporcionar respuestas completas a las listas y resúmenes ideales con mayor solapamiento, a pesar de una precisión ligeramente inferior en los factoides (un área de mejora para futuras investigaciones).

Es importante señalar que las puntuaciones de BioASQ pueden variar significativamente de un lote a otro y que los resultados publicados son preliminares; sin embargo, los resultados de los sistemas probados fueron coherentes con los de los sistemas de mayor rendimiento en algunos lotes. Por ejemplo, las puntuaciones F1 en preguntas dicotómicas (sí/no) en la Fase A+ y las preguntas tipo lista en la Fase B se encontraban entre las más altas del lote 4 y 3, respectivamente, según la clasificación oficial como se puede consultar en el Anexo C.

A nivel general, el sistema participante es un sistema competitivo de rango medio-alto que demuestra estabilidad. Por último, es importante señalar que he recibido notificación de la aceptación del paper de las working notes para esta Tarea 13b de bioASQ.

12.3. Análisis de rendimiento del módulo de *question answering* visual

12.3.1. Detección de conceptos

El sistema desarrollado alcanzó una puntuación F1 de 0.3982 en la tarea de detección de conceptos, lo que indica un éxito moderado en la identificación de las etiquetas correctas de las imágenes. Este valor es inferior al F1 del mejor equipo (0.5888), lo que sugiere que, aunque mi método identifica muchos conceptos relevantes, omite algunos en comparación con el enfoque líder. Es importante señalar que el sistema tuvo un rendimiento mucho mejor en la métrica secundaria F1 (0.8329 frente a 0.9484 del mejor equipo), la cual se calcula sobre un subconjunto curado de conceptos clave. Estos resultados demuestran una buena cobertura de las características importantes de las imágenes y, con un mayor refinamiento (como una mejor desambiguación de conceptos similares o un mejor recall para hallazgos menos comunes) la F1 primaria del modelo podría acercarse al mejor resultado reportado.

12.3.2. Predicción de captions

El modelo de generación de captions (captions) alcanzó una puntuación global de 0.3043 en la tarea de predicción de captions. Esta métrica, que agrega varios aspectos de la evaluación, representa una brecha de aproximadamente el 10 % de la puntuación global del equipo líder (0.3432). El resultado sugiere que los captions generados por el modelo son en general efectivos, aunque existe una pequeña brecha respecto al mejor desempeño en este reto. El análisis de las métricas individuales proporciona mayor claridad sobre las fortalezas del sistema. El modelo logró una puntuación alta en similitud textual (0.8251) y un BERTScore (Recall) de 0.5953, casi igualando el BERTScore del mejor equipo (0.5977). Estas cifras indican que, en cuanto a solapamiento de contenido y redacción, nuestros captions se asemejan mucho a los informes de referencia y capturan eficazmente las observaciones descritas.

El sistema también obtuvo puntuaciones moderadas en ROUGE-1 (0.2389) y BLEURT (0.3094), lo que refleja un solapamiento razonable con el texto de referencia y una calidad general aceptable de los captions, según estas métricas. Además, con una puntuación promedio de relevancia de 0.4922, los captions generados capturaron una parte sustancial de la información clave de los informes de referencia. En conjunto, estos resultados destacan la fortaleza del sistema para producir descripciones coherentes y relevantes que se alinean con el contenido esperado del dataset.

A pesar de estas fortalezas, nuestro sistema de generación de captions muestra ciertas limitaciones en cuanto a precisión específica de dominio y alineación factual. La puntuación F1 de conceptos UMLS fue 0.1366, notablemente inferior al 0.1816 del mejor equipo, lo que indica que nuestros captions a menudo omiten o identifican erróneamente algunos términos médicos especializados o hallazgos específicos presentes en la referencia.

De forma similar, el AlignScore de 0.0964 (frente a 0.1375 de la mejor solución) sugiere que la alineación entre el contenido de la imagen y el caption podría mejorar; por ejemplo, algunas

descripciones generadas pueden incluir detalles que no están suficientemente respaldados por la evidencia visual.

El promedio de factualidad de 0.1165 también queda por detrás del resultado líder (0.1596), reflejando inconsistencias factuales ocasionales o pequeñas alucinaciones en el texto generado. Estas diferencias señalan oportunidades de mejora: integrar bases de conocimiento médico, refinar la extracción de características visuales o incorporar mecanismos explícitos de verificación factual podrían ayudar al modelo a captar detalles clínicos con mayor precisión y aumentar la veracidad de los captions.

En general, nuestros resultados en la predicción de captions son competitivos en calidad lingüística y relevancia, pero ponen de manifiesto la necesidad de mejoras específicas en la captura de información médica y la garantía de exactitud factual.

12.3.3. Explicabilidad

En la tarea de explicabilidad, un experto humano evaluó la calidad de las explicaciones generadas por cada sistema, considerando tanto el caption textual como la visualización asociada para cada imagen. Las explicaciones de mi sistema lograron una puntuación global de 2.6 sobre 5, en comparación con el 3.2 del mejor equipo. Esto indica un desempeño respetable del método, aunque por debajo del enfoque mejor valorado. Analizando los componentes de la evaluación de explicabilidad, mi enfoque muestra una fortaleza particular en la explicación textual. La puntuación media de calidad del caption fue de 3.2, lo que sugiere que la claridad y utilidad clínica de los captions generados fueron bien valoradas, prácticamente al mismo nivel que el sistema líder en este aspecto.

De forma aún más destacada, mi método obtuvo una puntuación de enfoque del caption de 4.1, mucho más alta que la del mejor equipo (3.3). Esta puntuación excepcionalmente alta indica que los evaluadores humanos consideraron que mis captions se centraron eficazmente en los hallazgos relevantes de la imagen, enfocando de manera precisa el contenido clínico clave a explicar. En otras palabras, el sistema destacó por dirigir la explicación hacia los detalles importantes, lo que constituye una fortaleza crítica en el contexto médico.

Por otro lado, la componente visual de las explicaciones recibió una evaluación relativamente baja. La puntuación media de visualización fue de 1.9, considerablemente inferior a la del mejor equipo (2.8). Esto sugiere que mi estrategia durante el reto para las explicaciones visuales no fue tan clara o informativa como se esperaba. En términos prácticos, lo que sucedió fue que utilicé un método experimental que mezclaba técnicas de PLN biomédico con técnicas de visión computacional y, como usaba más de un modelo, no respondía a explicar el black-box del modelo.

El no haber proporcionado en el reto los resultados de Grad-CAM generó que la metodología del mejor equipo fue valorada en 4.0, mientras que mi enfoque obtuvo una puntuación inferior en este aspecto. Al mejorar la sinergia entre el caption y su justificación visual las futuras versiones del sistema podrían ofrecer una experiencia de explicabilidad más completa y convincente.

Esto me llevó a incorporar Grad-CAM directamente del modelo LLaVA-LLaMA en la versión unificada del sistema que se presenta en este trabajo; es decir, incorporé de forma proactiva el feedback recibido para mejorar el sistema.

En resumen, los resultados de explicabilidad subrayan un fuerte enfoque y relevancia en los captions textuales, lo que representa un punto destacado del sistema, y al mismo tiempo revelan oportunidades claras de mejora en la componente visual de la explicación. Aprovechar estas fortalezas y abordar los problemas de claridad visual en trabajos futuros permitirá proporcionar explicaciones para imágenes médicas más equilibradas y efectivas.

12.3.4. Análisis global de módulo de Q&A visual

En general los resultados del módulo de Q&A visual muestran un rendimiento competitivo en las tareas de detección de conceptos y generación de captions. Aunque con margen de mejora

en la alineación semántica y la precisión factual, el sistema ha sido evaluado en séptimo y sexto lugar entre 75 y 77 participantes (Anexo D), quienes han enviado 107 y 156 submissions para las tareas de concepts y caption, respectivamente («AI4MediaBench», s.f.). Por otra parte, la puntuación de explicabilidad indica que el sistema es capaz de generar explicaciones textuales relevantes y centradas en los hallazgos, pero necesita mejorar la claridad visual para proporcionar explicaciones más completas, lo que se ha abordado en la versión unificada del sistema.

Además, es importante destacar que he sido notificada de la aceptación del paper de las working notes del reto ImageCLEFmedical 2025 (Damm & et al., 2025), donde se presentan los resultados de este módulo de Q&A visual. Esto valida la calidad del trabajo realizado y la relevancia de los resultados obtenidos, los que serán presentados en el congreso CLEF 2025 que se llevará a cabo en Madrid en septiembre de 2025.

12.4. Análisis de latencias del sistema

Para evaluar la efectividad de la integración multimodal, se realizaron pruebas durante 48 horas como se especifica en la sección de Resultados.

El análisis de rendimiento temporal del sistema multimodal revela diferencias sustanciales en las latencias de procesamiento que reflejan la complejidad arquitectónica inherente a cada componente. La búsqueda vectorial mediante Qdrant presenta una latencia media de 3 segundos con una desviación estándar de 0.5 segundos, indicando un rendimiento relativamente estable con baja variabilidad (coeficiente de variación del 17 %). Esta latencia considerable sugiere el procesamiento de un espacio vectorial de alta dimensionalidad con millones de embeddings biomédicos, donde la precisión semántica se prioriza sobre la velocidad de respuesta.

El flujo visual exhibe expectablemente la mayor latencia del sistema con 56 segundos en promedio y una desviación estándar de 9 segundos. Esta latencia sustancial es arquitectónicamente justificada por la ejecución coordinada y secuencial de cuatro métodos de explicabilidad computacionalmente intensivos: Grad-CAM, mapas de atención, descenso de gradiente y generación de bounding boxes. La variabilidad moderada (coeficiente de variación del 16 %) indica un comportamiento predecible del sistema, donde las diferencias en latencia se correlacionan principalmente con la complejidad de las imágenes médicas procesadas y la especificidad de las consultas visuales.

El flujo textual demuestra una eficiencia operacional significativamente superior con una latencia media de 17 segundos y una desviación estándar de 3 segundos (coeficiente de variación del 18 %). Aunque esta latencia es considerablemente mayor que el procesamiento textual convencional, refleja la complejidad del pipeline agéntico que incluye extracción de entidades NER, enriquecimiento semántico UMLS, evaluación contextual para retrieval de literatura PubMed, y fusión inteligente con la base vectorial Qdrant.

Desde una perspectiva de arquitectura distribuida, estos resultados validan el diseño agéntico al evidenciar una diferenciación clara entre modalidades: el flujo visual requiere 3.3x más tiempo que el textual, justificando la implementación de decisiones contextuales para activación selectiva de componentes. La consistencia en las desviaciones estándar (todas entre 16-18 % de coeficiente de variación) indica un sistema robusto con comportamiento predecible, crucial para aplicaciones biomédicas donde la confiabilidad temporal es fundamental para la experiencia del usuario clínico.

Capacidad de respuesta: >99 % para preguntas de tipo textual y visual. La aplicación prácticamente no falla en las pruebas de sistema, pues tiene mecanismos de recuperación ante errores de API u otros.

12.5. Análisis cualitativo de la integración Multimodal

12.5.1. Experiencia de usuario (UX)

La aplicación desarrollada presenta una interfaz unificada que permite al usuario interactuar con contenido multimodal de manera intuitiva y eficiente. El sistema soporta dos modalidades principales de entrada: consultas textuales tradicionales sobre literatura biomédica y análisis de imágenes médicas con preguntas específicas.

En las consultas textuales, el usuario puede formular preguntas médicas complejas de tipo *yes/no*, *list*, *factoid* y *summary* que son procesadas por un pipeline de recuperación híbrido, combinando una base de datos vectorial local (Qdrant) con la API de PubMed como respaldo. Esta estrategia asegura una alta precisión y exhaustividad en la recuperación de información relevante, optimizando la latencia y mejorando la experiencia del usuario.

La experiencia de usuario se caracteriza por un flujo de trabajo simplificado donde este puede seleccionar entre tres tipos de análisis visual: (1) detección de conceptos médicos, (2) generación de descripciones radiológicas, y (3) localización espacial de estructuras anatómicas. Esta taxonomía de tareas refleja las necesidades reales de los profesionales médicos en diferentes contextos clínicos.

La interfaz implementa un sistema de retroalimentación progresiva que informa al usuario sobre el estado del procesamiento, especialmente relevante dado que el análisis multimodal requiere múltiples etapas computacionales. El sistema proporciona estimaciones de tiempo e indicadores de progreso específicos para cada modalidad, mejorando significativamente la experiencia de usuario en comparación con sistemas de procesamiento tradicionales.

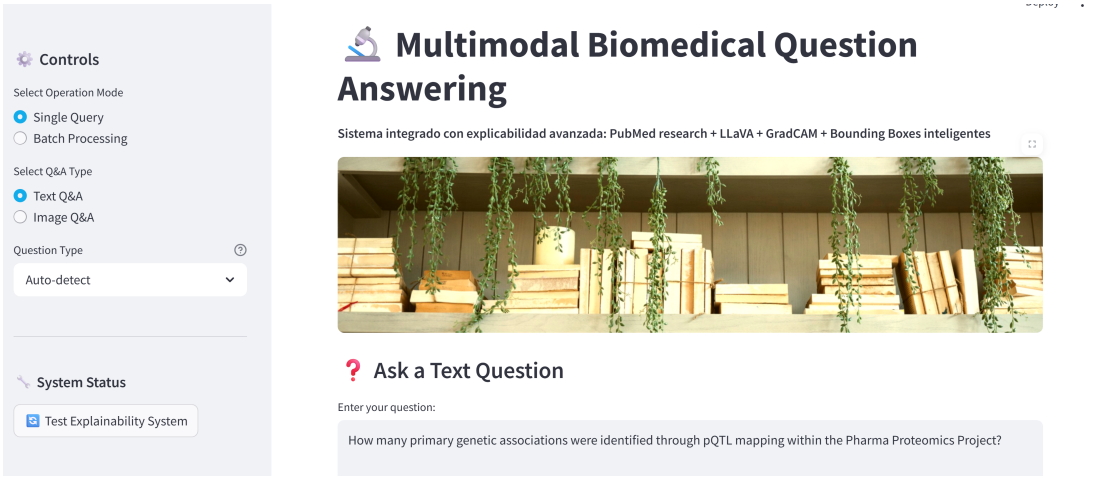
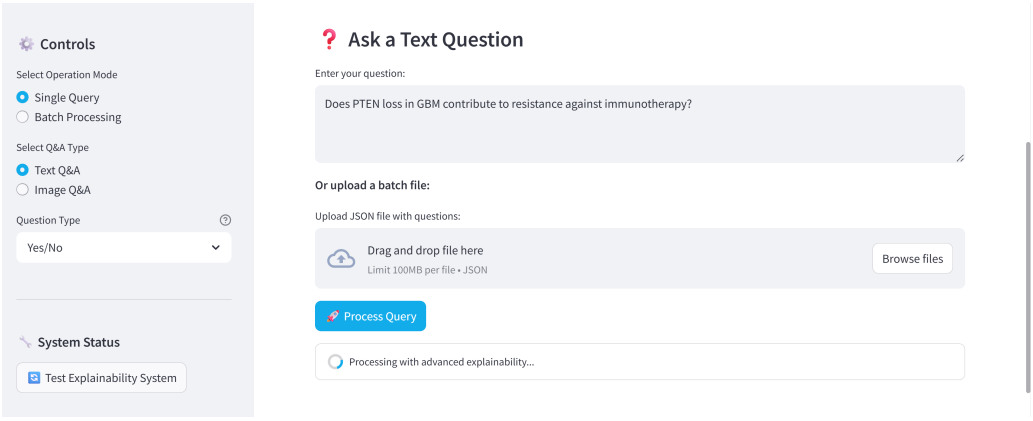


Figura 9: Interfaz principal del Sistema Multimodal de Q&A Biomédico. Permite selección entre preguntas de texto y preguntas de imágenes.

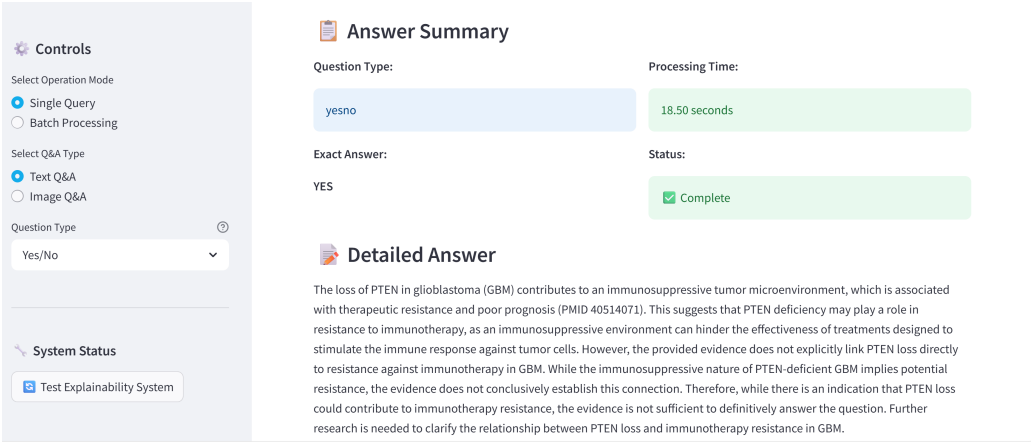
12.5.2. Caso de uso de Q&A textual

A continuación, se presenta un caso de uso representativo del sistema de Question Answering textual, ilustrando su capacidad para responder preguntas complejas sobre literatura biomédica:

El sistema tiene incorporado un chatbot en el flujo de Q&A textual. Cuando el usuario olvida ingresar el tipo de pregunta aparece un chatbot integrado con Streamlit que evalúa qué tipo de pregunta es (*yes/no*, *list*, *factoid* o *summary*) y le presenta al usuario su opción. Si el usuario no está de acuerdo puede cambiarla.

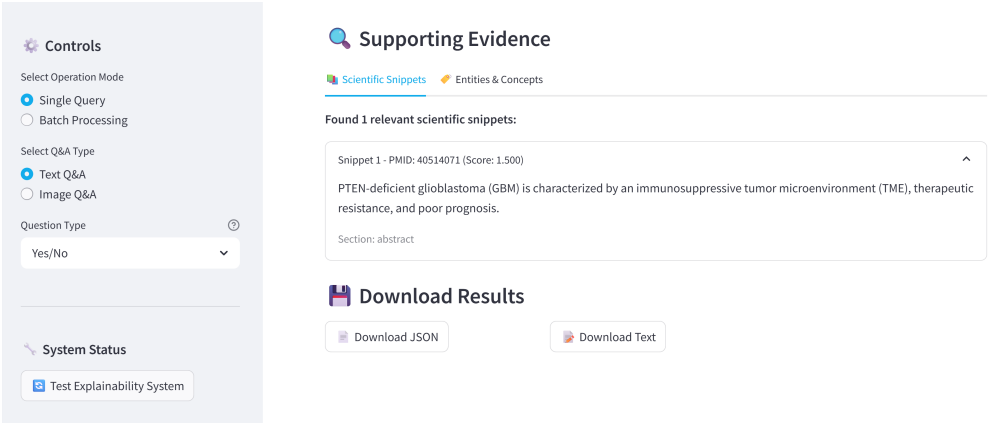


(a) Paso 1 de flujo textual: El sistema procesa la pregunta realizada.

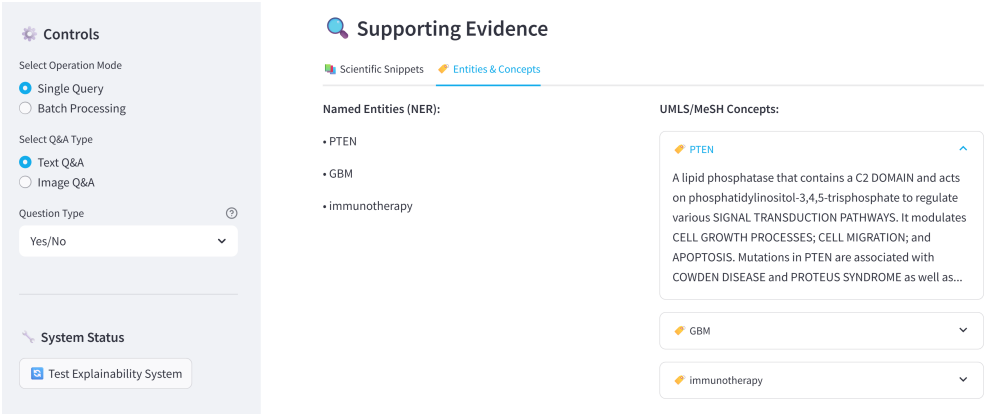


(b) Paso 2 de flujo textual: Entrega respuesta exacta según tipo de pregunta (Yes/No) y la respuesta ideal (Detailed Answer) con los números de PMID.

Figura 10: Flujo textual del sistema - Parte I: Procesamiento de pregunta y generación de respuesta.



(a) Paso 3 de flujo textual: Se pueden ver los snippets con el PMID del artículo en el que se basa la respuesta.



(b) Paso 4 de flujo textual: Se pueden ver las NER extraídas y los conceptos UMLS/MeSH.

Figura 11: Flujo textual del sistema - Parte II: Visualización de evidencia y análisis conceptual.

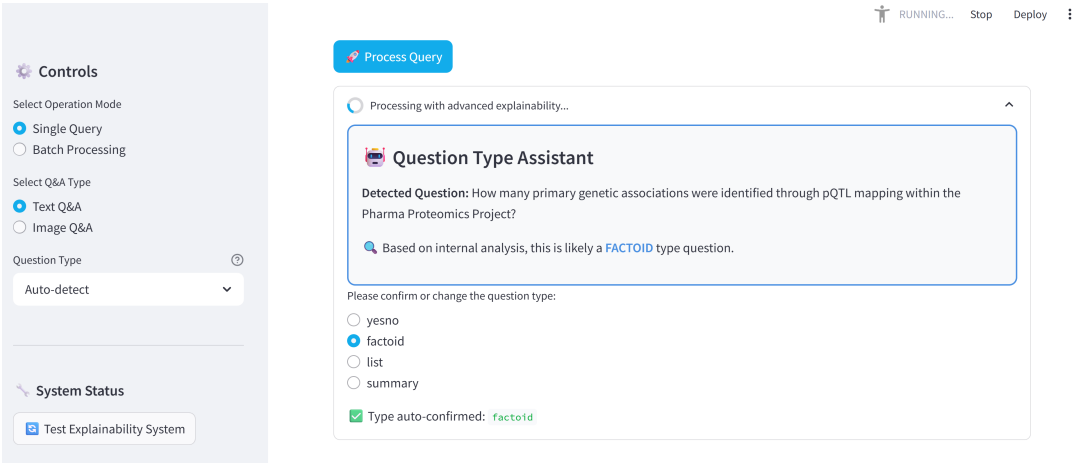


Figura 12: Chatbot: aparece de forma contextual en la interfaz de usuario del sistema.

12.5.3. Caso de uso de Q&A visual

En el flujo de Q&A visual, el usuario puede seleccionar entre tres tipos de análisis: detección de conceptos, generación de captions y localización espacial de estructuras anatómicas. El sistema proporciona una interfaz intuitiva que permite al usuario cargar imágenes médicas y recibir respuestas detalladas sobre los hallazgos presentes en las mismas. Por lo tanto, a continuación, se presenta un caso de uso representativo del sistema de Question Answering visual, ilustrando su capacidad para analizar imágenes médicas:

12.5.4. Explicabilidad de imágenes médicas

En el flujo de Q&A visual también se pueden ver los mapas de calor generados por el sistema, lo que permite contextualizar las respuestas del modelo. Esto puede ser muy útil a nivel clínico, ya que los profesionales de la salud pueden verificar visualmente las áreas de interés en la imagen médica y comprender mejor cómo el modelo llegó a sus conclusiones.

12.5.5. Fusión de modalidades y coordinación

Un aspecto interesante para analizar es que el sistema **MultimodalBioQA** implementa una **arquitectura híbrida de fusión** que opera en múltiples niveles jerárquicos, representando una aproximación novedosa en sistemas de Question Answering biomédico:

Fusión a nivel de modelo individual

El modelo LLaVA-LLaMA implementa **fusión temprana/intermedia** dentro de su arquitectura interna las características visuales extraídas por el vision encoder (Vision Transformer), las que se proyectan al espacio de embeddings del modelo de lenguaje mediante una capa de proyección lineal. Esta integración ocurre en las capas intermedias del transformer, permitiendo que el modelo de lenguaje procese simultáneamente información visual y textual durante la generación autorregresiva.

Formalmente, dado un conjunto de tokens visuales $\mathbf{V} = \{v_1, v_2, ..., v_n\}$ y tokens textuales $\mathbf{T} = \{t_1, t_2, ..., t_m\}$, el modelo realiza:

$$\mathbf{H}_{\text{fused}} = \text{LLM}([\mathbf{V}; \mathbf{T}]) \tag{1}$$

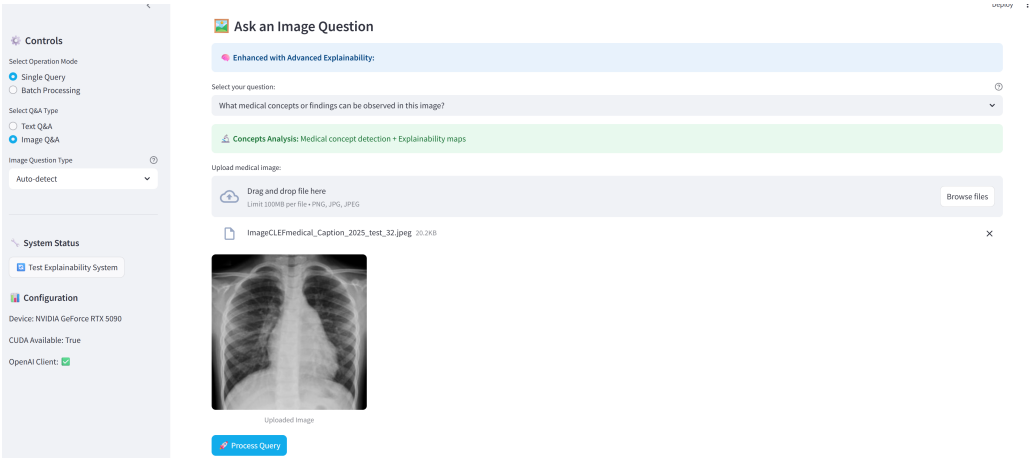
donde $[\cdot;\cdot]$ denota concatenación en la dimensión de secuencia.

Fusión a nivel de arquitectura del sistema

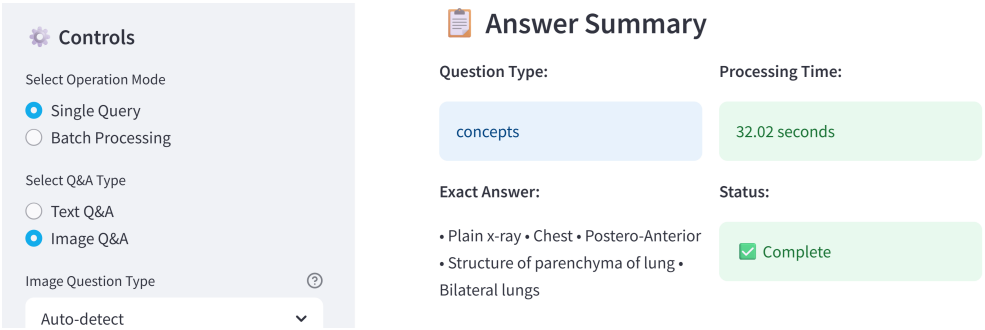
El sistema completo emplea una estrategia de **fusión tardía** entre pipelines especializados, donde las modalidades textuales y visuales se procesan mediante rutas independientes antes de la integración final de resultados. Esta decisión arquitectónica permite:

- Optimización independiente de cada pipeline especializado
- Escalabilidad modular del sistema
- Trazabilidad diferenciada por modalidad
- Flexibilidad en la gestión de recursos computacionales

Para consultas textuales, el sistema emplea una estrategia de búsqueda híbrida que prioriza la base de datos vectorial local Qdrant antes de recurrir a la API de PubMed como mecanismo de respaldo. Este enfoque optimiza la latencia mientras se mantiene la exhaustividad de la búsqueda.

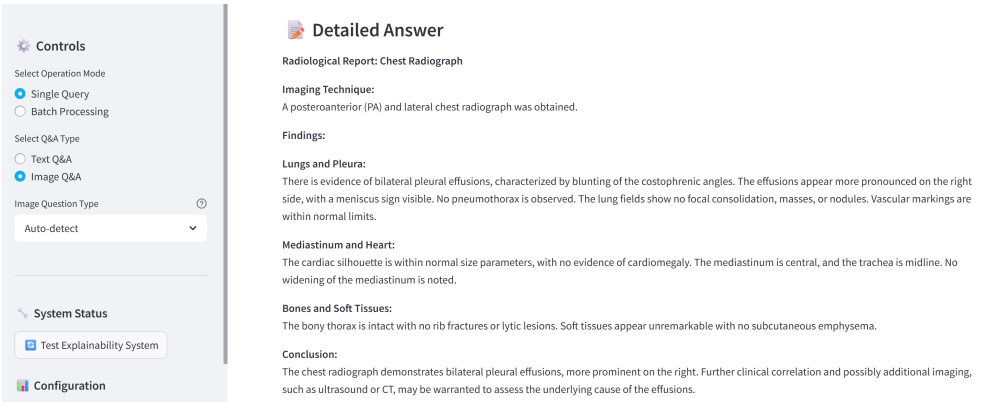


(a) Paso 1 de flujo visual: Se selecciona la pregunta y se carga la imagen.



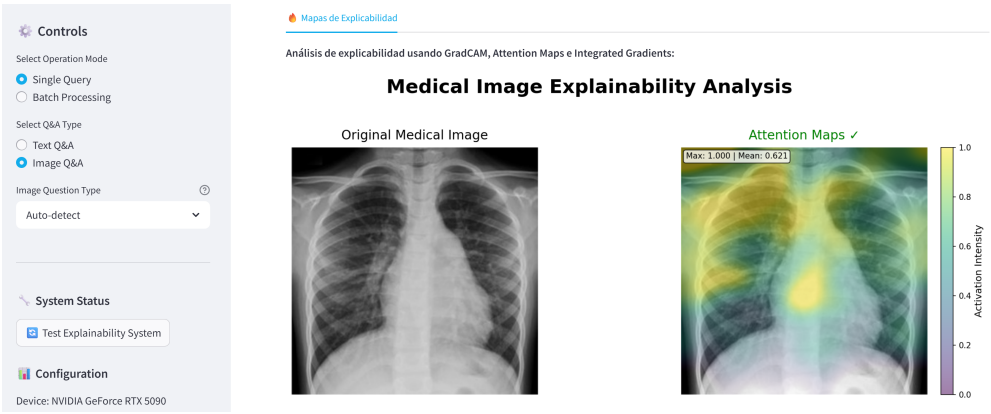
(b) Paso 2 de flujo visual: Se detectan los conceptos presentes en la imagen.

(c) Paso 3 de flujo visual: Se predice la caption de la imagen.

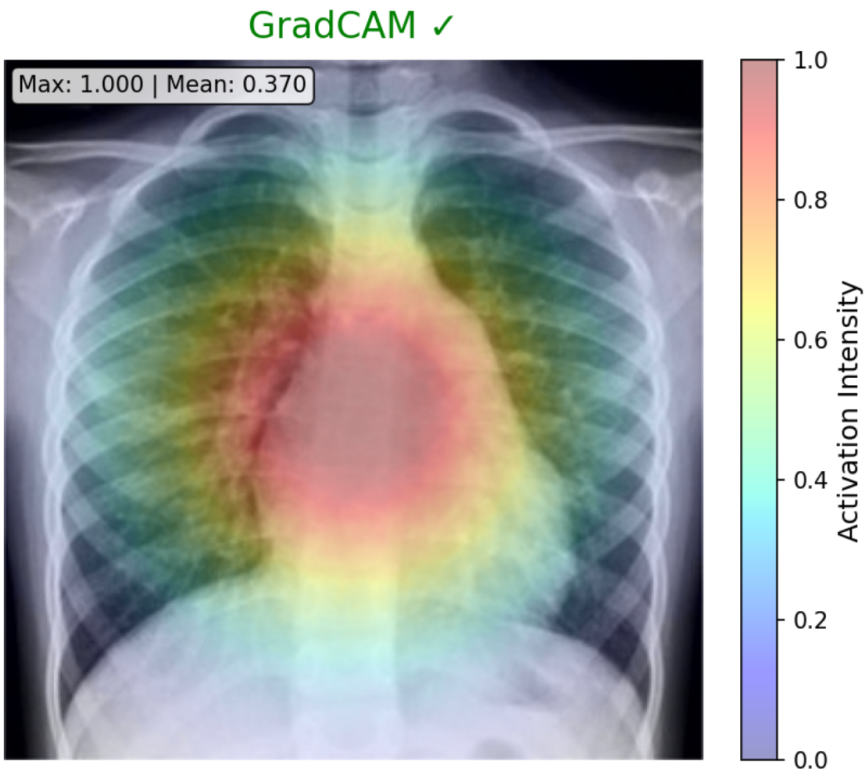


(d) Paso 4 de flujo visual: Se entrega una respuesta detallada (ideal).

Figura 13: Flujo visual del sistema: interacción paso a paso con la imagen.



(a) Mapas de calor: Se selecciona la pregunta y se carga la imagen.



(b) Mapas de calor: Grad-CAM.

Figura 14: Visualización de mapas de calor para explicabilidad visual en el sistema.

En el procesamiento visual, el sistema integra múltiples técnicas de explicabilidad (GradCAM, Attention Maps, Integrated Gradients) con generación de bounding boxes utilizando GPT-4o Vision API y Segment Anything (SAM). La coordinación entre estos componentes se gestiona mediante un sistema de metadatos o productos intermedios de datos que permite la trazabilidad completa del proceso de inferencia.

12.5.6. Coherencia textual-visual y fusión semántica

En cuanto a la **coherencia textual-visual**, esta se gestiona mediante un sistema de metadatos estructurados que permite la trazabilidad completa del proceso de inferencia. Son varios archivos .JSON que el sistema produce para ir dejando registro del paso a paso e incluso algunos de ellos son descargables en la interfaz de usuario.

La **fusión semántica final** se realiza a nivel del Writer Agent, que sintetiza información procedente de ambas modalidades utilizando prompts estructurados que preservan la coherencia interdisciplinaria.

El sistema mantiene la separación clara entre respuestas exactas (`exact_answer`) e ideales (`ideal_answer`), facilitando la evaluación automatizada según estándares de benchmarks biomédicos como BioASQ (Tsatsaronis et al., 2015).

Para comprobar la coherencia inter-modelo podemos comparar si la respuesta ideal o detallada del sistema de Q&A visual, que fue generada por el Writer Agent utilizando la API de GTP-4o se condice con el caption generado por el modelo LLaVA-LLaMA. En este caso de ejemplo que la radiografía de tórax que estamos utilizando podemos ver que la respuesta ideal del sistema de Q&A visual es:

> *"Radiological Report: Chest Radiograph*
>
> *Imaging Technique:*
> *A posteroanterior (PA) and lateral chest radiograph was obtained.*
>
> *Findings:*
> *Lungs and Pleura:*
> *There is evidence of bilateral pleural effusions, characterized by blunting of the costophrenic angles. The effusions appear more pronounced on the right side, with a meniscus sign visible. No pneumothorax is observed. The lung fields show no focal consolidation, masses, or nodules. Vascular markings are within normal limits."*

Esto coincide con la caption generada por el modelo LLaVA-LLaMA, que es:

> *"Exact Answer (Caption): Chest radiograph showing bilateral pleural effusions."*

Con este ejemplo se puede apreciar que al existir esta coherencia entre la respuesta ideal del sistema de Q&A visual y el caption generado por LLaVA-LLaMA nos demuestra la efectividad de la fusión semántica y la integración de modalidades en el sistema. Esto es muy importante porque nos asegura que las respuestas generadas sean consistentes y relevantes para el contexto clínico.

Flujo de procesamiento multimodal

Para ilustrar la integración completa del sistema, consideremos una consulta representativa de cada modalidad y observemos los diagramas de flujo correspondientes.

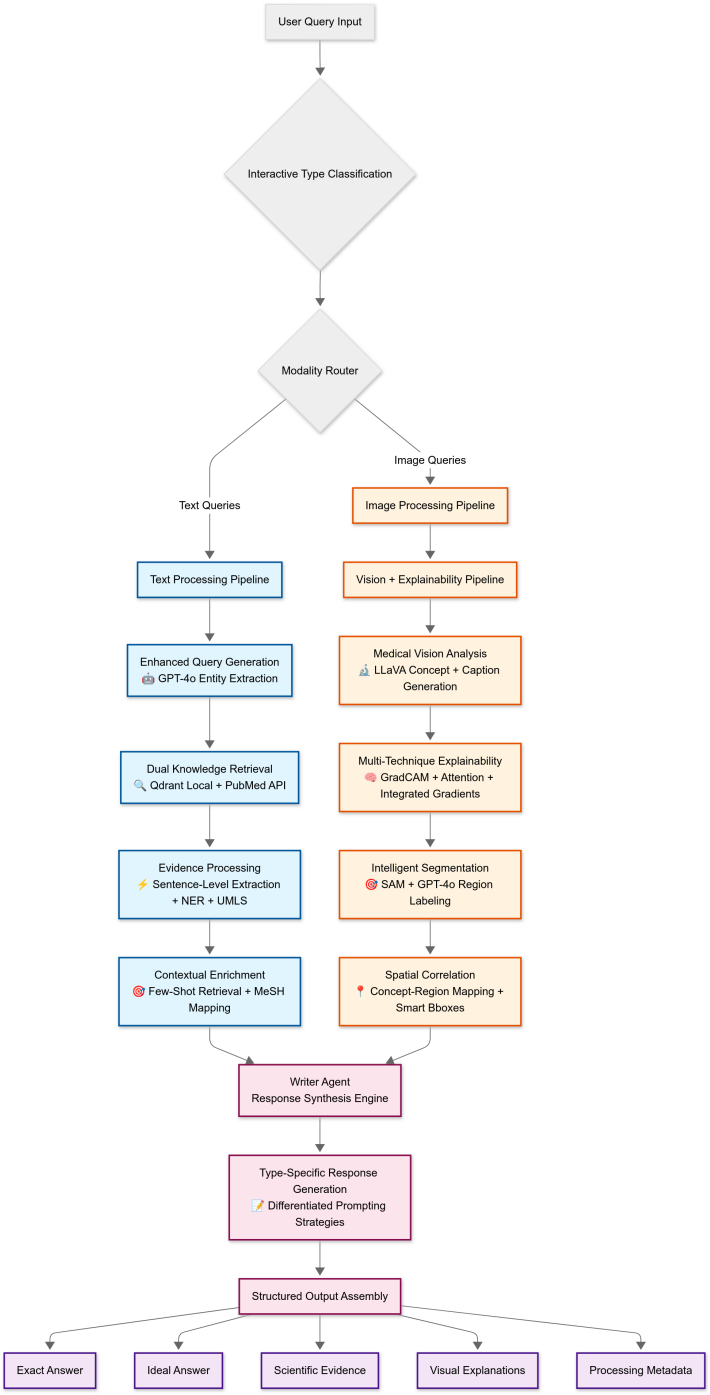


Figura 15: Arquitectura general del sistema multimodal de Q&A biomédica. Flujo de datos bidireccional mostrando: (a) Pipeline textual con búsqueda híbrida Qdrant-PubMed, (b) Pipeline visual con LLaVA y explicabilidad, (c) Fusión tardía en Writer Agent para síntesis final.

Flujo Textual

Pregunta: "List primary sclerosing cholangitis effector genes and biological mechanisms"
El procesamiento sigue la siguiente secuencia:

1. **Extracción de entidades**: GPT-4o identifica "primary sclerosing cholangitis", "effector genes", "biological mechanisms"
2. **Búsqueda híbrida**: Qdrant local ($n = 10$ snippets, scores: 0.259-0.240) + PubMed API fallback ($n = 4$ nuevos artículos)
3. **Enriquecimiento semántico**: Consulta UMLS/MeSH para contexto terminológico
4. **Síntesis**: Writer Agent genera respuesta estructurada con citas PMID

Como se muestra en las figuras, el pipeline de procesamiento textual sigue una secuencia de cuatro etapas principales.

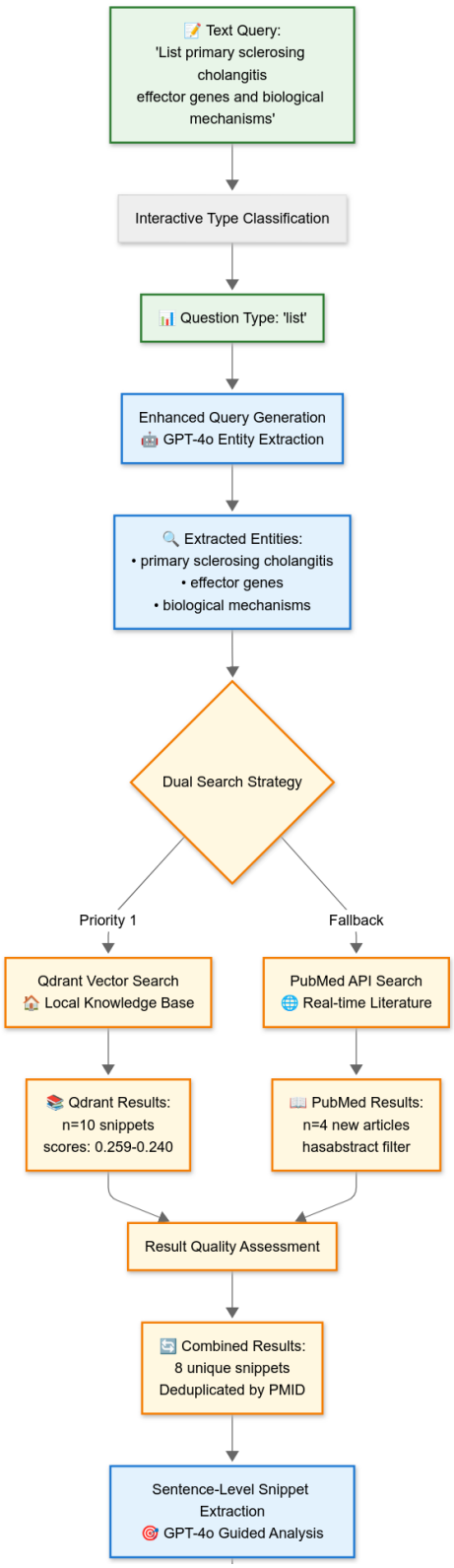


Figura 16: Diagrama detallado del procesamiento textual, Parte 1. El flujo ilustra las primeras dos de cuatro etapas secuenciales: (a) extracción de entidades biomédicas utilizando GPT-4o, (b) búsqueda híbrida que combina bases de datos vectoriales Qdrant con PubMed.

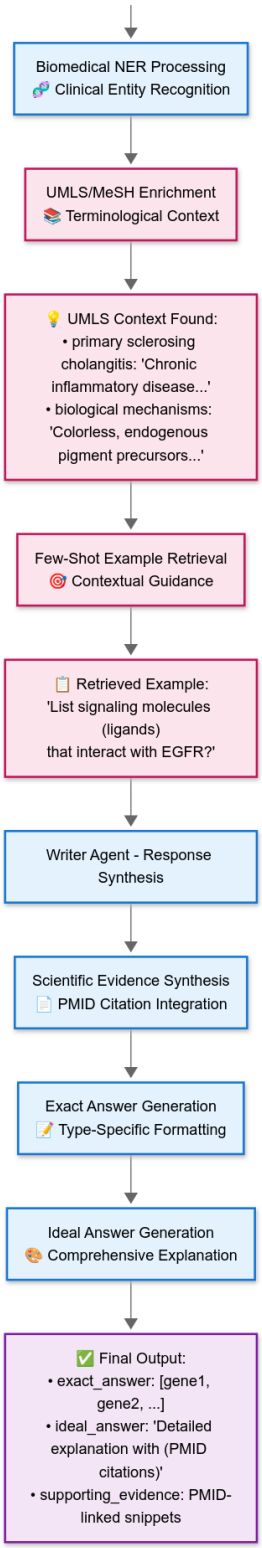


Figura 17: Diagrama detallado del procesamiento textual, Parte 2. El flujo ilustra las segundas dos de cuatro etapas secuenciales: (c) enriquecimiento semántico mediante conceptos UMLS, y (d) síntesis final de respuestas con citas bibliográficas.

Flujo visual

Pregunta: "Where are the key structures or findings located in this medical image?"

El procesamiento visual integra múltiples componentes especializados:

1. **Análisis conceptual**: LLaVA-LLaMA identifica conceptos médicos mediante fusión temprana visión-lenguaje.
2. **Explicabilidad avanzada**: Generación simultánea de GradCAM, Attention Maps e Integrated Gradients.
3. **Localización espacial**: GPT-4o Vision API + SAM generan bounding boxes automáticas con fallback a modelo fine-tuneado para dar coordenadas y que se encuentra a nivel local.
4. **Síntesis radiológica**: Writer Agent genera reporte clínico estructurado como informe radiológico.

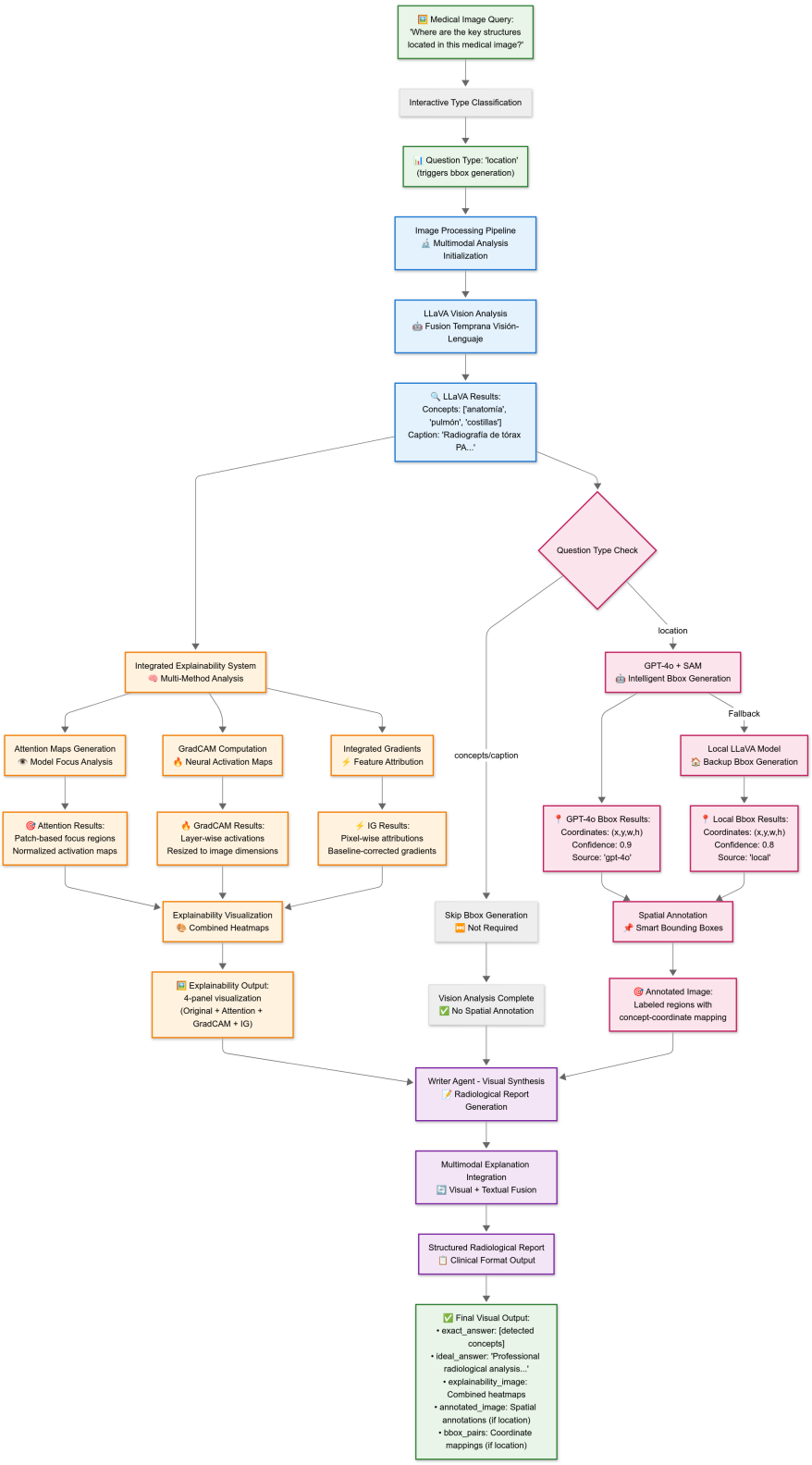


Figura 18: Flujo completo de procesamiento visual: (a) imagen de entrada, (b) extracción de conceptos con LLaVA, (c) explicabilidad multimethod, (d) localización con GPT-4o + SAM, (e) síntesis radiológica final.

12.5.7. Explicabilidad y trazabilidad

El sistema implementa mecanismos de interpretabilidad diferenciados según la modalidad de entrada, representando uno de los aspectos más innovadores de la integración desarrollada.

Interpretabilidad textual

Un aspecto en el que fui más allá de los requisitos del reto de bioASQ fue en la implementación de un sistema de trazabilidad textual que permite al usuario verificar la fuente de cada afirmación generada. Esto se logra mediante la inclusión de identificadores PMID en la respuesta ideal (que es un resumen). Los PMID referencian a artículos científicos específicos dependiendo de los snippets obtenidos. Es por este motivo que cada respuesta ideal generada por el Writer Agent en el flujo de Q&A textual incluye citas inline que permiten al usuario rastrear la evidencia científica subyacente. Esto es sumamente importante en el área médica porque cada afirmación en la respuesta ideal está respaldada por fragmentos científicos identificables, permitiendo la verificación independiente de las fuentes.

Además, el sistema integra contexto UMLS/MeSH para la desambiguación terminológica, proporcionando definiciones contextuales de conceptos médicos complejos. Esta integración semántica aparte de ser útil como contexto en los prompts para generar las respuestas, también facilita la interpretación clínica al proporcionar el conocimiento ontológico necesario para la comprensión de términos especializados.

Interpretabilidad visual

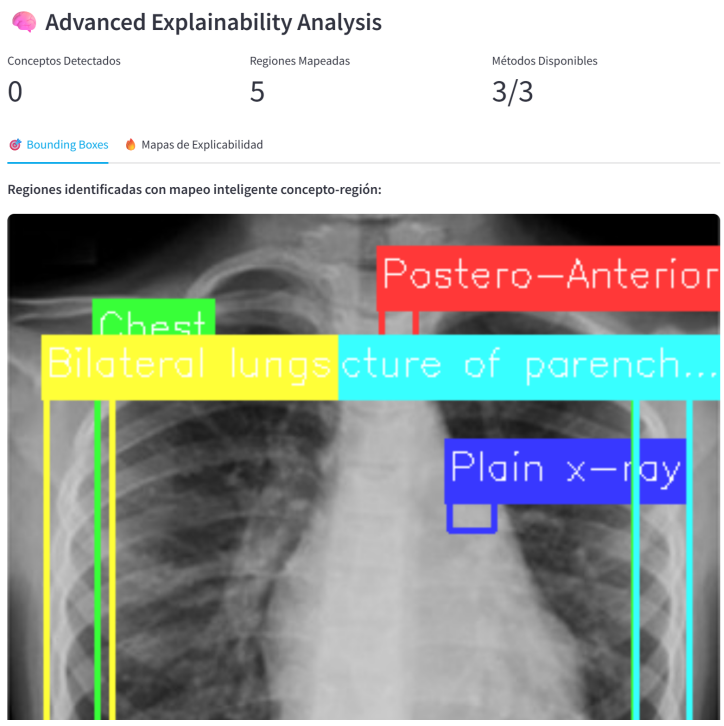
La explicabilidad visual representa una contribución técnica importante de este sistema, ya que integra múltiples técnicas complementarias que proporcionan diferentes perspectivas del proceso de inferencia:

- **Attention maps**: Visualización de patrones de atención del modelo LLaVA fine-tuned, mostrando regiones de la imagen que contribuyen más significativamente a la generación de conceptos.
- **GradCAM**: Implementación optimizada con selección de capas objetivo y manejo robusto de gradientes para modelos multimodales.
- **Integrated gradients**: Cálculo de atribuciones pixel-wise con línea base optimizada y manejo de artefactos de integración numérica.
- **Bounding boxes**: Generación condicional de regiones de interés utilizando GPT-4o API + SAM con modelo LLaVA local como respaldo.

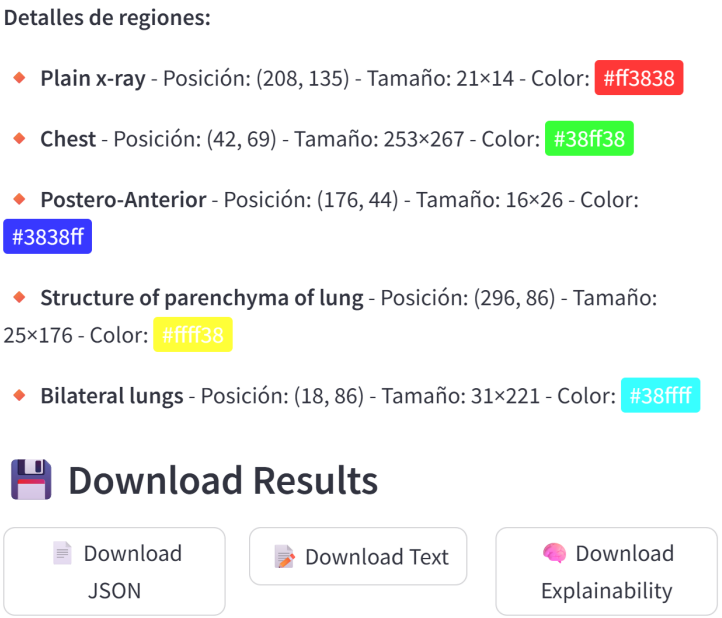
La integración de estos métodos proporciona una perspectiva multifacética del proceso de inferencia visual, permitiendo la validación cruzada de interpretaciones y aumentando la confianza clínica en los resultados.

12.5.8. Evaluación cualitativa de arquitectura agéntica

Los pipelines de procesamiento de lenguaje natural tradicionales son estáticos y monolíticos: aplican la misma secuencia de procesamiento independientemente de las características de entrada. En cambio, la arquitectura agéntica que he utilizado implementa una orquestación dinámica del flujo de trabajo, donde agentes especializados toman decisiones contextuales sobre qué estrategias de procesamiento aplicar basándose en las propiedades del texto. Esto es fundamentalmente diferente de los enfoques tradicionales, por ejemplo, en la integración de herramientas



(a) Bounding boxes para la imagen médica seleccionada.



(b) Regiones identificadas en la imagen con sus posiciones y etiquetas.

Figura 19: Análisis de bounding boxes en imágenes médicas.

heterogéneas, ya que el marco agéntico permite la integración fluida de herramientas complementarias (NER, grafos de conocimiento, métodos de explicabilidad, etc.) en un flujo de trabajo coordinado, en lugar de hacerlo como pasos de post-procesamiento independientes.

La diferenciación arquitectónica del sistema **MultimodalBioQA** se puede enmarcar en los siguientes elementos:

Orquestación adaptativa de modalidades

Los sistemas tradicionales procesan modalidades de forma secuencial y predeterminada. En cambio, la arquitectura agéntica del sistema **MultimodalBioQA** implementa decisión contextual inteligente:

Diferencia clave: La decisión de qué pathway activar (textual o visual) y cómo coordinar las explicaciones emerge del análisis contextual, no de reglas estáticas.

Explicabilidad multimodal coordinada

Para preguntas visuales - Agente de Explicabilidad: Aplica 4 métodos complementarios de forma orquestada:

- **Grad-CAM:** Localización de atención neuronal
- **Attention Maps:** Patrones de atención del transformer
- **Gradient Descent:** Optimización de características relevantes
- **Bounding Boxes:** Mapeo concepto-región anatómica

La coordinación sincronizada de estos 4 métodos y su integración coherente en la respuesta final requiere orquestación inteligente que no puede ser predeterminada.

Para preguntas textuales - Agente Writer (Grounding): Proporciona evidencia estructurada:

- PMIDs de artículos fuente
- Entidades NER extraídas
- Términos UMLS recuperados y enriquecidos

Enriquecimiento dinámico de contexto

El pipeline textual implementa cascada de decisiones inteligentes: El sistema implementa un flujo de decisión inteligente que comienza con la extracción de entidades biomédicas de la pregunta mediante el uso de GTPT-4o por parte del Agente Researcher. Estas entidades extraídas son posteriormente enriquecidas con términos UMLS por parte del agente Researcher, que consulta la API para obtener información contextual adicional. A continuación, el agente Researcher recupera artículos recientes relevantes mediante la API de Pubmed y fusiona esta información con la base de datos vectorial Qdrant existente para crear un contexto híbrido. Una vez establecido el contexto apropiado, se aplica few-shot seleccionando los ejemplos contextualmente relevantes que mejor se alineen con la pregunta y el contexto disponible. Finalmente, el agente Writer sintetiza la respuesta final incorporando tanto el contexto híbrido como los ejemplos seleccionados, asegurando que la respuesta esté fundamentada en evidencia verificable. Este flujo demuestra la toma de decisiones distribuida y contextual que caracteriza la arquitectura agéntica, donde cada agente especializado contribuye de manera coordinada al resultado final.

Ventajas arquitectónicas específicas

1. **Coherencia multimodal:** Coordinación inteligente entre explicabilidad visual (4 métodos) y grounding textual (PMIDs + NER + UMLS)
2. **Adaptabilidad contextual:** El sistema ajusta su estrategia de procesamiento basado en características específicas de cada pregunta biomédica
3. **Integración herramientas heterogéneas:** NER → UMLS → PubMed → Qdrant → Few-shot funciona como workflow coordinado, no como pasos independientes
4. **Grounding robusto:** Proporciona evidencia verificable (PMIDs, entidades NER, términos UMLS) de forma contextualmente relevante

El sistema **MultimodalBioQA** demuestra que la coordinación agéntica es arquitectónicamente necesaria para Q&A biomédico multimodal porque:

- **Complejidad irreducible:** La interacción entre modalidades, explicabilidad y grounding no puede ser predeterminada
- **Decisiones contextuales:** Cada consulta biomédica requiere decisiones específicas sobre enriquecimiento, retrieval, y explicabilidad
- **Optimización dinámica:** La eficiencia computacional requiere activación inteligente de componentes según necesidad

12.6. Conclusiones de análisis cualitativo

La integración multimodal desarrollada representa un avance hacia sistemas de Question Answering biomédico más explicables, demostrando la viabilidad técnica de combinar análisis textual y visual en una plataforma unificada que opera efectivamente tanto a nivel de modelo individual como de arquitectura de sistema.

Por lo tanto, mi conclusión es que la arquitectura agéntica no es una mejora incremental, sino una solución arquitectónica necesaria para la complejidad inherente de sistemas biomédicos multimodales con explicabilidad robusta y grounding verificable. La orquestación inteligente de agentes especializados permite abordar problemas que son arquitectónicamente irresolubles con enfoques tradicionales monolíticos.

12.6.1. Contribuciones principales

- **Búsqueda híbrida**: Estrategia Qdrant-first que optimiza latencia manteniendo exhaustividad.
- **Arquitectura híbrida de fusión**: Sistema con fusión multinivel (temprana en LLaVA + tardía en sistema) para Q&A biomédico.
- **Explicabilidad multimodal integrada**: Integración de 4 métodos complementarios de explicabilidad visual con trazabilidad textual completa.
- **Arquitectura agéntica**: Permite coherencia multimodal, adaptabilidad contextual, integración de herramientas heterogéneas y grounding robusto.

13. Desafíos y trabajo futuro

El análisis cualitativo revela varias categorías de limitaciones que caracterizan el estado actual de la integración multimodal:

13.1. Limitaciones de modalidad visual

- **Alucinaciones conceptuales**: El modelo LLaVA ocasionalmente genera conceptos médicos plausibles, pero no presentes en la imagen analizada, particularmente en imágenes con artefactos o calidad subóptima.
- **Inconsistencias en bounding boxes**: La generación de coordenadas espaciales mediante GPT-4o muestra variabilidad en la precisión, especialmente para estructuras anatómicas con límites difusos.
- **Limitaciones de resolución**: El procesamiento de imágenes de alta resolución requiere redimensionamiento que puede resultar en pérdida de detalles diagnósticos críticos.

13.2. Limitaciones de integración cross-modal

- **Conflictos semánticos**: En casos donde la información visual sugiere hallazgos que contradicen el conocimiento textual previo, el sistema carece de mecanismos sofisticados de resolución de conflictos. Este es un aspecto para mejorar en futuras versiones del sistema.
- **Coherencia temporal**: Para casos clínicos evolutivos, el sistema no mantiene coherencia temporal entre análisis secuenciales de la misma patología. Esta es un área de mejora importante, ya que la evolución de una patología es clave para el diagnóstico y tratamiento médico.
- **Calibración de incertidumbre**: Las métricas de confianza entre modalidades no están calibradas, dificultando la interpretación de inconsistencias intermodales.

13.3. Limitaciones de escalabilidad

- **Dependencia de APIs externas**: La integración con GPT-4o introduce latencia variable y dependencia de servicios terceros. Por este motivo se implementó un modelo LLaVA local que permite al sistema funcionar sin depender de la API de OpenAI, pero aún está a nivel experimental.
- **Gestión de memoria**: El procesamiento simultáneo de múltiples modalidades puede resultar en limitaciones de memoria GPU para análisis batch.
- **Paralelización limitada**: La arquitectura actual no optimiza completamente el procesamiento paralelo de componentes independientes.

13.4. Mejoras técnicas propuestas

Se podrían realizar múltiples mejoras al sistema **MultimodalBioQA**, ya que es un área de investigación activa. Sin embargo, las que propongo son de mi interés personal y profesional, ya que creo constituyen un aporte en el área de la IA biomédica.

1. **Explicabilidad temporal**: Extender análisis de explicabilidad a imágenes médicas temporales (videos o series de imágenes) para capturar dinámicas evolutivas.
2. **Multimodalidad avanzada**: Integración de datos genómicos y proteómicos.

3. **Knowledge-Graph**: Construcción de un grafo de conocimiento (knowledge-graph) propio alimentado con ontologías médicas especializadas. Comencé a trabajar en este aspecto durante el reto de BioASQ con una base de grafos en neo4j que alimenté con datos de UMLS, pero no pude mantener el sistema debido a falta de recursos. Sin embargo, creo que es un aspecto clave para mejorar tanto la interpretabilidad del sistema como la calidad de los resultados.

4. **Personalización clínica**: Adaptación de respuestas basada en contexto clínico específico. Sería interesante personalizar el sistema como apoyo radiológico en un hospital o clínica, ya que muy motivante a nivel técnico la adopción de protocolos clínicos específicos en forma de guardrails que guíen al sistema a responder de forma más contextualizada y precisa para el contexto clínico en el que se esté utilizando.

13.5. Investigación futura

En concordancia con lo expresado en la sección anterior, propongo las siguientes líneas de investigación futura para abordar las limitaciones identificadas y potenciar el sistema:

1. **Ayuda al diagnóstico con explicabilidad temporal**: Desarrollo de modelos que integren análisis temporal de imágenes médicas para capturar dinámicas evolutivas de patologías, mejorando la capacidad diagnóstica del sistema. Es algo que ya he comenzado a investigar y que creo puede ser un aporte significativo al área de la IA biomédica. También puede servir como validación del diagnóstico médico, ya que permite al profesional de la salud verificar la evolución de una patología a través del tiempo.

2. **Validación clínica**: Estudios con profesionales médicos para validar utilidad clínica.

3. **Multimodalidad extendida**: Esta es mi mayor línea de investigación futura, ya que creo que la integración de datos genómicos y proteómicos con imágenes médicas y texto clínico puede revolucionar el campo de la IA biomédica. Esto permitiría que el sistema realizara un análisis más completo y preciso de los pacientes, lo que a su vez mejoraría la capacidad diagnóstica y pronóstica del sistema. En esta misma línea de investigación pienso integrar los grafos para mantener las relaciones entre los datos y mejorar la interpretabilidad del sistema.

14. Conclusiones

En síntesis, se ha presentado **MultimodalBioQA**, un sistema de *question answering* biomédico multimodal con explicabilidad incorporada, que representa un avance en la capacidad de acceder y comprender la información biomédica. Entre sus contribuciones más relevantes se incluyen las siguientes:

14.1. Contribuciones académicas y metodológicas

He resumido las contribuciones académicas y metodológicas del sistema de Q&A Biomédico Multimodal con Explicabilidad en las siguientes categorías:

1. **Integración de explicabilidad multimodal**: Implementación que combina LLaVA, GradCAM, *Attention Maps* e *Integrated Gradients* en un pipeline unificado para análisis y explicabilidad a nivel médico.

2. **Mapeo Concepto-Región**: Método que correlaciona automáticamente conceptos médicos extraídos por LLMs con regiones espaciales específicas de imágenes.

3. **Extracción de snippets guiada por LLM**: Metodología innovadora que utiliza GPT-4o para extraer evidencia científica a nivel de oración, lo que constituye una alternativa a las limitaciones de métodos basados en *similarity scoring*.

14.2. Contribuciones tecnológicas

También considero que el sistema **MultimodalBioQA** ha realizado contribuciones tecnológicas que pueden ser de utilidad para la comunidad científica y profesional en el área de la IA biomédica:

1. **Arquitectura híbrida de búsqueda**: Combinación efectiva de búsqueda vectorial local con APIs en tiempo real para maximizar cobertura y actualidad de la información.
2. **Pipeline de ingesta masiva**: Sistema capaz de procesar millones de artículos PubMed con uso de memoria constante mediante sistema de *batching*.
3. **Sistema de Few-Shot Learning contextual**: Implementación de recuperación semántica de ejemplos para few-shot learning que mejora significativamente la consistencia de respuestas en dominios especializados.

Desde mi punto de vista, este sistema de Q&A Biomédico Multimodal con Explicabilidad representa una contribución al campo de IA médica al integrar con éxito múltiples modalidades de información (texto e imagen) y proporcionar explicabilidad a través de un enfoque unificado que combina:

1. **Procesamiento multimodal** con análisis de texto e imagen especializado.
2. **Explicabilidad multitécnica** que combina diversos métodos complementarios de interpretabilidad.
3. **Recuperación de información con IA generativa** mediante extracción guiada por LLM.
4. **Arquitectura multiagente escalable** con optimizaciones específicas para datos biomédicos masivos.

La combinación de estas innovaciones crea un sistema que no solo proporciona respuestas precisas, sino que también ofrece explicabilidad integrada, estableciendo un modelo para herramientas de IA en medicina que equilibren rendimiento, interpretabilidad y utilidad clínica.

En mi opinión se han cumplido los resultados esperados, ya que se logró precisión en la recuperación de información biomédica mediante arquitectura RAG híbrida. Además, se ha demostrado la capacidad de procesamiento multimodal de flujos textuales y visuales con explicabilidad integrada para análisis de imágenes médicas mediante la implementación de un sistema unificado multiagente para procesamiento biomédico textual y visual.

Por lo tanto, puedo decir que he quedado conforme con este trabajo, ya que cumplió con los objetivos planteados inicialmente demostrando que es posible desarrollar un sistema de *question answering* biomédico multimodal con explicabilidad incorporada, lo que representa un avance en la capacidad de acceder y comprender la información biomédica.

Referencias

AI4MediaBench. (s.f.).

Al-hammuri, K., Gebali, F., Kanan, A., & Chelvan, I. T. (2023). Vision Transformer Architecture and Applications in Digital Health: A Tutorial and Survey. *Visual Computing for Industry, Biomedicine, and Art*, *6*(1), 14. https://doi.org/10.1186/s42492-023-00140-9

Amugongo, L. M., Mascheroni, P., Brooks, S., Doering, S., & Seidel, J. (2025). Retrieval augmented generation for large language models in healthcare: A systematic review. *PLOS Digit Health*, *4*(6), e0000877. https://doi.org/10.1371/journal.pdig.0000877

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., & Sherlock, G. (2000). Gene Ontology: Tool for the unification of biology. *Nat Genet*, *25*(1), 25-29. https://doi.org/10.1038/75556

Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI. *Information Fusion*, *58*, 82-115. https://doi.org/10.1016/j.inffus.2019.12.012

Beddiar, D.-R., Oussalah, M., & Seppänen, T. (2023). Automatic captioning for medical imaging (MIC): A rapid review of literature. *Artif Intell Rev*, *56*(5), 4019-4076. https://doi.org/10.1007/s10462-022-10270-w

Bi, Y., Abrol, A., Fu, Z., & Calhoun, V. D. (2024). A multimodal vision transformer for interpretable fusion of functional and structural neuroimaging data. *Hum Brain Mapp*, *45*(17), e26783. https://doi.org/10.1002/hbm.26783

Bodenreider, O. (2004). The Unified Medical Language System (UMLS): Integrating Biomedical Terminology. *Nucleic Acids Research*, *32*(suppl_1), D267-D270. https://doi.org/10.1093/nar/gkh061

Borys, K., Schmitt, Y. A., Nauta, M., Seifert, C., Krämer, N., Friedrich, C. M., & Nensa, F. (2023). Explainable AI in medical imaging: An overview for clinical practitioners - Beyond saliency-based XAI approaches. *Eur J Radiol*, *162*, 110786. https://doi.org/10.1016/j.ejrad.2023.110786

Boulahia, S. Y., Amamra, A., Madi, M. R., & Daikh, S. (2021). Early, intermediate and late fusion strategies for robust deep learning-based multimodal action recognition. *Machine Vision and Applications*, *32*(6), 121. https://doi.org/10.1007/s00138-021-01249-8

Bradshaw, J. M. (Ed.). (1997). *Software Agents*. MIT Press.

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., Nori, H., Palangi, H., Ribeiro, M. T., & Zhang, Y. (2023, abril). Sparks of Artificial General Intelligence: Early Experiments with GPT-4. https://doi.org/10.48550/arXiv.2303.12712

Cai, L., Fang, H., & Li, Z. (2023). Pre-Trained Multilevel Fuse Network Based on Vision-Conditioned Reasoning and Bilinear Attentions for Medical Image Visual Question Answering. *J. Supercomput.*, *79*(12), 13696-13723. https://doi.org/10.1007/s11227-023-05195-2

Cálem, J., Moreira, C., & Jorge, J. (2024). Intelligent systems in healthcare: A systematic survey of explainable user interfaces. *Comput Biol Med*, *180*, 108908. https://doi.org/10.1016/j.compbiomed.2024.108908

Chandak, P., Huang, K., & Zitnik, M. (2023). Building a knowledge graph to enable precision medicine. *Sci Data*, *10*(1), 67. https://doi.org/10.1038/s41597-023-01960-3

Chen, Q., Hu, Y., Peng, X., Xie, Q., Jin, Q., Gilson, A., Singer, M. B., Ai, X., Lai, P.-T., Wang, Z., Keloth, V. K., Raja, K., Huang, J., He, H., Lin, F., Du, J., Zhang, R., Zheng, W. J., Adelman, R. A., ... Xu, H. (2025). Benchmarking large language models for biomedical natural language processing applications and recommendations. *Nat Commun*, *16*(1), 3280. https://doi.org/10.1038/s41467-025-56989-2

Croatti, A., Gabellini, M., Montagna, S., & Ricci, A. (2020). On the Integration of Agents and Digital Twins in Healthcare. *J Med Syst*, *44*(9), 161. https://doi.org/10.1007/s10916-020-01623-5

Damm, H., & et al. (2025). Overview of ImageCLEFmedical 2025 - Medical Concept Detection and Interpretable Caption Generation. *CEUR-WS.org*.

Dettmers, T., Pagnoni, A., Holtzman, A., & Zettlemoyer, L. (2023, mayo). QLoRA: Efficient Finetuning of Quantized LLMs [Comment: Extended NeurIPS submission]. https://doi.org/10.48550/arXiv.2305.14314

Durfee, E., & Lesser, V. (1991). Partial Global Planning: A Coordination Framework for Distributed Hypothesis Formation. *IEEE Transactions on Systems, Man, and Cybernetics*, *21*(5), 1167-1183. https://doi.org/10.1109/21.120067

Ferber, J. (1998). *Multi-agent systems : An introduction to distributed artificial intelligence*. Harlow : Addison-Wesley.

Finin, T., Fritzson, R., McKay, D., & McEntire, R. (1994). KQML as an agent communication language. *Proceedings of the Third International Conference on Information and Knowledge Management - CIKM '94*, 456-463. https://doi.org/10.1145/191246.191322

Gao, Y., Li, R., Croxford, E., Caskey, J., Patterson, B. W., Churpek, M., Miller, T., Dligach, D., & Afshar, M. (2025). Leveraging Medical Knowledge Graphs Into Large Language Models for Diagnosis Prediction: Design and Application Study. *JMIR AI*, *4*(1), e58670. https://doi.org/10.2196/58670

Gaudet-Blavignac, C., Foufi, V., Bjelogrlic, M., & Lovis, C. (2021). Use of the Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT) for Processing Free Text in Health Care: Systematic Scoping Review. *Journal of Medical Internet Research*, *23*(1), e24594. https://doi.org/10.2196/24594

Genesereth, M. R., & Nilsson, N. J. (2014). *Logical Foundations of Artificial Intelligence*. Morgan Kaufmann.

Guarrasi, V., Aksu, F., Caruso, C. M., Feola, F. D., Rofena, A., Ruffini, F., & Soda, P. (2025). A Systematic Review of Intermediate Fusion in Multimodal Deep Learning for Biomedical Applications. *Image and Vision Computing*, *158*, 105509. https://doi.org/10.1016/j.imavis.2025.105509

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A Survey of Methods for Explaining Black Box Models. *ACM Comput. Surv.*, *51*(5), 93:1-93:42. https://doi.org/10.1145/3236009

Gupta, D., Suman, S., & Ekbal, A. (2021). Hierarchical Deep Multi-Modal Network for Medical Visual Question Answering. *Expert Systems with Applications*, *164*, 113993. https://doi.org/10.1016/j.eswa.2020.113993

Huang, S.-C., Shen, L., Lungren, M. P., & Yeung, S. (2021). GLoRIA: A Multimodal Global-Local Representation Learning Framework for Label-efficient Medical Image Recognition. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 3922-3931. https://doi.org/10.1109/ICCV48922.2021.00391

Ionescu, B. (2025). Overview of ImageCLEF 2025: Multimedia Retrieval in Medical, Social Media and Content Recommendation Applications. *Springer Lecture Notes in Computer Science (LNCS*.

Isern, D., Sánchez, D., & Moreno, A. (2010). Agents applied in health care: A review. *Int J Med Inform*, *79*(3), 145-166. https://doi.org/10.1016/j.ijmedinf.2010.01.003

Jennings, N. R. (1999). On agent-based software engineering I. *Artificial Intelligence*, (117 (2000)), 277-296.

Jin, Q., Yuan, Z., Xiong, G., Yu, Q., Ying, H., Tan, C., Chen, M., Huang, S., Liu, X., & Yu, S. (2022). Biomedical Question Answering: A Survey of Approaches and Challenges. *ACM Comput. Surv.*, *55*(2), 35:1-35:36. https://doi.org/10.1145/3490238

Jing, X. (2021a). The Unified Medical Language System at 30 Years and How It Is Used and Published: Systematic Review and Content Analysis. *JMIR Med Inform*, *9*(8), e20675. https://doi.org/10.2196/20675

Jing, X. (2021b). The Unified Medical Language System at 30 Years and How It Is Used and Published: Systematic Review and Content Analysis. *JMIR Medical Informatics*, *9*(8), e20675. https://doi.org/10.2196/20675

Kang, T., Perotte, A., Tang, Y., Ta, C., & Weng, C. (2021). UMLS-based data augmentation for natural language processing of clinical research literature. *Journal of the American Medical Informatics Association*, *28*(4), 812-823. https://doi.org/10.1093/jamia/ocaa309

Kell, G., Roberts, A., Umansky, S., Qian, L., Ferrari, D., Soboczenski, F., Wallace, B. C., Patel, N., & Marshall, I. J. (2024). Question Answering Systems for Health Professionals at the Point of Care—a Systematic Review. *Journal of the American Medical Informatics Association*, *31*(4), 1009-1024. https://doi.org/10.1093/jamia/ocae015

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., Dollár, P., & Girshick, R. (2023, abril). Segment Anything [Comment: Project web-page: https://segment-anything.com]. https://doi.org/10.48550/arXiv.2304.02643

Krithara, A., Nentidis, A., Bougiatiotis, K., & Paliouras, G. (2023a). BioASQ-QA: A Manually Curated Corpus for Biomedical Question Answering. *Scientific Data*, *10*(1), 170.

Krithara, A., Nentidis, A., Bougiatiotis, K., & Paliouras, G. (2023b). BioASQ-QA: A manually curated corpus for Biomedical Question Answering. *Sci Data*, *10*(1), 170. https://doi.org/10.1038/s41597-023-02068-4

Kumar, S., Rani, S., Sharma, S., & Min, H. (2024). Multimodality Fusion Aspects of Medical Diagnosis: A Comprehensive Review. *Bioengineering (Basel)*, *11*(12), 1233. https://doi.org/10.3390/bioengineering11121233

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2020). BioBERT: A Pre-Trained Biomedical Language Representation Model for Biomedical Text Mining. *Bioinformatics*, *36*(4), 1234-1240. https://doi.org/10.1093/bioinformatics/btz682

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. *Advances in Neural Information Processing Systems*, *33*, 9459-9474.

Li, Y., Daho, M. E. H., Conze, P.-H., Zeghlache, R., Boité, H. L., Tadayoni, R., Cochener, B., Lamard, M., & Quellec, G. (2024, abril). A Review of Deep Learning-Based Information Fusion Techniques for Multimodal Medical Image Classification. https://doi.org/10.48550/arXiv.2404.15022

Liang, Z., Mao, J., Lu, K., & Li, G. (2021). Finding citations for PubMed: A large-scale comparison between five freely available bibliographic data sources. *Scientometrics*, *126*(12), 9519-9542. https://doi.org/10.1007/s11192-021-04191-8

Liu, H., Li, C., Wu, Q., & Lee, Y. J. (2023, diciembre). Visual Instruction Tuning [Comment: NeurIPS 2023 Oral; project page: https://llava-vl.github.io/]. https://doi.org/10.48550/arXiv.2304.08485

Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., & Neubig, G. (2023). Pre-Train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing. *ACM Comput. Surv.*, *55*(9), 195:1-195:35. https://doi.org/10.1145/3560815

Markus, A. F., Kors, J. A., & Rijnbeek, P. R. (2021). The role of explainability in creating trustworthy artificial intelligence for health care: A comprehensive survey of the terminology, design choices, and evaluation strategies. *J Biomed Inform*, *113*, 103655. https://doi.org/10.1016/j.jbi.2020.103655

Matsumoto, N., Moran, J., Choi, H., Hernandez, M. E., Venkatesan, M., Wang, P., & Moore, J. H. (2024). KRAGEN: A Knowledge Graph-Enhanced RAG Framework for Biomedical Problem Solving Using Large Language Models. *Bioinformatics*, *40*(6), btae353. https://doi.org/10.1093/bioinformatics/btae353

Mialon, G., Dessì, R., Lomeli, M., Nalmpantis, C., Pasunuru, R., Raileanu, R., Rozière, B., Schick, T., Dwivedi-Yu, J., Celikyilmaz, A., Grave, E., Lecun, Y., & Scialom, T. (2023). Augmented Language Models: A Survey [Publisher Copyright: © 2023, Transactions on Machine Learning Research. All rights reserved.]. *Transactions on Machine Learning Research*, *2023*.

Mueller, S. T., Hoffman, R., & Clancey, W. (s.f.). (PDF) Explanation in Human-AI Systems: A Literature Meta-Review, Synopsis of Key Ideas and Publications, and Bibliography for Explainable AI. https://doi.org/10.48550/arXiv.1902.01876

Murphy, R. (2019). *Introduction to AI Robotics*.

Nentidis, A., Katsimpras, G., Krithara, A., Krallinger, M., Rodríguez-Ortega, M., Rodriguez-López, E., Loukachevitch, N., Sakhovskiy, A., Tutubalina, E., Dimitriadis, D., Tsoumakas, G., Giannakoulas, G., Bekiaridou, A., Samaras, A., Maria Di Nunzio, F., Giorgio, Marchesin, S., Martinelli, M., Silvello, G., & Paliouras, G. (2025). Overview of BioASQ 2025: The Thirteenth BioASQ Challenge on Large-Scale Biomedical Semantic Indexing and Question Answering. En J. Carrillo-de-Albornoz, J. Gonzalo, L. Plaza, A. G. S. de Herrera, J. Mothe, F. Piroi, P. Rosso, D. Spina, G. Faggioli & N. Ferro (Eds.), *Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Sixteenth International Conference of the CLEF Association (CLEF 2025)*.

Nentidis, A., Katsimpras, G., Krithara, A., López, S. L., Farré-Maduell, E., Gasco, L., Krallinger, M., & Paliouras, G. (2023). Overview of BioASQ 2023: The Eleventh BioASQ Challenge on Large-Scale Biomedical Semantic Indexing and Question

Answering [Comment: 24 pages, 12 tables, 3 figures. CLEF2023. arXiv admin note: text overlap with arXiv:2210.06852]. https://doi.org/10.1007/978-3-031-42448-9_19

Nicholson, D. N., & Greene, C. S. (2020). Constructing knowledge graphs and their biomedical applications. *Computational and Structural Biotechnology Journal*, *18*, 1414-1428. https://doi.org/10.1016/j.csbj.2020.05.017

Novoa, J., Chagoyen, M., Benito, C., Moreno, F. J., & Pazos, F. (2023). PMIDigest: Interactive Review of Large Collections of PubMed Entries to Distill Relevant Information. *Genes (Basel)*, *14*(4), 942. https://doi.org/10.3390/genes14040942

Oates, B. J. (2006). *Researching Information Systems and Computing*. SAGE Publications Ltd.

ONU. (s.f.). Salud.

Organización Mundial de la Salud. (2021). *Estrategia Mundial Sobre Salud Digital 2020-2025* (1st ed). World Health Organization.

Öztürk, E., & Mesut, A. (2024). PERFORMANCE ANALYSIS OF CHROMA, QDRANT, AND FAISS DATABASES. *UNSP*. https://doi.org/10.70456/TBRN3643

Pandey, H., Amod, A., & Shivang. (2024, julio). Advancing Healthcare Automation: Multi-Agent System for Medical Necessity Justification [Comment: Accepted at BioNLP2024]. https://doi.org/10.48550/arXiv.2404.17977

Park, J. S., O'Brien, J. C., Cai, C. J., Morris, M. R., Liang, P., & Bernstein, M. S. (2023, abril). Generative Agents: Interactive Simulacra of Human Behavior.

Robinson, P. N., Köhler, S., Bauer, S., Seelow, D., Horn, D., & Mundlos, S. (2008). The Human Phenotype Ontology: A Tool for Annotating and Analyzing Human Hereditary Disease. *The American Journal of Human Genetics*, *83*(5), 610-615. https://doi.org/10.1016/j.ajhg.2008.09.017

Rückert, J., Bloch, L., Brüngel, R., Idrissi-Yaghir, A., Schäfer, H., Schmidt, C. S., Koitka, S., Pelka, O., Abacha, A. B., G Seco de Herrera, A., Müller, H., Horn, P. A., Nensa, F., & Friedrich, C. M. (2024). ROCOv2: Radiology Objects in COntext Version 2, an Updated Multimodal Image Dataset. *Sci Data*, *11*(1), 688. https://doi.org/10.1038/s41597-024-03496-6

Schick, T., Dwivedi-Yu, J., Dessi, R., Raileanu, R., Lomeli, M., Hambro, E., Zettlemoyer, L., Cancedda, N., & Scialom, T. (2023). Toolformer: Language Models Can Teach Themselves to Use Tools. *Thirty-Seventh Conference on Neural Information Processing Systems*.

Schouten, D., Nicoletti, G., Dille, B., Chia, C., Vendittelli, P., Schuurmans, M., Litjens, G., & Khalili, N. (2025). Navigating the Landscape of Multimodal AI in Medicine: A Scoping Review on Technical Challenges and Clinical Applications. *Medical Image Analysis*, *105*, 103621. https://doi.org/10.1016/j.media.2025.103621

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2020). Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization [Comment: This version was published in International Journal of Computer Vision (IJCV) in 2019; A previous version of the paper was published at International Conference on Computer Vision (ICCV'17)]. *Int J Comput Vis*, *128*(2), 336-359. https://doi.org/10.1007/s11263-019-01228-7

Simon, B. D., Ozyoruk, K. B., Gelikman, D. G., Harmon, S. A., & Türkbey, B. (2024). The future of multimodal artificial intelligence models for integrating imaging and clinical metadata: A narrative review. *Diagn Interv Radiol*. https://doi.org/10.4274/dir.2024.242631

Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., Chung, H. W., Scales, N., Tanwani, A., Cole-Lewis, H., Pfohl, S., Payne, P., Seneviratne, M., Gamble, P., Kelly, C., Babiker, A., Schärli, N., Chowdhery, A., Mansfield, P., Demner-Fushman, D., ... Natarajan, V. (2023). Large language models encode clinical knowledge. *Nature*, *620*(7972), 172-180. https://doi.org/10.1038/s41586-023-06291-2

Smith, R. G. (1980). The Contract Net Protocol: High-Level Communication and Control in a Distributed Problem Solver. *IEEE TRANSACTIONS ON COMPUTERS*, (12).

Soman, K., Rose, P. W., Morris, J. H., Akbas, R. E., Smith, B., Peetoom, B., Villouta-Reyes, C., Cerono, G., Shi, Y., Rizk-Jackson, A., Israni, S., Nelson, C. A., Huang, S., & Baranzini, S. E. (2024). Biomedical Knowledge Graph-Optimized Prompt Generation for Large Language Models. *Bioinformatics*, *40*(9), btae560. https://doi.org/10.1093/bioinformatics/btae560

Song, B., Li, F., Liu, Y., & Zeng, X. (2021). Deep learning methods for biomedical named entity recognition: A survey and qualitative comparison. *Briefings in Bioinformatics*, *22*(6), bbab282. https://doi.org/10.1093/bib/bbab282

Stahlschmidt, S. R., Ulfenborg, B., & Synnergren, J. (2022). Multimodal Deep Learning for Biomedical Data Fusion: A Review. *Briefings in Bioinformatics*, *23*(2), bbab569. https://doi.org/10.1093/bib/bbab569

Stone, P., & Veloso, M. (2000). Multiagent Systems: A Survey from a Machine Learning Perspective. *Autonomous Robots*, *8*(3), 345-383. https://doi.org/10.1023/A:1008942012299

Stuart, R., & Peter, N. (2021). *Artificial Intelligence: A Modern Approach*.

Tambe, M. (1997). Towards Flexible Teamwork. *Journal of Artificial Intelligence Research*, *7*, 83-124. https://doi.org/10.1613/jair.433

Tariq, A., Banerjee, I., Trivedi, H., & Gichoya, J. (2025). Multimodal Artificial Intelligence Models for Radiology. *BJR|Artificial Intelligence*, *2*(1), ubae017. https://doi.org/10.1093/bjrai/ubae017

Thirunavukarasu, A. J., Ting, D. S. J., Elangovan, K., Gutierrez, L., Tan, T. F., & Ting, D. S. W. (2023). Large language models in medicine. *Nat Med*, *29*(8), 1930-1940. https://doi.org/10.1038/s41591-023-02448-8

Tsatsaronis, G., Balikas, G., Malakasiotis, P., Partalas, I., Zschunke, M., Alvers, M. R., Weissenborn, D., Krithara, A., Petridis, S., Polychronopoulos, D., Almirantis, Y., Pavlopoulos, J., Baskiotis, N., Gallinari, P., Artiéres, T., Ngomo, A.-C. N., Heino, N., Gaussier, E., Barrio-Alvers, L., ... Paliouras, G. (2015). An Overview of the BIOASQ Large-Scale Biomedical Semantic Indexing and Question Answering Competition. *BMC Bioinformatics*, *16*(1), 138. https://doi.org/10.1186/s12859-015-0564-6

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ukasz Kaiser, Ł., & Polosukhin, I. (2017). Attention Is All You Need. *Advances in Neural Information Processing Systems*, *30*.

Vuokko, R., Vakkuri, A., & Palojoki, S. (2023). Systematized Nomenclature of Medicine–Clinical Terminology (SNOMED CT) Clinical Use Cases in the Context of Electronic Health Record Systems: Systematic Literature Review. *JMIR Medical Informatics*, *11*(1), e43750. https://doi.org/10.2196/43750

Wang, L., Ma, C., Feng, X., Zhang, Z., Yang, H., Zhang, J., Chen, Z., Tang, J., Chen, X., Lin, Y., Zhao, W. X., Wei, Z., & Wen, J.-R. (2023, agosto). A Survey on Large

Language Model based Autonomous Agents. https://doi.org/10.1007/s11704-024-40231-1

Warner, E., Lee, J., Hsu, W., Syeda-Mahmood, T., Kahn Jr., C. E., Gevaert, O., & Rao, A. (2024). Multimodal Machine Learning in Image-Based and Clinical Biomedicine: Survey and Prospects. *Int J Comput Vis*, *132*(9), 3753-3769. https://doi.org/10.1007/s11263-024-02032-8

Wei, J., Bosma, M., Zhao, V. Y., Guu, K., Yu, A. W., Lester, B., Du, N., Dai, A. M., & Le, Q. V. (2022, febrero). Finetuned Language Models Are Zero-Shot Learners [Comment: Version 5. Find list of changes in Appendix F (page 35)]. https://doi.org/10.48550/arXiv.2109.01652

Wooldridge, M. (2009). *An Introduction to MultiAgent Systems, 2nd Edition | Wiley*.

Wooldridge, M., & Jennings, N. R. (1995). Intelligent agents: Theory and practice. *The Knowledge Engineering Review*, *10*(2), 115-152. https://doi.org/10.1017/S0269888900008122

Wu, Q., Bansal, G., Zhang, J., Wu, Y., Li, B., Zhu, E., Jiang, L., Zhang, X., Zhang, S., Liu, J., Awadallah, A. H., White, R. W., Burger, D., & Wang, C. (2023, agosto). AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversation.

Xi, Z., Chen, W., Guo, X., He, W., Ding, Y., Hong, B., Zhang, M., Wang, J., Jin, S., Zhou, E., Zheng, R., Fan, X., Wang, X., Xiong, L., Zhou, Y., Wang, W., Jiang, C., Zou, Y., Liu, X., ... Gui, T. (2023, septiembre). The Rise and Potential of Large Language Model Based Agents: A Survey.

Xiao, H., Zhou, F., Liu, X., Liu, T., Li, Z., Liu, X., & Huang, X. (2025). A Comprehensive Survey of Large Language Models and Multimodal Large Language Models in Medicine. *Information Fusion*, *117*, 102888. https://doi.org/10.1016/j.inffus.2024.102888

Xie, X., Zhang, X., Tang, X., Zhao, J., Xiong, D., Ouyang, L., Yang, B., Zhou, H., Ling, B. W.-K., & Teo, K. L. (2025). MACTFusion: Lightweight Cross Transformer for Adaptive Multimodal Medical Image Fusion. *IEEE J Biomed Health Inform*, *29*(5), 3317-3328. https://doi.org/10.1109/JBHI.2024.3391620

Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., & Cao, Y. (2023). REACT: SYNERGIZING REASONING AND ACTING IN LANGUAGE MODELS. *11th International Conference on Learning Representations, ICLR 2023*.

Yin, S., Fu, C., Zhao, S., Li, K., Sun, X., Xu, T., & Chen, E. (2024). A Survey on Multimodal Large Language Models. *Natl Sci Rev*, *11*(12), nwae403. https://doi.org/10.1093/nsr/nwae403

Zhang, W., & Zhang, J. (2025). Hallucination Mitigation for Retrieval-Augmented Large Language Models: A Review. *Mathematics*, *13*(5), 856. https://doi.org/10.3390/math13050856

Zhang, Y., Sui, X., Pan, F., Yu, K., Li, K., Tian, S., Erdengasileng, A., Han, Q., Wang, W., Wang, J., Wang, J., Sun, D., Chung, H., Zhou, J., Zhou, E., Lee, B., Zhang, P., Qiu, X., Zhao, T., & Zhang, J. (2025). A comprehensive large-scale biomedical knowledge graph for AI-powered data-driven biomedical research. *Nat Mach Intell*, *7*(4), 602-614. https://doi.org/10.1038/s42256-025-01014-w

Zhao, X., Liu, S., Yang, S.-Y., & Miao, C. (2025). MedRAG: Enhancing Retrieval-augmented Generation with Knowledge Graph-Elicited Reasoning for Healthcare Copilot. *Proceedings of the ACM on Web Conference 2025*, 4442-4457. https://doi.org/10.1145/3696410.3714782

Zhou, H.-Y., Yu, Y., Wang, C., Zhang, S., Gao, Y., Pan, J., Shao, J., Lu, G., Zhang, K., & Li, W. (2023). A transformer-based representation-learning model with unified

processing of multimodal input for clinical diagnostics. *Nat Biomed Eng*, *7*(6), 743-755. https://doi.org/10.1038/s41551-023-01045-x

Zhuang, Y., Zhang, J., Lu, R., He, K., & Li, X. (2024). MedNER: Enhanced Named Entity Recognition in Medical Corpus via Optimized Balanced and Deep Active Learning. *ACM Trans. Intell. Syst. Technol.*, *15*(5), 1-24. https://doi.org/10.1145/3678178

Acrónimos

A continuación, se presentan los acrónimos utilizados en este documento. La lista fue elaborada por la autora a partir de los términos más relevantes del proyecto y su contexto.

- **AI** Artificial Intelligence (Inteligencia Artificial)
- **API** Application Programming Interface
- **BERT** Bidirectional Encoder Representations from Transformers
- **BioBERT** Biomedical BERT
- **BioNER** Biomedical Named Entity Recognition
- **BioNEs** Biomedical Named Entities
- **BioNLP** Biomedical Natural Language Processing
- **CLIP** Contrastive Language-Image Pre-training
- **CNN** Convolutional Neural Network
- **CSS** Cascading Style Sheets
- **CUDA** Compute Unified Device Architecture
- **CUI** Concept Unique Identifier
- **CV** Computer Vision (Visión por Computadora)
- **DICOM** Digital Imaging and Communications in Medicine
- **DOI** Digital Object Identifier
- **EHR** Electronic Health Record
- **F1** F1 Score (Medida F1)
- **GPU** Graphics Processing Unit
- **GPT** Generative Pre-trained Transformer
- **GradCAM** Gradient-weighted Class Activation Mapping
- **GSC** Gold-Standard Corpora
- **HKG** Healthcare Knowledge Graphs (Grafos de Conocimiento Sanitario)
- **HQS-VQA** Hierarchical Question Segregation based Visual Question Answering
- **HTML** HyperText Markup Language
- **IA** Inteligencia Artificial
- **ICD** International Classification of Diseases (Clasificación Internacional de Enfermedades)
- **IDF** Inverse Document Frequency (Frecuencia Inversa de Documentos)
- **JSON** JavaScript Object Notation
- **JSONL** JavaScript Object Notation Lines

- **KG** Knowledge Graph
- **LIME** Local Interpretable Model-agnostic Explanations
- **LLaVA** Large Language and Vision Assistant
- **LLM** Large Language Model
- **LoRA** Low-Rank Adaptation
- **LSTM** Long Short-Term Memory
- **LVLM** Large Vision-Language Model
- **maF1** Macro-averaged F1
- **MeSH** Medical Subject Headings
- **MIC** Medical Image Captioning
- **ML** Machine Learning
- **MLP** Multi-Layer Perceptron (Perceptrón Multicapa)
- **MRR** Mean Reciprocal Rank (Rango Recíproco Medio)
- **NCBI** National Center for Biotechnology Information
- **NEN** Named Entity Normalization
- **NER** Named Entity Recognition
- **NIH** National Institutes of Health
- **NLP** Natural Language Processing
- **OOV** Out-of-Vocabulary
- **PDF** Portable Document Format
- **PIL** Python Imaging Library
- **PLN** Procesamiento de Lenguaje Natural
- **PMID** PubMed Identifier
- **Q&A** Question and Answer
- **QA** Question Answering
- **RAG** Retrieval-Augmented Generation
- **RGPD** Reglamento General de Protección de Datos
- **ROCO** Radiology Objects in Context
- **ROUGE** Recall-Oriented Understudy for Gisting Evaluation
- **RxNorm** Prescription Drug Nomenclature
- **SAM** Segment Anything Model
- **SHAP** SHapley Additive exPlanations

- **SNOMED CT** Systematized Nomenclature of Medicine - Clinical Terms
- **SVM** Support Vector Machine (Máquina de Vectores de Soporte)
- **UI** User Interface
- **UMLS** Unified Medical Language System
- **URL** Uniform Resource Locator
- **VB-MVQA** Vision-Bilinear Multi-level Visual Question Answering
- **ViT** Vision Transformer
- **VQA** Visual Question Answering (Preguntas y Respuestas Visuales)
- **VQA-Med** Visual Question Answering Medical
- **VRAM** Video Random Access Memory
- **XAI** Explainable Artificial Intelligence (Inteligencia Artificial Explicable)
- **XML** eXtensible Markup Language

Anexos

A. Recursos en línea

El producto de este trabajo está disponible en Github para su evaluación.

- Repositorio de GitHub
- Modelo fine-tunado LLaVA-LLaMA 3 8B
- Modelo fine-tunado LLaVA-Mistral 7B (fallback)

B. Ejemplo de pregunta de tipo lista

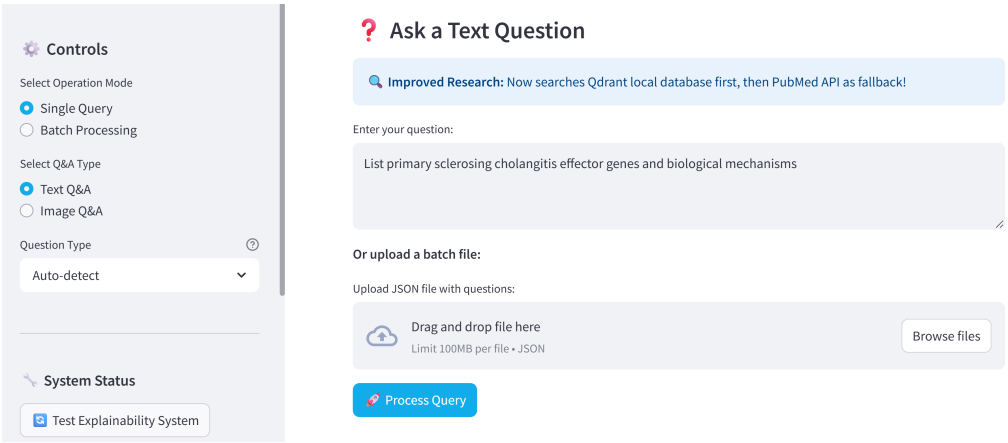


Figura 20: Paso 1 de flujo textual (Pregunta tipo "Lista"): Se introduce la pregunta.

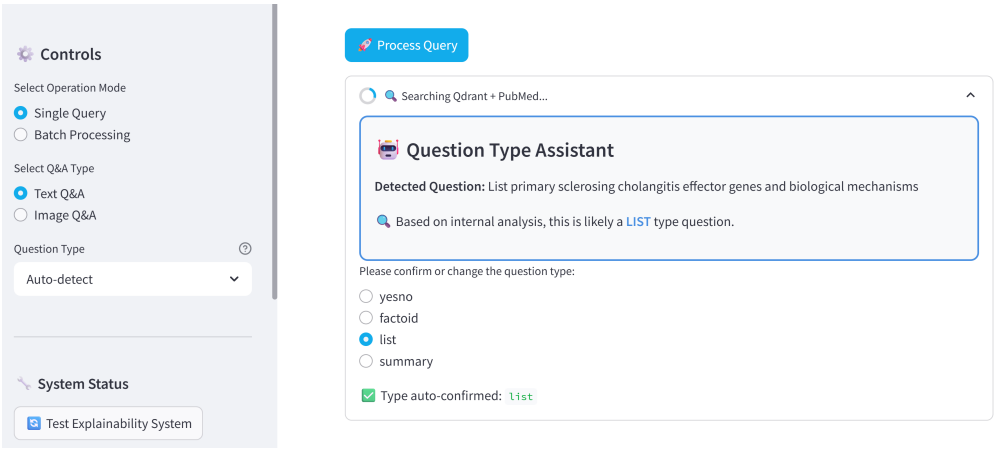


Figura 21: Paso 2 de flujo textual (Pregunta tipo "Lista"): Como no se seleccionó el tipo de pregunta aparece el chatbot para decir que según su análisis es una pregunta de tipo lista y solicitar confirmación.

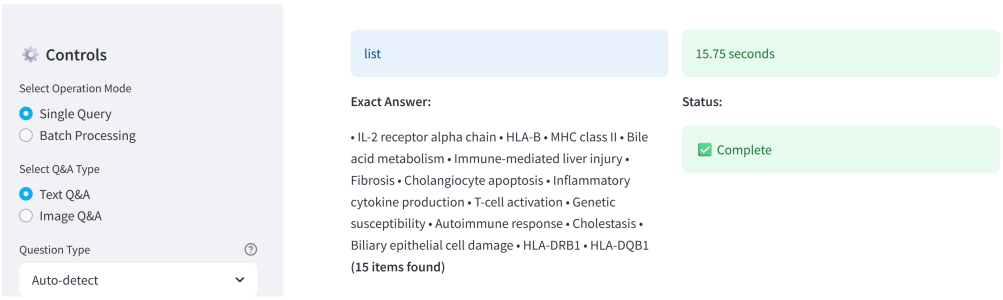


Figura 22: Paso 3 de flujo textual (Pregunta tipo "Lista"): Respuesta exacta (Exact Answer) con el listado de elementos identificados en respuesta a la pregunta.

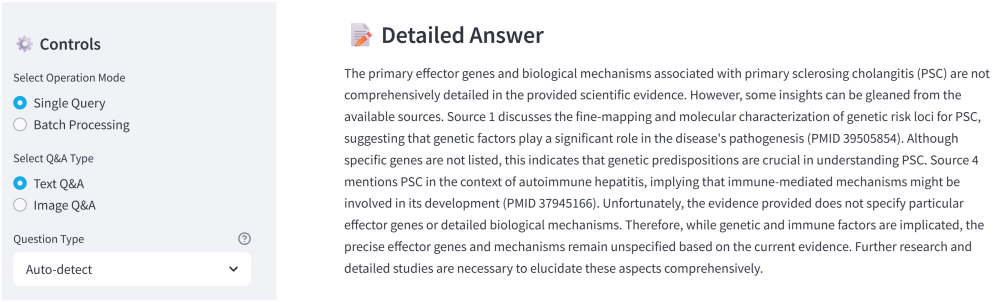


Figura 23: Paso 4 de flujo textual (Pregunta tipo "Lista"): Respuesta ideal (Detailed Answer) con un resumen en respuesta a la pregunta y con los PMID en los que se basa la respuesta.

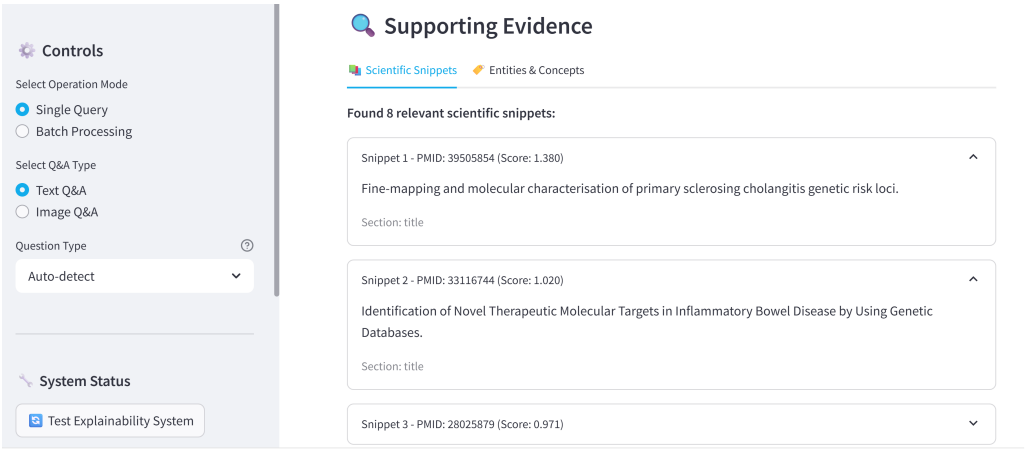


Figura 24: Paso 5 de flujo textual (Pregunta tipo "Lista"): Snippets identificados con el PMID del artículo en el que se basa la respuesta y su puntuación.

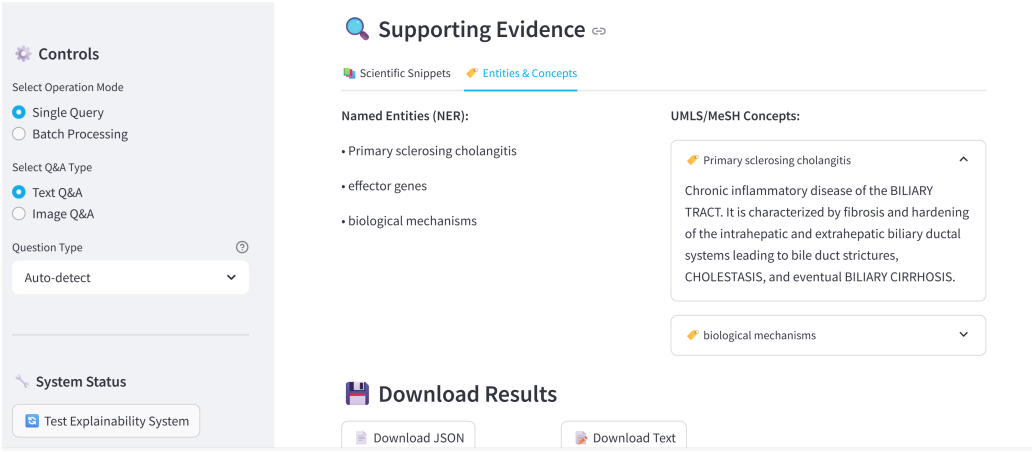


Figura 25: Paso 6 de flujo textual (Pregunta tipo "Lista"): NER biomédicas y términos UMLS/MeSH identificados por el sistema.

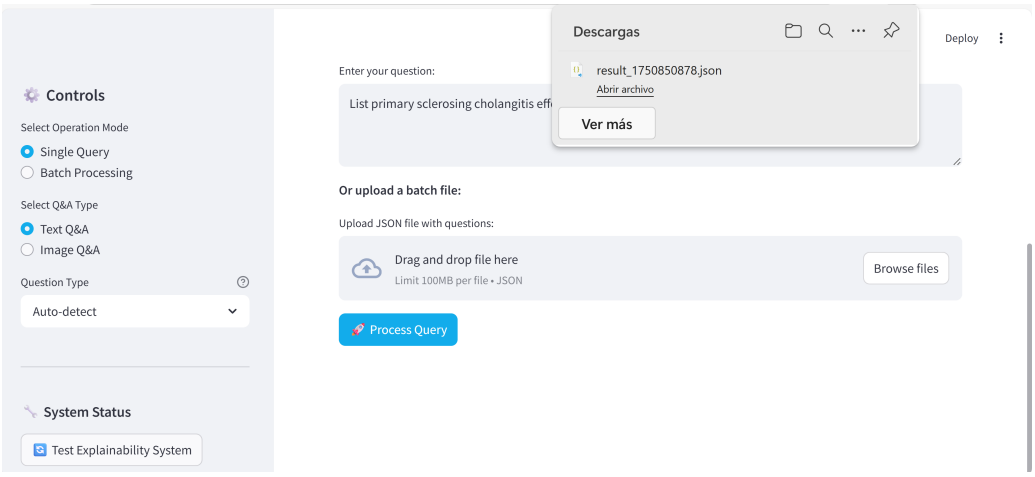
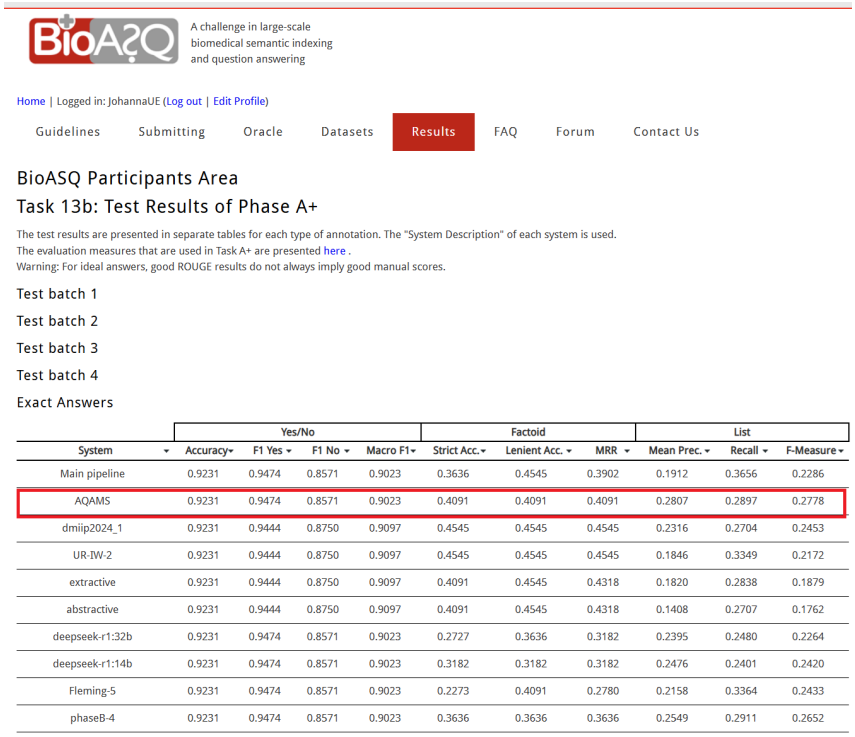


Figura 26: Paso 7 de flujo textual (Pregunta tipo "Lista"): Descarga del archivo .JSON con los metadatos.

C. Resultados preliminares de bioASQ Task 13b 2025



A challenge in large-scale biomedical semantic indexing and question answering

Home | Logged in: JohannaUE (Log out | Edit Profile)

Guidelines | Submitting | Oracle | Datasets | **Results** | FAQ | Forum | Contact Us

BioASQ Participants Area

Task 13b: Test Results of Phase A+

The test results are presented in separate tables for each type of annotation. The "System Description" of each system is used.
The evaluation measures that are used in Task A+ are presented here .
Warning: For ideal answers, good ROUGE results do not always imply good manual scores.

Test batch 1

Test batch 2

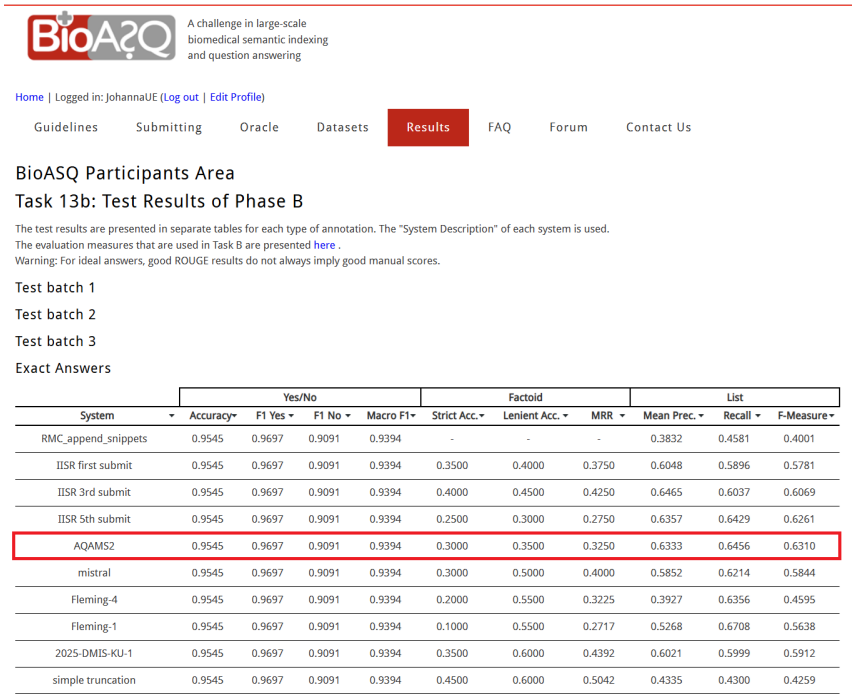
Test batch 3

Test batch 4

Exact Answers

	Yes/No				Factoid			List		
System	Accuracy	F1 Yes	F1 No	Macro F1	Strict Acc.	Lenient Acc.	MRR	Mean Prec.	Recall	F-Measure
Main pipeline	0.9231	0.9474	0.8571	0.9023	0.3636	0.4545	0.3902	0.1912	0.3656	0.2286
AQAMS	0.9231	0.9474	0.8571	0.9023	0.4091	0.4091	0.4091	0.2807	0.2897	0.2778
dmiip2024_1	0.9231	0.9444	0.8750	0.9097	0.4545	0.4545	0.4545	0.2316	0.2704	0.2453
UR-IW-2	0.9231	0.9444	0.8750	0.9097	0.4545	0.4545	0.4545	0.1846	0.3349	0.2172
extractive	0.9231	0.9444	0.8750	0.9097	0.4091	0.4545	0.4318	0.1820	0.2838	0.1879
abstractive	0.9231	0.9444	0.8750	0.9097	0.4091	0.4545	0.4318	0.1408	0.2707	0.1762
deepseek-r1:32b	0.9231	0.9474	0.8571	0.9023	0.2727	0.3636	0.3182	0.2395	0.2480	0.2264
deepseek-r1:14b	0.9231	0.9474	0.8571	0.9023	0.3182	0.3182	0.3182	0.2476	0.2401	0.2420
Fleming-5	0.9231	0.9474	0.8571	0.9023	0.2273	0.4091	0.2780	0.2158	0.3364	0.2433
phaseB-4	0.9231	0.9474	0.8571	0.9023	0.3636	0.3636	0.3636	0.2549	0.2911	0.2652

Figura 27: Resultados preliminares de Fase A+ (Batch 4): El sistema participante obtuvo buenos resultados en este batch.



A challenge in large-scale biomedical semantic indexing and question answering

Home | Logged in: JohannaUE (Log out | Edit Profile)

Guidelines | Submitting | Oracle | Datasets | **Results** | FAQ | Forum | Contact Us

BioASQ Participants Area

Task 13b: Test Results of Phase B

The test results are presented in separate tables for each type of annotation. The "System Description" of each system is used.
The evaluation measures that are used in Task B are presented here .
Warning: For ideal answers, good ROUGE results do not always imply good manual scores.

Test batch 1

Test batch 2

Test batch 3

Exact Answers

	Yes/No				Factoid			List		
System	Accuracy	F1 Yes	F1 No	Macro F1	Strict Acc.	Lenient Acc.	MRR	Mean Prec.	Recall	F-Measure
RMC_append_snippets	0.9545	0.9697	0.9091	0.9394	-	-	-	0.3832	0.4581	0.4001
IISR first submit	0.9545	0.9697	0.9091	0.9394	0.3500	0.4000	0.3750	0.6048	0.5896	0.5781
IISR 3rd submit	0.9545	0.9697	0.9091	0.9394	0.4000	0.4500	0.4250	0.6465	0.6037	0.6069
IISR 5th submit	0.9545	0.9697	0.9091	0.9394	0.2500	0.3000	0.2750	0.6357	0.6429	0.6261
AQAMS2	0.9545	0.9697	0.9091	0.9394	0.3000	0.3500	0.3250	0.6333	0.6456	0.6310
mistral	0.9545	0.9697	0.9091	0.9394	0.3000	0.5000	0.4000	0.5852	0.6214	0.5844
Fleming-4	0.9545	0.9697	0.9091	0.9394	0.2000	0.5500	0.3225	0.3927	0.6356	0.4595
Fleming-1	0.9545	0.9697	0.9091	0.9394	0.1000	0.5500	0.2717	0.5268	0.6708	0.5638
2025-DMIS-KU-1	0.9545	0.9697	0.9091	0.9394	0.3500	0.6000	0.4392	0.6021	0.5999	0.5912
simple truncation	0.9545	0.9697	0.9091	0.9394	0.4500	0.6000	0.5042	0.4335	0.4300	0.4259

Figura 28: Resultados preliminares de Task B (Batch 3): Puntaje alto en preguntas de tipo List.

D. Resultados de ImageCLEF medical 2025

Results

The tables below contain only the best runs of each owner on ai4mediabench, for a complete list of all runs please see the Google Sheets files for Concept Detection and for Caption Prediction.

Concept Detection

ID	Owner	Submission Name	F1	F1 secondary
1980	AUEB NLP Group	ensemble_dual_thr_3_5monte_eff_eff_.zip	0.5888	0.9484
1725	DeepLens	submission	0.5766	0.9299
1505	mapan	submission	0.5660	0.9298
1892	UIT-Oggy	submission	0.5613	0.9104
1508	DS4DH	submission.csv	0.5225	0.8672
1774	sakthiii	submission	0.4003	0.9082
1903	JJ-VMed	submission	0.3982	0.8329
1807	UMUTeam	submission_with_unknown_clean	0.2398	0.5377
1942	LekshmiscopeVIT	submission.csv	0.1494	0.2298

Caption Prediction

ID	owner	Submission Name	Overall	Similarity	BERTScore (Recall)	ROUGE-1	BLEURT	Relevance Average	UMLS Concept F1	AlignScore	Factuality Average
1681	UMUTeam	submission.zip	0.3432	0.9271	0.5977	0.2594	0.3230	0.5268	0.1816	0.1375	0.1596
1520	DS4DH	submission.csv.zip	0.3362	0.9016	0.6067	0.2516	0.3096	0.5174	0.1682	0.1417	0.1549
1900	AI Stat Lab	submission.zip	0.3229	0.8919	0.5823	0.2440	0.3173	0.5089	0.1524	0.1213	0.1369
1914	UIT-Oggy	submission_ep2_cleaned.zip	0.3211	0.8798	0.5951	0.2535	0.3020	0.5076	0.1672	0.1021	0.1346
1403	AUEB NLP Group	2-instruct-blip-ft.zip	0.3068	0.7947	0.5884	0.2176	0.3030	0.4759	0.1429	0.1325	0.1377
1896	JJ-VMed	submission.zip	0.3043	0.8251	0.5953	0.2389	0.3094	0.4922	0.1366	0.0964	0.1165
1890	sakthiii	submission	0.2746	0.7957	0.5553	0.1607	0.2806	0.4481	0.1094	0.0928	0.1011
1815	csmorgan	Qwen_2B_Submission_1.zip	0.2315	0.5704	0.5180	0.1598	0.2385	0.3717	0.0741	0.1087	0.0914

Explainability Task - Human Evaluation Results

Team	Caption readability	Clinical appropriateness of caption	Caption level of detail	Caption focus	Mean caption rating	Visual-text coherence	Completeness of visualization	Visualization focus	Mean visualization rating	Appropriateness of Methodology	Overall
AUEB NLP Group	4.5	2.7	2.6	3.3	3.3	3.1	2.8	2.6	2.8	4.0	3.2
JJ-VMed	3.4	2.4	2.8	4.1	3.2	1.9	1.9	1.9	1.9	2.0	2.6

Figura 29: Resultados del sistema participante en ImageCLEF medical Caption 2025.