

Descripción del estado de salud de cultivos mediante recuperación multimodal de información

Karen D. Ramon-Cristancho¹, Germán Ríos-Toledo¹, J. A. de Jesús Osuna-Coutiño^{1,*},
Abiel Aguilar-González², Fernando Pech-May³, Madaín Pérez-Patricio¹,
Juan Pablo Francisco Posadas-Durán⁴, Nestor A. Morales-Navarro¹

¹ Tecnológico Nacional de México/IT de Tuxtla Gutiérrez, Department of Science,
Tuxtla Gutiérrez, Chiapas, México

² Instituto Nacional de Astrofísica, Óptica y Electrónica,
Department of Computer Science, Cholula, México

³ Instituto Tecnológico Superior de los Ríos,
Department of Computer Science, Tabasco,
México

⁴ Instituto Politécnico Nacional (IPN),
Escuela Superior de Ingeniería Mecánica y Eléctrica (ESIME),
Unidad Zacatenco, Ciudad de México,
México

m24271381@tuxtla.tecnm.mx, osuna_antonio_77@hotmail.com, german.rt@tuxtla.tecnm.mx,
abel@inaoep.mx, fernando.pech@cinvestav.mx, madain.pp@tuxtla.tecnm.mx,
jposadasd@ipn.mx, nestor.mn@tuxtla.tecnm.mx

Resumen. En el contexto del procesamiento de lenguaje natural aplicado al ámbito agrícola, la descripción precisa e interpretable del estado de salud de los cultivos resulta fundamental para apoyar la toma de decisiones en el campo. Una alternativa para abordar la descripción de cultivos es el uso de modelos de lenguaje grandes (LLMs). En general, estos modelos pueden dar descripciones de múltiples áreas. Sin embargo, cuando se emplean sin acceso a conocimiento agrícola especializado ni a fuentes externas de información, su desempeño tiende a ser limitado. En consecuencia, las descripciones suelen ser genéricas, omiten detalles relevantes de los síntomas e en algunos casos, generan diagnósticos incorrectos o alucinaciones. Para mitigar estas limitaciones, los modelos LLMs suelen integrarse con Generación Aumentada por Recuperación (RAG), de modo que las respuestas se fundamenten en conocimiento agrícola externo y actualizado. Uno de los principales retos actuales consiste en diseñar estrategias de recuperación y selección de contexto en las que tanto la imagen como la pregunta funcionen como

señales para recuperar evidencia relevante. Motivados por lo anterior, proponemos un sistema basado en un modelo de lenguaje grande para describir el estado de salud de cultivos mediante una recuperación híbrida multimodal. Para eso, nuestra metodología consta de cinco etapas. Primero, la definición de las entradas del sistema. Segundo, el procesamiento del conocimiento externo mediante BGE y CLIP. Tercero, el procesamiento de la consulta del usuario. Cuarto, la recuperación híbrida mediante similitud texto–texto e imagen–texto y fusión con *reciprocal rank fusion* (RRF). Finalmente, la generación de la respuesta con un LLM multimodal. Los resultados experimentales muestran que la estrategia propuesta logra hasta un 2.32% de mejora relativa en F1-score respecto al modelo base, lo cual respalda su efectividad para mejorar la calidad de las respuestas generadas en el dominio agrícola.

Palabras clave. Mapeo detallado de inundaciones, mapeo de inundaciones con Sentinel-1, detección de inundaciones con U-Net.

Description of Crop Health Status Using Multimodal Information Retrieval

Abstract. In the context of natural language processing applied to agriculture, a precise and interpretable description of crop health is crucial for supporting field-level decision-making. Large Language Models (LLMs) offer a potential approach to describing crops; generally, these models are capable of providing descriptions across a wide range of domains. However, when used without access to specialized agricultural knowledge or external information sources, their performance tends to be limited. Consequently, the resulting descriptions are often generic, omit relevant details regarding symptoms, and—in some instances—lead to incorrect diagnoses or hallucinations. To mitigate these limitations, LLMs are frequently integrated with Retrieval-Augmented Generation (RAG) to ground their responses in up-to-date, external agricultural knowledge. A key current challenge lies in designing context retrieval and selection strategies where both the image and the query serve as signals to retrieve relevant evidence. Motivated by this, we propose a system based on a large language model to describe crop health using multimodal hybrid retrieval. Our methodology comprises five stages: first, defining the system inputs; second, processing external knowledge using BGE and CLIP; third, processing the user query; fourth, performing hybrid retrieval based on text-text and image-text similarity, combined via Reciprocal Rank Fusion (RRF); and finally, generating the response using a multimodal LLM. Experimental results demonstrate that the proposed strategy achieves a relative improvement of up to 2.32% in F1-score compared to the baseline model, validating its effectiveness in enhancing the quality of generated responses within the agricultural domain.

Keywords. Deep Flood mapping, Sentinel-1 to flood mapping, Flood detection with U-Net.

1. Introducción

La descripción del estado de salud de los cultivos se basa en la observación de síntomas visibles en las plantas para determinar si están sanas o afectadas por algún tipo de estrés [24]. Esto brinda a los agricultores una herramienta valiosa para la toma de decisiones sobre el manejo de sus plantas. Al identificar a tiempo signos de enfermedades, plagas o estrés ambiental, es

posible actuar con rapidez, reducir pérdidas y mejorar tanto la productividad como la calidad de la cosecha [46].

Esta práctica también promueve un uso más eficiente de insumos agrícolas, fortalece los sistemas productivos frente a amenazas externas y contribuye a un manejo sostenible y responsable con el ambiente [17]. Estos beneficios no solo representan una mejora teórica en el manejo agrícola, sino que ya se han aplicado en diversos contextos rurales para optimizar recursos [41], predicción de enfermedades [54] y fortalecer la sostenibilidad de los cultivos [38].

Tradicionalmente, la descripción del estado de salud de los cultivos se realiza mediante la observación visual directa [8, 51]. Este enfoque ha sido ampliamente utilizado por agricultores durante décadas debido a sus ventajas prácticas, ya que permite detectar tempranamente síntomas visibles como manchas, decoloraciones o deformaciones, y no requiere equipamiento costoso ni infraestructura tecnológica y puede aplicarse de forma inmediata en cualquier entorno agrícola [6].

Además, permite una respuesta rápida ante la detección de problemas, lo que resulta útil en contextos rurales con recursos limitados. Sin embargo, esta práctica también presenta limitaciones importantes. La precisión del diagnóstico depende en gran medida del criterio, la experiencia y el nivel de capacitación del observador. Esto puede generar errores de interpretación, especialmente en fases tempranas de las enfermedades, donde los síntomas son sutiles o se confunden con otros factores como deficiencias nutricionales [5, 11].

Posteriormente, se incorporaron metodologías basadas en análisis visuales, sensores ópticos y procesamiento clásico de imágenes [8]. Entre estas técnicas están el uso de imágenes hiperespectrales (HSI) [6, 34], que capturan información en múltiples bandas del espectro electromagnético para detectar alteraciones fisiológicas no visibles al ojo humano; el análisis de imágenes asistido por computadora (CBIA) [28], que emplea algoritmos para procesar y segmentar imágenes con el fin de identificar patrones asociados a plagas o enfermedades; los Diagramas de Área Estándar (SADs) [13], que cuantifican características visuales como forma, textura y color de los síntomas presentes

en las plantas; y la teledetección hiperespectral (HRS) [47], que permiten registrar imágenes con gran nivel de detalle espacial para identificar anomalías en hojas, tallos o frutos. Sin embargo, estas metodologías presentan problemas. Suelen funcionar bien solo en ambientes controlados y se suelen enfocar a un solo tipo de cultivos [8, 7]. Esto hace que sea complicado aplicarlas en el campo, sobre todo en zonas rurales donde no hay mucha tecnología ni personal capacitado.

Actualmente, una de las alternativas más relevantes para la descripción del estado de salud de los cultivos es el uso de técnicas de visión por computadora basadas en modelos de Aprendizaje Profundo (deep learning) [15]. Las redes neuronales convolucionales (CNNs) han sido el enfoque predominante en el análisis de imágenes, debido a que emplean operaciones de convolución para extraer características visuales relevantes. En agricultura, estas redes se utilizan para analizar imágenes de plantas y detectar automáticamente síntomas de enfermedades, plagas o estrés fisiológico, aproximándose al razonamiento visual de un experto. Este enfoque ha demostrado resultados prometedores en la clasificación de enfermedades foliares, segmentación de áreas afectadas y reconocimiento de patrones visuales complejos en diversos cultivos [45, 48]. Sin embargo, su implementación en escenarios reales aún presenta desafíos importantes, como la necesidad de grandes volúmenes de imágenes etiquetadas para entrenar los modelos, la sensibilidad a variaciones ambientales como cambios de iluminación o ángulos de captura, y la baja capacidad de generalización a nuevas enfermedades o especies no vistas durante el entrenamiento. Además, en muchos enfoques basados en CNNs la salida suele limitarse a la predicción de una etiqueta de clase, sin proporcionar una explicación textual de los síntomas observados, lo que reduce su interpretabilidad para usuarios no especializados y limita su utilidad práctica en el campo agrícola [15, 58].

En paralelo a los avances en visión por computadora, han surgido los modelos de lenguaje grande (LLMs), con alto potencial en procesamiento de lenguaje natural y tareas multimodales [53]. Estos modelos se basan en la arquitectura Transformer

[49], la cual permite capturar relaciones entre elementos de una secuencia sin depender de su posición. En consecuencia, los Transformers pueden emplearse en tareas discriminativas, como la clasificación, y en tareas generativas, debido a su capacidad para comprender y generar lenguaje [14, 37]. Estos modelos pueden adaptarse a tareas específicas mediante una técnica conocida como ajuste fino [1], que consiste en entrenar un modelo previamente generalizado con ejemplos concretos de un dominio, como el agrícola. Así, el modelo se adapta al contexto, generando respuestas precisas y alineadas con el conocimiento experto [43, 2, 57]. Sin embargo, los modelos de lenguaje grande (LLMs) presentan requerimientos aún más exigentes que otros enfoques de deep learning, ya que necesitan volúmenes de datos anotados significativamente mayores y una infraestructura computacional mucho más robusta para su entrenamiento y ajuste. Estas exigencias incluyen el uso de GPUs especializadas, almacenamiento de gran capacidad y tiempos de entrenamiento prolongados [29, 19].

Además, presentan la limitación de que el conocimiento incorporado durante el ajuste fino no se actualiza de manera dinámica, por lo que es necesario reentrenar el modelo cada vez que surge nueva información [52, 25], lo que impacta en la viabilidad práctica de su implementación en entornos agrícolas reales, especialmente en contextos con limitaciones tecnológicas.

En trabajos recientes se ha explorado la actualización dinámica del conocimiento en LLMs mediante la generación aumentada por recuperación (RAG), un enfoque que integra información externa proveniente de bases de conocimiento o documentos especializados [25, 21]. Esta alternativa permite especializar al modelo sin necesidad de reentrenamiento completo. Esto evita aplicar ajuste fino, lo que reduce de manera considerable los requerimientos computacionales y al mismo tiempo, favorece que los modelos generales se adapten mejor a un dominio particular como el agrícola. RAG permite acceder a conocimiento actualizado y específico, ampliando la precisión de las respuestas y disminuyendo el riesgo de alucinaciones. Sin embargo, una limitación clave es que RAG se ha diseñado

principalmente para trabajar con texto, sin incorporar información visual. Además, su rendimiento y confiabilidad suelen depender de la calidad y organización del conocimiento disponible, así como de la eficacia del retriever al seleccionar el contexto más relevante; en aplicaciones sensibles puede requerirse verificación humana para reducir errores y garantizar respuestas más consistentes [4, 44].

Motivados por los resultados de los enfoques existentes, en este trabajo se propone un sistema RAG multimodal que integra tanto la imagen del cultivo como la pregunta del usuario para recuperar información contextual especializada. A diferencia de otros trabajos en agricultura donde la imagen se utiliza únicamente para clasificación o detección de enfermedades, aquí la imagen también cumple un papel clave en la etapa de recuperación. Específicamente, se aprovechan las características visuales para ayudar a encontrar documentos relevantes que complementen la pregunta del usuario, facilitando así una recuperación más precisa desde bases de datos vectoriales y/o relacionales. Para lograrlo, se emplean dos mecanismos de recuperación: uno basado en texto mediante embeddings BGE y otro guiado por la imagen utilizando CLIP para proyectar la imagen a un espacio multimodal donde puede compararse con texto mediante similitud, lo cual permite capturar tanto la intención lingüística del usuario como las evidencias visuales presentes en la planta. Ambas fuentes se combinan mediante la técnica de fusión recíproca rank fusion (RRF), que pondera la relevancia de los fragmentos desde ambas modalidades y selecciona los más útiles para la generación de respuesta. Este enfoque multimodal permite mejorar la precisión del sistema RAG, al reducir la dependencia exclusiva del texto, lo que representa un aporte diferencial frente a las soluciones RAG tradicionales aplicadas al ámbito agrícola.

2. Trabajos relacionados

Los sistemas RAG combinan técnicas de recuperación de información con modelos generativos de lenguaje, permitiendo generar respuestas

fundamentadas en documentos relevantes recuperados en tiempo real [25]. Esta capacidad ha demostrado ser especialmente valiosa frente a los desafíos del ajuste fino, ofreciendo una alternativa más eficiente y dinámica para actualizar el conocimiento en los modelos. Estudios recientes muestran beneficios concretos de RAG en agricultura. Microsoft [4] creó una plataforma que extrae conocimiento de documentos de Estados Unidos, Brasil e India y genera preguntas y respuestas por región; así logra recomendaciones contextualizadas y evita reentrenamientos costosos manteniendo la información actualizada. De forma similar, AgroLLM [44] y ShizishanGPT [55] integran fuentes agronómicas diversas para responder consultas locales con mayor precisión. En conjunto, estas experiencias sugieren que RAG mejora la exactitud y la pertinencia de las respuestas en escenarios agrícolas dinámicos. Sin embargo, la mayoría de estos sistemas aplicados en el contexto agrícola siguen limitándose al uso exclusivo de texto, sin aprovechar información visual como imágenes de cultivos, enfermedades o condiciones del suelo. Esta limitación marca una importante línea de investigación abierta: la integración de RAG con tareas de VQA. Al combinar imágenes con conocimiento recuperado desde bases especializadas, estos sistemas podrían ofrecer respuestas más completas y precisas, especialmente en escenarios donde los datos visuales son clave para el diagnóstico y la toma de decisiones.

Recientes investigaciones han comenzado a combinar RAG con tareas de VQA, al reconocer que el contenido visual por sí solo no siempre permite generar respuestas correctas, especialmente cuando se requiere conocimiento funcional o contextual no visible en la imagen. En estos casos, la recuperación de documentos especializados permite complementar la información visual y mejorar el razonamiento multimodal más allá de lo que ofrecen los píxeles. Modelos como RA-VQA [26] muestran que añadir recuperación de conocimiento externo aumenta la precisión al enlazar la imagen con textos o descripciones enciclopédicas. De forma complementaria, ConceptBERT [16] y KAT [18] conectan imagen y lenguaje mediante conceptos y representaciones semánticas, lo que

permite inferir relaciones no visibles de forma directa, como funciones, usos y conocimientos de sentido común. Estas estrategias amplían el rango de preguntas que pueden responderse cuando la evidencia no está explícita en los píxeles. Trabajos como REVEAL [20] amplían este enfoque al usar la imagen y la pregunta para recuperar textos externos y luego fusionarlos con la entrada visual mediante atención cruzada, logrando respuestas mejor contextualizadas. Asimismo, SB-EVQA [9] clasifica arquitecturas multimodales y muestra que la alineación explícita entre visión y texto suele tener mejor rendimiento en preguntas complejas. Sin embargo, la mayoría de estos trabajos se enfocan en responder preguntas que requieren conocimientos generales, como cultura popular, historia o ciencias naturales. Aunque efectivamente utilizan RAG para complementar información visual con conocimiento externo, en la mayoría de los casos estas aplicaciones tienden a centrarse en razonamiento abstracto o contextual amplio, sin especializarse en dominios técnicos específicos como la agricultura [27, 39, 10].

Por otro lado, *reciprocal rank fusion* (RRF) [12] es una técnica de agregación de rankings utilizada para combinar resultados provenientes de múltiples recuperadores. En RAG, se ha empleado para fusionar evidencia recuperada a partir de reformulaciones de la consulta [35], mientras que en escenarios multimodales se ha explorado como mecanismo para integrar señales visuales y textuales en la recuperación. Sin embargo, su aplicación para fusionar explícitamente recuperación textual y recuperación guiada por imagen en sistemas VQA+RAG especializados en diagnóstico agrícola sigue siendo limitada, lo que motiva su adopción en este trabajo.

En el dominio agrícola, [31] propuso un sistema híbrido para el diagnóstico de enfermedades del café que combina detección visual mediante YOLOv8, recuperación de información con RAG y generación de respuestas con LLMs. El sistema está diseñado para detectar enfermedades en hojas de café y generar respuestas detalladas en lenguaje natural utilizando modelos como Llama-3. Esta propuesta demuestra el potencial de los sistemas RAG multimodales en aplicaciones agrícolas reales, especialmente al abordar la generación de

respuestas explicativas y contextuales guiadas por información visual y textual.

Sin embargo, el componente de visión basado en YOLOv8 depende de entrenamiento supervisado para clases predefinidas; por tanto, incorporar nuevas enfermedades o cultivos requiere reentrenamiento y nuevas anotaciones, lo que limita su escalabilidad. Por otro lado, la evaluación del sistema se centra únicamente en el rendimiento de detección medido mediante la métrica *mean Average Precision* (mAP), sin analizar el impacto directo del módulo RAG en tareas VQA. Además, si no se detecta una enfermedad, el sistema no activa el pipeline RAG y el modelo LLM responde sin análisis guiado, lo que puede reducir la precisión o interpretabilidad.

Nosotros previamente hemos trabajado en ICQP (Identificación, Clasificación, Cuantificación y Predicción) de cultivos. En [42] y [33] propusimos un pipeline basado en una sola imagen RGB y reconstrucción 3D, donde primero se genera un mapa de profundidad con una red entrenada en geometría de la planta, luego se reconstruye la estructura tridimensional usando un modelo de cámara pinhole, y finalmente se clasifica el nivel de estrés mediante una red neuronal convolucional (CNN).

Este enfoque no invasivo permite detectar alteraciones morfológicas tempranas, lo que lo hace útil para el monitoreo en campo. Sin embargo, al estar limitado a etiquetas de clase derivadas exclusivamente de la imagen el sistema no puede responder preguntas abiertas ni ofrecer explicaciones detalladas, lo que restringe su utilidad en contextos donde la interpretación es clave para la toma de decisiones.

Por otro lado, en [50] presentamos un enfoque supervisado que combina un generador y un clasificador en bucle para crear y depurar imágenes sintéticas, con el objetivo de ampliar conjuntos de datos reducidos y mejorar modelos de clasificación en agricultura. Aunque este tipo de propuestas automatiza el entrenamiento y mejora la precisión visual, siguen centradas exclusivamente en la tarea de clasificación, sin abordar la integración de información textual ni la generación de respuestas interpretables para el usuario.

A diferencia de los trabajos mencionados, la metodología propuesta integra explícitamente la evidencia visual con recuperación de conocimiento externo especializado en agricultura para responder preguntas en lenguaje natural donde el sistema no solo observa la imagen. Para ello, proponemos un enfoque RAG multimodal con RRF que fusiona recuperación textual y visual para priorizar evidencia relevante y construir un contexto más completo y coherente. Este contexto es aprovechado por el modelo generativo, que produce respuestas claras, fundamentadas y adaptadas a la consulta del usuario.

Aún son escasos los estudios que apliquen VQA con RAG en contextos agrícolas, donde la interpretación visual requiere conocimientos técnicos específicos sobre plagas, enfermedades u otras variables agronómicas. Esta brecha evidencia la necesidad de sistemas multimodales orientados al diagnóstico visual agrícola con recuperación contextualizada, abriendo nuevas posibilidades para la toma de decisiones en el campo.

3. Metodología

En esta sección se describe el sistema RAG multimodal propuesto, el cual integra recuperación semántica textual y recuperación basada en información visual para mejorar la descripción del estado de salud de los cultivos. La estrategia combina dos mecanismos complementarios: la recuperación texto–texto mediante embeddings BGE y la recuperación imagen–texto utilizando CLIP, cuyos rankings se fusionan mediante *reciprocal rank fusion* (RRF) antes de la generación de la respuesta con un LLM multimodal.

La metodología se organiza en cinco etapas principales: (1) Entradas del sistema, (2) Procesamiento de conocimiento externo, (3) Procesamiento de la consulta, (4) Recuperación híbrida y (5) Respuesta. Cada una de estas etapas se detalla en las subsecciones siguientes. La Figura 1 muestra el diagrama general del sistema propuesto.

3.1. Entradas del sistema

En esta sección se describen los elementos que constituyen las entradas del sistema RAG multimodal. En particular, se define la base documental externa utilizada como fuente de conocimiento, así como el par multimodal (q, x) compuesto por la pregunta textual y la imagen proporcionadas por el usuario. Estos componentes representan el punto de partida del sistema y determinan la información disponible para las etapas posteriores de procesamiento, recuperación y generación.

3.1.1. Documentos externos

Las fuentes de conocimiento utilizadas para la construcción del sistema RAG corresponden al conjunto de preguntas y respuestas del dataset PlantVillageVQA, específicamente del Nivel 3 (*razonamiento e inferencia de orden superior*) [32], que agrupa categorías orientadas a razonamiento causal, enfermedad específica, descripción integral y análisis contrafactual. Con el fin de evitar fuga de información y garantizar una evaluación justa, se realizó un proceso de filtrado para asegurar que ninguna pregunta empleada como fuente de conocimiento coincidiera con las utilizadas en la etapa de evaluación experimental. De esta manera, la base documental externa se construyó exclusivamente a partir de pares pregunta–respuesta distintos a los evaluados, evitando solapamientos directos entre recuperación y prueba.

3.1.2. Pregunta e imagen

La entrada del sistema está compuesta por el par multimodal (q, x) , donde q representa la consulta textual que realiza el usuario acerca de la imagen de entrada y expresa qué información desea obtener, guiando el proceso posterior del sistema. Por su parte, la imagen x corresponde a la evidencia visual asociada a cada pregunta y representa el cultivo o parte de la planta observada en campo, conteniendo patrones, texturas y variaciones de color relevantes para caracterizar su estado de salud. En este trabajo, tanto las preguntas q como las imágenes x provienen del dataset PlantVillageVQA (Nivel 3), de modo

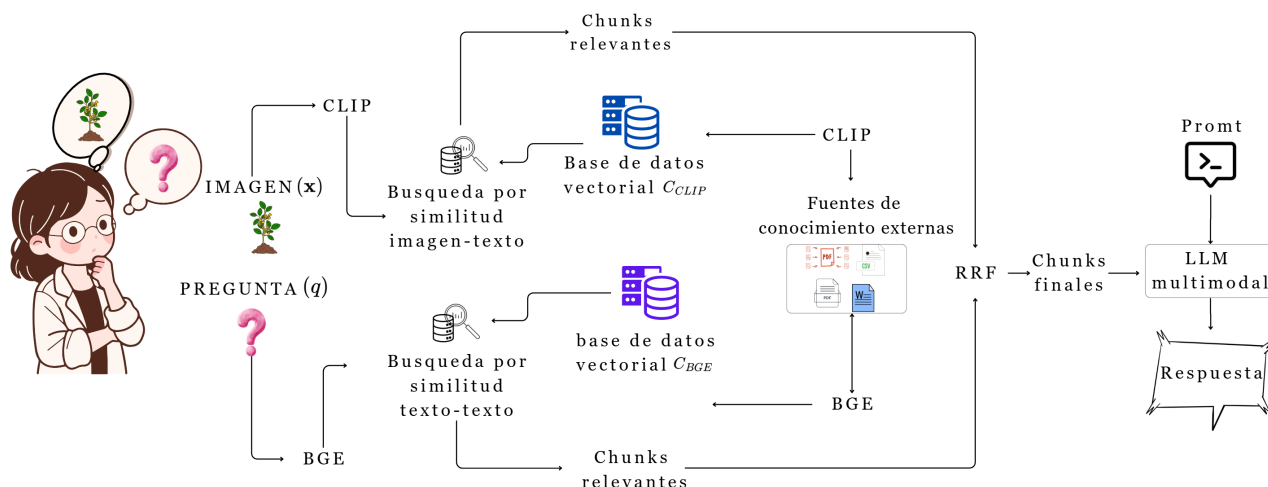


Fig. 1. Diagrama general del sistema RAG multimodal con recuperación híbrida (BGE/CLIP) y fusión por RRF

que cada consulta está asociada a su respectiva imagen y refleja escenarios de diagnóstico y razonamiento propios del dominio agrícola.

3.2. Procesamiento de conocimiento externo

Con el fin de habilitar la recuperación texto–texto e imagen–texto propuesta en este trabajo, la base documental externa debe transformarse en representaciones vectoriales compatibles con los modelos de embeddings empleados. En esta sección se describe el proceso de segmentación del corpus, la generación de embeddings con distintos encoders y la construcción de colecciones vectoriales que servirán como soporte para la etapa de recuperación híbrida.

3.2.1. Codificador de texto BGE (Documentos externos)

Los documentos de la fuente de conocimiento externa se procesan mediante LangChain [30, 22]. El contenido se divide en fragmentos de texto o *chunks* utilizando *RecursiveCharacterTextSplitter* [23]. Cada fragmento, denotado como K_i , tiene aproximadamente 500 caracteres y mantiene una superposición de 100 caracteres con el fragmento siguiente. Esta configuración permite conservar

continuidad semántica entre fragmentos consecutivos sin generar unidades de recuperación excesivamente extensas.

Posteriormente, cada chunk K_i se codifica mediante el modelo BGE (BAAI General Embedding) [3]. BGE utiliza una arquitectura basada en Transformer para obtener una representación vectorial que captura el contenido semántico del fragmento. En este trabajo, cada texto queda representado mediante un embedding de 768 dimensiones.

Los embeddings resultantes se normalizan antes de ser almacenados, lo que permite posteriormente realizar comparaciones mediante similitud coseno.

Finalmente, los embeddings obtenidos se almacenan en una colección vectorial en ChromaDB, indexada mediante un grafo jerárquico HNSW (Hierarchical Navigable Small World), de modo que cada chunk K_i queda representado por un identificador, su embedding y los metadatos asociados.

Denotamos esta colección BGE como $C_{BGE} = \{id_i, K_i, u_i, meta_i\}_{i=1}^N$, donde id_i es el identificador único del chunk K_i , u_i es el embedding BGE normalizado y $meta_i$ los metadatos correspondientes.

3.2.2. Codificador de texto CLIP (Documentos externos)

Los mismos chunks K_i utilizados en la colección BGE se codifican adicionalmente mediante el encoder textual de CLIP (clip-ViT-L-14), implementado utilizando SentenceTransformers [40, 36]. A diferencia de BGE, cuyo propósito en este sistema es representar la similitud semántica entre textos, CLIP proyecta el contenido textual a un espacio multimodal compartido en el que imágenes y textos pueden compararse directamente. Cada chunk se procesa mediante el encoder textual de CLIP y se transforma en un embedding normalizado e_i .

Estas representaciones textuales normalizadas, junto con el contenido K_i y los metadatos meta_i , se almacenan en una colección vectorial específica para CLIP, $\mathcal{C}_{\text{CLIP}} = \{\text{id}_i, K_i, e_i, \text{meta}_i\}_{i=1}^N$, la cual se utiliza posteriormente para la recuperación imagen–texto en el espacio multimodal definido por el modelo.

3.3. Procesamiento de la pregunta

En esta sección se describe el procesamiento aplicado a la consulta del usuario, compuesta por la pregunta textual y la imagen asociada. El objetivo es transformar este par multimodal en representaciones vectoriales compatibles con las colecciones previamente construidas, de modo que puedan ejecutarse las etapas de recuperación texto–texto e imagen–texto.

3.3.1. Codificador de texto BGE (pregunta)

La pregunta original q se procesa utilizando el mismo encoder BGE empleado previamente para representar los documentos. Como resultado, se obtiene un embedding normalizado de la consulta, denotado como u_q , directamente comparable con los embeddings u_i de los chunks almacenados en la colección BGE. Esta representación se utiliza posteriormente durante la recuperación texto–texto.

3.3.2. Codificador de texto CLIP (imagen)

Para habilitar la recuperación de chunks textuales a partir de una imagen, empleamos el encoder visual de CLIP basado en Vision Transformer (ViT) [14]. El ViT divide la imagen en pequeños parches y los procesa mediante un Transformer para obtener una representación global de su contenido visual. CLIP proyecta posteriormente esta representación al mismo espacio multimodal utilizado por su encoder textual.

Después de su normalización, se obtiene un embedding visual k , que puede compararse directamente con los embeddings textuales e_i almacenados en la colección CLIP. De esta manera, el sistema puede recuperar fragmentos de texto utilizando directamente la información visual de la imagen, sin requerir una conversión completa de la imagen a lenguaje natural.

3.4. Recuperación híbrida

En esta sección se describe el mecanismo de recuperación híbrida propuesto, el cual constituye el núcleo del sistema RAG multimodal. A partir de las representaciones vectoriales de la consulta y de la base documental externa, se realizan dos procesos de recuperación complementarios: uno basado en similitud texto–texto mediante BGE y otro basado en similitud imagen–texto utilizando CLIP. Cada uno genera un ranking independiente de fragmentos candidatos. Posteriormente, estos rankings se combinan mediante el esquema de fusión Reciprocal Rank Fusion (RRF), con el objetivo de obtener un conjunto final de chunks que integre evidencia semántica y visual de manera coherente.

3.4.1. Búsqueda por similitud texto–texto

Una vez obtenido el embedding de la consulta u_q , realizamos la recuperación texto–texto sobre $\mathcal{C}_{\text{BGE}}$ para identificar los chunks más cercanos semánticamente. La similitud semántica entre la consulta y cada chunk se calcula como la similitud coseno entre estos embeddings:

$$\text{sim}(u_q, u_i) = \cos(u_q, u_i) = u_q^\top u_i. \quad (1)$$

Los valores $\text{sim}(u_q, u_i)$ para cada chunk K_i se ordenan de mayor a menor, de modo que los fragmentos con puntuaciones más altas corresponden a los textos más similares a la consulta. A partir de este ranking se seleccionan los chunks. En este trabajo se utilizan los 8 primeros.

3.4.2. Búsqueda por similitud imagen-texto

En el caso de la recuperación imagen-texto, la imagen se codifica mediante el encoder visual de CLIP explicado en la Subsubsección 3.3.2, generando un embedding visual normalizado k , mientras que cada chunk K_i está representado por el embedding textual normalizado e_i descrito en la Subsubsección 3.2.2. La similitud imagen-texto se modela como

$$\text{sim}(k, e_i) = \cos(k, e_i) = k^\top e_i, \quad (2)$$

donde los valores $\text{sim}(k, e_i)$ para cada chunk K_i se ordenan de mayor a menor, de modo que los chunks con puntuaciones más altas corresponden a los textos más similares a la imagen de la consulta. A partir de este ranking se seleccionan los chunks. En este trabajo se utilizan los 8 primeros.

De este modo, nuestra metodología dispone de dos listas ordenadas de candidatos (ocho fragmentos procedentes de BGE y ocho procedentes de CLIP), que se combinan posteriormente mediante el esquema de fusión RRF para obtener un subconjunto final de chunks con mayor soporte combinado.

3.4.3. Fusión de clasificaciones mediante Reciprocal Rank Fusion (RRF)

Una vez obtenidas las dos listas ordenadas de candidatos con mayor similitud texto-texto e imagen-texto, se combinan mediante el método RRF. Denotamos con D el conjunto de chunks candidatos y por R el conjunto de rankings que deseamos fusionar. En nuestro caso, $R = \{\text{text}, \text{img}\}$, donde *text* representa el ranking texto-texto obtenido con BGE y *img* el ranking imagen-texto obtenido con CLIP. Cada $r \in R$

asigna a cada chunk $d \in D$ una posición $r(d)$ en su lista ordenada de relevancia.

Para cada chunk $d \in D$, su puntuación RRF se define como

$$\text{RRFscore}(d) = \sum_{r \in R} \frac{1}{m + r(d)}, \quad (3)$$

donde $r(d)$ denota la posición del chunk d en el ranking r , comenzando en 1 para el elemento más relevante, y m es una constante de suavizado que controla el peso relativo de las posiciones más profundas del ranking. En este trabajo se fija $m = 60$ siguiendo la formulación original de RRF [12].

De este modo, cada chunk recibe una contribución inversamente proporcional a su posición en cada una de las listas, los fragmentos que aparecen en los primeros lugares de algún ranking obtienen contribuciones mayores, mientras que aquellos situados en posiciones más bajas reciben contribuciones menores. La suma sobre $r \in R$ hace que los chunks que son consistentemente bien posicionados tanto por BGE como por CLIP acumulen una puntuación RRF más alta que aquellos respaldados solo por una de las modalidades. Finalmente, los chunks candidatos se ordenan de mayor a menor según $\text{RRFscore}(d)$ y se seleccionan los cuatro chunks con mayor puntuación $C = \{B_1, B_2, B_3, B_4\}$, que constituyen el contexto final proporcionado al modelo generativo para la síntesis de la respuesta.

3.5. Generación de respuestas

A partir del contexto final recuperado, junto con la pregunta y la imagen correspondientes, se construye un mensaje de entrada estructurado que se envía a un LLM multimodal para generar la respuesta.

3.5.1. Prompt

Una vez obtenidos los chunks finales seleccionados mediante la fusión RRF, se construye el *prompt* de entrada textual para el modelo generativo. En primer lugar, se cargan instrucciones de control desde un archivo externo, las cuales definen el rol del modelo, las restricciones para el uso de la información recuperada y las

reglas de priorización del contenido incluido en [REFERENCES] (Anexo A). Las instrucciones se redactaron en inglés para mantener consistencia con PlantVillageVQA, cuyas preguntas, respuestas de referencia y contenidos empleados para construir la base de conocimiento se encuentran en este idioma. Esta decisión evita introducir variaciones derivadas de la traducción y permite conservar las condiciones lingüísticas originales durante la generación y evaluación de las respuestas.

Posteriormente, el contenido textual se estructura incorporando el contexto recuperado C y la pregunta original q , delimitados mediante las etiquetas [REFERENCES] y [QUESTION]. De esta manera, el LLM recibe un *prompt* compuesto por instrucciones, evidencia recuperada y la consulta del usuario, condicionando la generación de la respuesta en función de dichos elementos.

3.5.2. Respuesta

La respuesta final se genera mediante un LLM multimodal (GPT-4o), condicionado por la imagen x y por el *prompt* que integra la pregunta q y el contexto recuperado C . La imagen se incorpora como entrada visual junto con el texto estructurado mediante mensajes de tipo chat en configuraciones basadas en API, incluyendo `image_url` en formato `base64`. Las respuestas se generan en inglés para mantener consistencia con el idioma original de las preguntas, los documentos recuperados y las respuestas de referencia de PlantVillageVQA. De esta manera, se evitan posibles variaciones semánticas introducidas por un proceso de traducción previo a la evaluación con BERTScore. Finalmente, cada respuesta generada se almacena en un archivo independiente asociado a su correspondiente par de imagen y pregunta.

4. Métricas

La evaluación cuantitativa se realiza comparando cada respuesta generada con su referencia correspondiente, reportando resultados por categoría y promedios globales. Debido a que el conjunto de preguntas incluye respuestas abiertas y expresiones que pueden transmitir el

mismo significado utilizando palabras diferentes, se emplea BERTScore como métrica principal. Esta métrica permite evaluar la similitud semántica entre las respuestas generadas y sus respectivas referencias más allá de las coincidencias léxicas exactas.

4.1. BertScore

En este trabajo se utiliza BERTScore [56] como métrica principal para evaluar la calidad de las respuestas generadas. Esta métrica compara la respuesta y la referencia mediante embeddings contextualizados de BERT, capturando su similitud semántica. A diferencia de las métricas clásicas que se basan en coincidencias exactas, BERTScore permite identificar cuándo se transmite el mismo significado utilizando diferentes palabras, frases u orden gramatical, ofreciendo así una evaluación más flexible y alineada con la comprensión humana.

BERTScore proporciona tres métricas principales: precisión, recall y F1-score, calculadas en función de la similitud entre los embeddings de las palabras de la respuesta generada y la referencia. En la Figura 2 se muestra el proceso de evaluación mediante BERTScore, desde la obtención de los embeddings hasta el cálculo de la similitud mediante promedios ponderados.

El funcionamiento de BERTScore inicia transformando los tokens de las frases a embeddings contextualizados, de modo que cada representación incorpora información del contexto en el que aparece el token. Para ello, se emplea un modelo de lenguaje preentrenado basado en Transformers, cuyos embeddings permiten capturar relaciones semánticas entre los tokens. Esto permite evaluar la similitud entre expresiones que pueden diferir a nivel léxico, pero presentar un significado similar (ver Figura 2a).

Posteriormente, se calcula la similitud coseno entre los embeddings contextualizados de los tokens de la referencia y de la respuesta generada, formando una matriz de similitud que permite determinar las mejores correspondencias semánticas entre ambas secuencias (ver Figura 2b). Para cada token de una secuencia se identifica

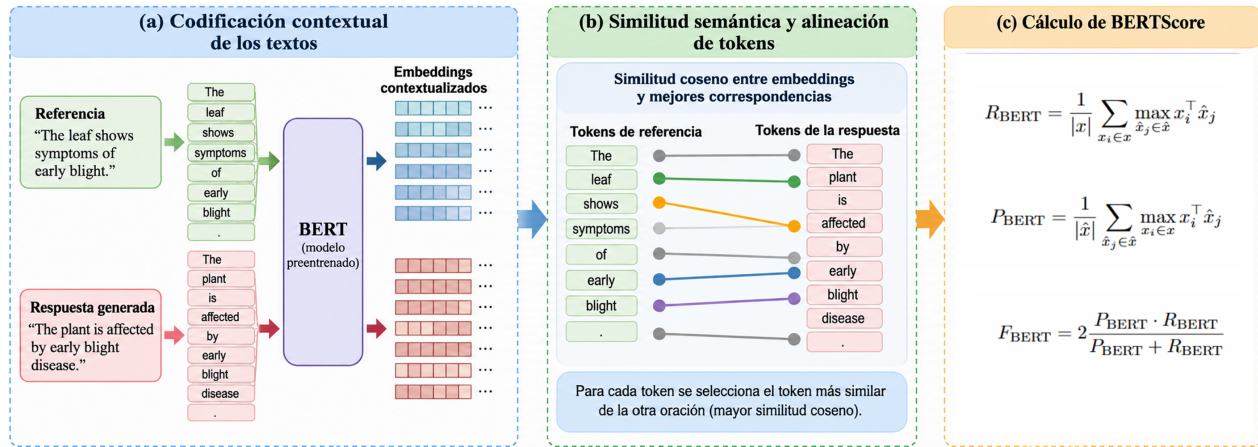


Fig. 2. Flujo general del cálculo de la métrica BERTScore basada en embeddings contextualizados

el token de la otra secuencia que presenta la mayor similitud.

A partir de estas correspondencias, BERTScore calcula la precisión y el recall considerando ambos sentidos de comparación (ver Figura 2c). La precisión promedia, para cada token de la respuesta generada, su mayor similitud con los tokens de la referencia. Por su parte, el recall promedia, para cada token de la referencia, su mayor similitud con los tokens de la respuesta generada. Finalmente, el F1-score se obtiene mediante la media armónica de precisión y recall. De esta forma, BERTScore proporciona una medida de similitud semántica entre la respuesta generada y la referencia, basada en sus representaciones contextuales.

Con base en la Figura 2, las métricas de BERTScore para una referencia (ground truth) x y una candidata $\hat{x}$ (respuesta generada) se definen como:

$$R_{BERT} = \frac{1}{|x|} \sum_{x_i \in x} \max_{\hat{x}_j \in \hat{x}} x_i^\top \hat{x}_j, \quad (4)$$

$$P_{BERT} = \frac{1}{|\hat{x}|} \sum_{\hat{x}_j \in \hat{x}} \max_{x_i \in x} x_i^\top \hat{x}_j, \quad (5)$$

$$F_{BERT} = 2 \frac{P_{BERT} \cdot R_{BERT}}{P_{BERT} + R_{BERT}}. \quad (6)$$

En estas expresiones, x denota la respuesta de referencia (ground truth) y $\hat{x}$ la respuesta generada, mientras que $|x|$ representa el número de tokens de la referencia y $|\hat{x}|$ representa el número de tokens de la respuesta generada. Los términos $x_i \in x$ y $\hat{x}_j \in \hat{x}$ corresponden a los embeddings contextualizados asociados a los tokens de cada secuencia.

El producto $x_i^\top \hat{x}_j$ representa la similitud entre ambos embeddings, equivalente a la similitud coseno cuando los embeddings se encuentran normalizados, y el operador $\max$ selecciona para cada token su mejor correspondencia en la otra secuencia.

5. Resultados

En esta sección se presentan los resultados medidos para distintas configuraciones de un sistema orientado a la descripción del estado de salud de cultivos. Se evalúan cuatro categorías de preguntas (Razonamiento contrafactual, enfermedad específica, razonamiento causal y descripción completa). Estas categorías provienen del subconjunto *razonamiento e inferencia de orden superior* con base en los datos de PlantVillageVQA [32]. En el subconjunto *enfermedad específica*, se

solicita un diagnóstico directo. En la *descripción completa*, se requiere una evaluación general de la planta y su condición. En el *razonamiento causal*, se busca inferir la causa o agente asociado a los síntomas. Finalmente, en el *razonamiento contrafactual*, se plantean escenarios hipotéticos para identificar qué rasgos serían distintos si la planta estuviera sana. A partir de ese subconjunto se construyeron dos particiones, una destinada a la evaluación y otra destinada a conformar la base de conocimiento externa usada en la recuperación, garantizando que no hubiera preguntas repetidas entre ambas.

5.1. Enfoques

Se analizan tres modelos representativos para VQA en el dominio agrícola, que difieren principalmente en su capacidad de integración de visión, lenguaje y nivel de razonamiento. LLaVA-1.6 y Qwen2-VL corresponden a LLMs multimodales orientados al seguimiento de instrucciones, capaces de producir explicaciones más naturales y fluidas a partir de la combinación de imagen y texto, pero frente a evidencia visual ambigua o consultas poco específicas tienden a introducir información no sustentada por la imagen o por conocimiento especializado. GPT-4o representa un LLM multimodal de gran escala con fuerte capacidad de razonamiento y generación, lo que favorece respuestas más coherentes y detalladas incluso si la imagen aporta evidencia parcial.

5.2. Resultados globales con BERTScore

La Tabla 1 presenta los resultados obtenidos con BERTScore para los modelos LLaVA-1.6, Qwen2-VL y GPT-4o bajo diferentes configuraciones. La Figura 3a complementa esta comparación mostrando el F1-score promedio de cada configuración, mientras que las Figuras 3b–d muestran el comportamiento por categoría de pregunta.

En términos globales, la incorporación de recuperación de información mejoró el F1-score promedio en los tres modelos evaluados. En LLaVA-1.6, el valor aumentó de 0.8567 en la configuración Base a 0.8700 con RAG, lo que representa una mejora relativa del 1.55 %. De manera similar, Qwen2-VL incrementó su

resultado de 0.8596 a 0.8712, equivalente a una mejora relativa del 1.35 %. Estos resultados indican que el contexto externo recuperado beneficia a modelos con diferentes capacidades de visión y lenguaje, al proporcionar información adicional para la generación de las respuestas.

En GPT-4o, la configuración Base obtuvo un F1-score promedio de 0.8590, mientras que la incorporación de RAG aumentó este valor a 0.8752. La configuración propuesta, que integra recuperación textual con BGE, recuperación visual con CLIP y fusión de rankings mediante RRF, alcanzó el mejor resultado global, con un F1-score de 0.8789. Esto corresponde a una mejora relativa del 2.32 % respecto a la configuración Base y del 0.42 % respecto a RAG. Aunque el incremento respecto a RAG es menor, muestra que la incorporación de la señal visual en la recuperación y su combinación con la recuperación textual aporta información complementaria al contexto utilizado por el modelo generativo.

El análisis por categoría muestra que el efecto de la metodología propuesta no fue uniforme. En razonamiento contrafactual, GPT-4o con RAG obtuvo un F1-score de 0.8950, mientras que la configuración propuesta alcanzó 0.8949. La diferencia es mínima, por lo que ambas configuraciones pueden considerarse prácticamente equivalentes en esta categoría. No obstante, ambas superaron claramente el valor de 0.8701 obtenido por GPT-4o Base, lo que evidencia que la recuperación de contexto resulta especialmente útil para responder preguntas sobre escenarios hipotéticos.

En identificación de enfermedades específicas, la configuración propuesta obtuvo un F1-score de 0.8845, superando a GPT-4o con RAG, que alcanzó 0.8809, y a GPT-4o Base, con 0.8631. Este resultado representa una mejora relativa del 2.48 % frente a la configuración Base. El incremento sugiere que combinar la pregunta con las características visuales de la planta durante la recuperación ayuda a seleccionar fragmentos con información más relacionada con los síntomas y las enfermedades observadas.

La mayor contribución de la recuperación multimodal se observó en razonamiento causal. En esta categoría, el F1-score aumentó de 0.8559 con GPT-4o Base a 0.8704 con RAG y finalmente

Tabla 1. Comparación del rendimiento entre diferentes modelos, configuraciones y categorías de preguntas

Modelo	Configuración	Categoría de pregunta	Precisión	Recall	F1-score
LLaVA-1.6	Base	Razonamiento contrafactual	0.8668	0.8812	0.8739
		enfermedad específica	0.8582	0.8622	0.8599
		Razonamiento causal	0.8477	0.8562	0.8517
		Descripción completa	0.8308	0.8527	0.8414
		Promedio	0.8509	0.8631	0.8567
	RAG	Razonamiento contrafactual	0.8779	0.8927	0.8851
		enfermedad específica	0.8835	0.8729	0.8779
		Razonamiento causal	0.8638	0.8738	0.8683
		Descripción completa	0.8330	0.8652	0.8485
		Promedio	0.8646	0.8762	0.8700
Qwen2-VL	Base	Razonamiento contrafactual	0.8677	0.8823	0.8749
		enfermedad específica	0.8561	0.8671	0.8613
		Razonamiento causal	0.8471	0.8612	0.8538
		Descripción completa	0.8384	0.8591	0.8484
		Promedio	0.8523	0.8674	0.8596
	RAG	Razonamiento contrafactual	0.8720	0.8886	0.8802
		enfermedad específica	0.8805	0.8755	0.8777
		Razonamiento causal	0.8633	0.8792	0.8708
		Descripción completa	0.8405	0.8728	0.8561
		Promedio	0.8641	0.8790	0.8712
GPT-4o	Base	Razonamiento contrafactual	0.8609	0.8798	0.8701
		enfermedad específica	0.8520	0.8752	0.8631
		Razonamiento causal	0.8450	0.8676	0.8559
		Descripción completa	0.8352	0.8594	0.8468
		Promedio	0.8483	0.8705	0.8590
	RAG	Razonamiento contrafactual	0.8918	0.8985	0.8950
		enfermedad específica	0.8860	0.8765	0.8809
		Razonamiento causal	0.8601	0.8817	0.8704
		Descripción completa	0.8410	0.8691	0.8545
		Promedio	0.8697	0.8815	0.8752
	Our	Razonamiento contrafactual	0.8911	0.8990	0.8949
		enfermedad específica	0.8858	0.8840	0.8845
		Razonamiento causal	0.8790	0.8853	0.8817
		Descripción completa	0.8421	0.8680	0.8546
		Promedio	0.8745	0.8841	0.8789

a 0.8817 con la configuración propuesta. Esto supone una mejora relativa del 3.01 % respecto

a la configuración Base y del 1.30 % respecto a RAG. Este comportamiento indica que la fusión

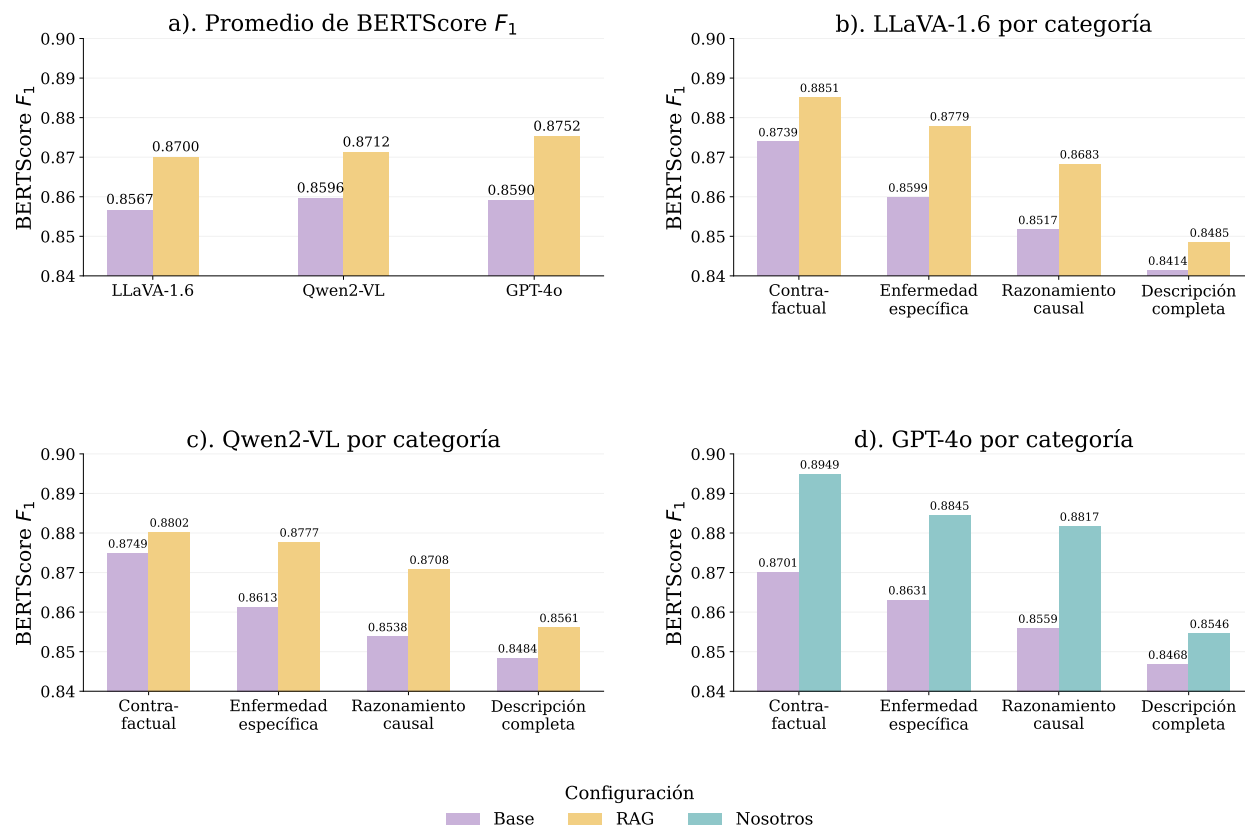


Fig. 3. Resultados obtenidos mediante BERTScore. La figura (a) presenta el F1-score promedio de las configuraciones evaluadas. Las figuras (b) y (c) comparan las configuraciones Base y RAG para LLaVA-1.6 y Qwen2-VL, respectivamente, mientras que la figura (d) compara la configuración Base con la propuesta para GPT-4o

mediante RRF resulta particularmente útil cuando la respuesta requiere relacionar los síntomas visibles con sus posibles causas o agentes asociados, debido a que permite integrar evidencia recuperada tanto desde la consulta textual como desde la imagen.

Por otro lado, la descripción completa presentó los valores más bajos en todos los modelos y configuraciones. La configuración propuesta alcanzó un F1-score de 0.8546, frente a 0.8545 con RAG y 0.8468 con GPT-4o Base.

Aunque se observa una mejora respecto a la configuración Base, el resultado es prácticamente igual al obtenido con RAG. Esta menor diferencia puede deberse a que las preguntas de descripción completa requieren integrar simultáneamente diversos aspectos de la planta, como su apariencia

general, síntomas, condición sanitaria y posibles causas, lo que incrementa la variabilidad y extensión de las respuestas generadas.

En conjunto, los resultados muestran que el uso de conocimiento externo mejora el desempeño de los tres modelos evaluados. Además, la configuración propuesta obtuvo el mayor F1-score promedio y mostró su principal ventaja en las categorías de identificación de enfermedades específicas y razonamiento causal.

Por tanto, la recuperación guiada por imagen no sustituye a la recuperación textual, sino que aporta una señal complementaria que permite construir un contexto más informativo para responder preguntas que dependen directamente de los síntomas visuales del cultivo.

6. Conclusiones

Hoy en día, diferentes trabajos han mostrado que incorporar conocimiento externo puede mejorar la respuesta a preguntas complejas. Sin embargo, gran parte de estos trabajos se ha centrado en dominios generales, mientras que en el ámbito agrícola la literatura es aún limitada, particularmente en VQA multimodal apoyado por RAG. Esto es relevante porque, en consultas agrícolas, especialmente cuando el usuario no es experto, es común que la pregunta omita detalles importantes y que la imagen no siempre muestre información explícita. En esas condiciones, el sistema puede generar respuestas incompletas o poco confiables si no cuenta con contexto adicional. Por lo tanto, el reto es construir un agente que integre evidencia visual con conocimiento externo consultable, reduzca el riesgo de alucinaciones y produzca respuestas más útiles y consistentes para describir el estado de salud de los cultivos. Para abordar estas limitaciones, en este trabajo se propuso un RAG multimodal para VQA agrícola que utiliza tanto la imagen como la consulta para recuperar evidencia contextual especializada y genera respuestas explicativas en lenguaje natural.

La evaluación se realizó mediante BERTScore en cuatro categorías de preguntas del subconjunto de razonamiento e inferencia de orden superior de PlantVillageVQA. Los resultados mostraron mejoras consistentes al incorporar conocimiento externo en los tres modelos evaluados. La metodología propuesta alcanzó el mejor F1-score promedio con GPT-4o, obteniendo 0.8789, lo que representa una mejora relativa del 2.32% respecto a la configuración Base y del 0.42% respecto a RAG. La mayor contribución se observó en las preguntas de razonamiento causal e identificación de enfermedades específicas, lo que sugiere que combinar evidencia textual y visual resulta especialmente útil cuando la respuesta requiere relacionar los síntomas observados con una enfermedad o con sus posibles causas.

En conjunto, los resultados muestran que la recuperación guiada por imagen complementa la recuperación textual y puede contribuir a seleccionar un contexto más informativo para la

generación de respuestas. La fusión mediante RRF permite integrar ambas fuentes de evidencia dentro de una arquitectura modular, sin necesidad de realizar ajuste fino ni reentrenar el modelo generativo. Sin embargo, su aporte no fue uniforme en todas las categorías, ya que en razonamiento contrafactual y descripción integral los resultados fueron similares a los obtenidos mediante RAG. Esto indica que la utilidad de la evidencia visual recuperada depende de la naturaleza de la pregunta y de la información disponible en la imagen.

Los resultados obtenidos constituyen una validación inicial favorable de la metodología propuesta sobre PlantVillageVQA. Estos resultados muestran el potencial de combinar información textual y visual para recuperar conocimiento especializado y generar respuestas con mayor correspondencia semántica respecto a las referencias. Asimismo, la evaluación realizada establece una base experimental para continuar analizando el comportamiento del sistema en escenarios agrícolas más diversos.

Finalmente, se concluye que la combinación de recuperación texto-texto, recuperación imagen-texto y fusión mediante RRF constituye una alternativa viable para incorporar conocimiento agrícola especializado en sistemas VQA multimodales. La metodología puede servir como base para herramientas de apoyo destinadas al monitoreo de cultivos y la asistencia preliminar a usuarios no especializados. Como trabajo futuro, se propone ampliar la base de conocimiento con documentos agronómicos especializados, evaluar el sistema con imágenes capturadas en campo bajo diferentes condiciones visuales, complementar BERTScore con métricas orientadas a la exactitud diagnóstica y analizar las respuestas generadas con la participación de especialistas del sector agrícola.

A. Instrucciones del prompt

Este anexo presenta las instrucciones de control utilizadas para formular el *prompt* de entrada del modelo generativo. Estas reglas definen el rol del asistente y la manera en que debe priorizar el

contenido recuperado mediante RAG, el cual se denota como [REFERENCES].

You are an agricultural assistant. Your goal is to help answer plant health or related questions based on an image and a set of retrieved reference documents ([REFERENCES]).

CONSTRAINTS:

1. Use the image only to supplement information if needed.
2. Only generate a new answer yourself if [REFERENCES] is empty or completely irrelevant.
3. Your final answer MUST prioritize any content in [REFERENCES].
4. Do not explain or justify answers. Just respond directly and clearly.
5. Be literal. Do not filter or validate content from [REFERENCES].
6. If [REFERENCES] contains a direct or partial answer to the question, you MUST use it as-is. DO NOT correct, rephrase, interpret, or validate it.

Referencias

1. **Anisuzzaman, D., Malins, J. G., Friedman, P. A., Attia, Z. I. (2025).** Fine-tuning large language models for specialized use cases. *Mayo Clinic Proceedings: Digital Health*, Vol. 3, No. 1, pp. 100184.
2. **Awais, M., Salem Abdulla Alharthi, A. H., Kumar, A., Cholakkal, H., Anwer, R. M. (2025).** AgroGPT: Efficient agricultural vision-language model with expert tuning. *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 5687–5696. DOI: 10.1109/WACV61041.2025.00555.
3. **BAAI (2023).** BGE: BAAI general embedding. <https://huggingface.co/BAAI/bge-large-en-v1.5>. Model card and official documentation. Accessed: 2026-02-16.
4. **Balaguer, A., Benara, V., Cunha, R. L. d. F., Hendry, T., Holstein, D., Marsman, J., Mecklenburg, N., Malvar, S., Nunes, L. O., Padilha, R., et al. (2024).** RAG vs fine-tuning: pipelines, tradeoffs, and a case study on agriculture. *arXiv preprint arXiv:2401.08406*.
5. **Bock, C., El Jarroudi, M., Kouadio, L., Mackels, C., Chiang, K.-S., Delfosse, P. (2015).** Disease severity estimates—effects of rater accuracy and assessment methods for comparing treatments. *Plant Disease*, Vol. 99, No. 8, pp. 1104–1112.
6. **Bock, C., Poole, G., Parker, P., Gottwald, T. (2010).** Plant disease severity estimated visually, by digital photography and image analysis, and by hyperspectral imaging. *Critical reviews in plant sciences*, Vol. 29, No. 2, pp. 59–107.
7. **Bock, C. H., Barbedo, J. G., Del Ponte, E. M., Bohnenkamp, D., Mahlein, A.-K. (2020).** From visual estimates to fully automated sensor-based measurements of plant disease severity: status and challenges for improving accuracy. *Phytopathology Research*, Vol. 2, No. 1, pp. 9.
8. **Bock, C. H., Chiang, K.-S., Del Ponte, E. M. (2022).** Plant disease severity estimated visually: a century of research, best practices, and opportunities for improving methods and practices to maximize accuracy. *Tropical Plant Pathology*, Vol. 47, No. 1, pp. 25–42.
9. **Chen, B., Khare, A., Kumar, G., Akula, A., Narayana, P. (2025).** Seeing beyond: Enhancing visual question answering with multi-modal retrieval. *Proceedings of the 31st International Conference on Computational Linguistics: Industry Track*, pp. 410–421.
10. **Chen, Z., Huang, Y., Chen, J., Geng, Y., Fang, Y., Pan, J. Z., Zhang, N., Zhang, W. (2022).** Lako: Knowledge-driven visual question answering via late knowledge-to-text injection. *Proceedings of the 11th International Joint Conference on Knowledge Graphs*, pp. 20–29.
11. **Chiang, K.-S., Bock, C., El Jarroudi, M., Delfosse, P., Lee, I., Liu, H. (2016).** Effects of rater bias and assessment method on disease severity estimation with regard to hypothesis testing. *Plant Pathology*, Vol. 65, No. 4, pp. 523–535.

12. **Cormack, G. V., Clarke, C. L., Buettcher, S. (2009).** Reciprocal rank fusion outperforms condorcet and individual rank learning methods. Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval, pp. 758–759.
13. **Del Ponte, E. M., Pethybridge, S. J., Bock, C. H., Michereff, S. J., Machado, F. J., Spolti, P. (2017).** Standard area diagrams for aiding severity estimation: scientometrics, pathosystems, and methodological trends in the last 25 years. *Phytopathology*, Vol. 107, No. 10, pp. 1161–1174.
14. **Dosovitskiy, A. (2020).** An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.
15. **Ejaz, M. H., Bilal, M., Habib, U., Attique, M., Chung, T.-S. (2025).** Crop pest classification using deep learning techniques: A review. arXiv preprint arXiv:2507.01494.
16. **Gardères, F., Ziaeeafard, M., Abeloos, B., Lecue, F. (2020).** Conceptbert: Concept-aware representation for visual question answering. Findings of the Association for Computational Linguistics: EMNLP 2020, pp. 489–498.
17. **Group on Earth Observations (GEO) (2022).** Integrating earth observations into the formulation and implementation of national adaptation plans: Agriculture and food security. GEO Supplement to the UNFCCC NAP Technical Guidelines. Consultado el 2025-10-02.
18. **Gui, L., Wang, B., Huang, Q., Hauptmann, A., Bisk, Y., Gao, J. (2021).** Kat: A knowledge augmented transformer for vision-and-language. arXiv preprint arXiv:2112.08614.
19. **Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., Casas, D. d. L., Hendricks, L. A., Welbl, J., Clark, A., et al. (2022).** Training compute-optimal large language models. arXiv preprint arXiv:2203.15556.
20. **Hu, Z., Iscen, A., Sun, C., Wang, Z., Chang, K.-W., Sun, Y., Schmid, C., Ross, D. A., Fathi, A. (2023).** Reveal: Retrieval-augmented visual-language pre-training with multi-source multimodal knowledge memory. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 23369–23379.
21. **Izacard, G., Grave, E. (2020).** Leveraging passage retrieval with generative models for open domain question answering. arXiv preprint arXiv:2007.01282.
22. **LangChain (2025).** Langchain overview.
23. **LangChain (2026).** Recursive character text splitter. https://docs.langchain.com/oss/python/integrations/splitters/recursive_text_splitter. LangChain Official Documentation. Accessed: 2026-02-16.
24. **Lastres, L., Soza, F. (2009).** Manual de Sanidad Vegetal. Escuela Agrícola Panamericana (Zamorano), Carrera de Ciencia y Producción Agropecuaria, El Zamorano, Honduras. Programa PASOLAC; Programa PROMIPAC. Consultado el 2025-10-02.
25. **Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., et al. (2020).** Retrieval-augmented generation for knowledge-intensive NLP tasks. Advances in neural information processing systems, Vol. 33, pp. 9459–9474.
26. **Lin, W., Byrne, B. (2022).** Retrieval augmented visual question answering with outside knowledge. arXiv preprint arXiv:2210.03809.
27. **Long, X., Ma, Z., Hua, E., Zhang, K., Qi, B., Zhou, B. (2025).** Retrieval-augmented visual question answering via built-in autoregressive search engines. Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 39, No. 23, pp. 24723–24731.
28. **Lukáš, J., Stejskal, V. (2003).** Computer-based image analysis to estimate the area of a sticky trap occupied or contaminated by pests. *Plant Protection Science*, Vol. 39, No. 2, pp. 52.

29. Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., et al. (2020). Language models are few-shot learners. arXiv preprint arXiv:2005.14165, Vol. 1, No. 3, pp. 3.
30. Mavroudis, V. (2024). LangChain v0.3. DOI: 10.20944/preprints202411.0566.v1. Working paper or preprint.
31. Mondal, S. (2025). Vision meets language: A RAG-augmented YOLOv8 framework for coffee disease diagnosis and farmer assistance. arXiv preprint arXiv:2505.21544.
32. Nazmus Sakib, S., Haque, N., Zayed Hos-sain, M., Arman, S. E. (2025). PlantVillageV-QA: A visual question answering dataset for benchmarking vision-language models in plant science. arXiv e-prints, pp. arXiv-2508.
33. Pérez-Patricio, M., Osuna-Coutiño, J. d. J., Ríos-Toledo, G., Aguilar-González, A., Camas-Anzueto, J., Morales-Navarro, N., Velázquez-González, J. R., Cundapí-López, L. Á. (2024). Plant stress detection using a three-dimensional analysis from a single RGB image. *Sensors*, Vol. 24, No. 23, pp. 7860.
34. Pérez-Roncal, C., López-Maestresalas, A., Lopez-Molina, C., Jarén, C., Urrestarazu, J., Santesteban, L. G., Arazuri, S. (2020). Hyperspectral imaging to assess the presence of powdery mildew (*Erysiphe necator*) in cv. Carignan noir grapevine bunches. *Agronomy*, Vol. 10, No. 1, pp. 88.
35. Rackauckas, Z. (2024). Rag-fusion: a new take on retrieval-augmented generation. arXiv preprint arXiv:2402.03367.
36. Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. (2021). Learning transferable visual models from natural language supervision. *International conference on machine learning*, PMLR, pp. 8748–8763.
37. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, Vol. 1, No. 8, pp. 9.
38. Rajendran, G., Islam, R. (2017). Plant clinics in Bangladesh: are farmers losing less and feeding more. *CABI case study*, Vol. 19.
39. Ravi, S., Chinchure, A., Sigal, L., Liao, R., Schwartz, V. (2023). Vlc-bert: Visual question answering with contextualized commonsense knowledge. *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 1155–1165.
40. Reimers, N., Gurevych, I. (2024). Sentence transformers: Multilingual sentence, paragraph, and image embeddings.
41. Rezatec Limited (2021). Case study: Crop observation, management & production analysis services system (COMPASS). Case study. *Agricultural Innovation in Mexico and Argentina*. Consultado el 2025-10-02.
42. Ríos-Toledo, G., Pérez-Patricio, M., Cundapí-López, L. Á., Camas-Anzueto, J.-L., Morales-Navarro, N., Osuna-Coutiño, J. d. J. (2023). Plant stress recognition using deep learning and 3d reconstruction. *Mexican Conference on Pattern Recognition*, Springer, pp. 114–124.
43. Roumeliotis, K. I., Sapkota, R., Karkee, M., Tselikas, N. D., Nasiopoulos, D. K. (2025). Plant disease detection through multimodal large language models and convolutional neural networks. arXiv preprint arXiv:2504.20419.
44. Samuel, D. J., Skarga-Bandurova, I., Sikolia, D., Awais, M. (2025). AgroLLM: Connecting farmers and agricultural practices through large language models for enhanced knowledge transfer and practical application. arXiv preprint arXiv:2503.04788.
45. Shi, T., Liu, Y., Zheng, X., Hu, K., Huang, H., Liu, H., Huang, H. (2023). Recent advances in plant disease severity assessment using convolutional neural networks. *Scientific Reports*, Vol. 13, No. 1, pp. 2336.
46. Taylor, P. (2015). *Plantwise Diagnostic Field Guide*. CABI / Plantwise. A tool to diagnose

crop problems and make recommendations for their management. Consultado el 2025-10-02.

47. **Terentev, A., Dolzhenko, V., Fedotov, A., Eremenko, D. (2022).** Current state of hyperspectral remote sensing for early plant disease detection: A review. *Sensors*, Vol. 22, No. 3, pp. 757.
48. **Ullah, N., Khan, J. A., Alharbi, L. A., Raza, A., Khan, W., Ahmad, I. (2022).** An efficient approach for crops pests recognition and classification based on novel deeppestnet deep learning model. *IEEE Access*, Vol. 10, pp. 73019–73032.
49. **Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., Polosukhin, I. (2017).** Attention is all you need. *Advances in neural information processing systems*, Vol. 30.
50. **Velázquez-González, J. R., Pérez-Patricio, M., Osuna-Coutiño, J. D. J., Camas-Anzueto, J. L., Aguilar-González, A., Morales-Navarro, N., Hernández-Gutiérrez, C. A. (2024).** Supervisory configuration of deep learning networks for plant stress detection and synthetic dataset generation. *IEEE Access*.
51. **Venbrux, M., Crauwels, S., Rediers, H. (2023).** Current and emerging trends in techniques for plant pathogen detection. *Frontiers in Plant Science*, Vol. 14, pp. 1120968.
52. **Wang, S., Zhu, Y., Liu, H., Zheng, Z., Chen, C., Li, J. (2024).** Knowledge editing for large language models: A survey. *ACM Computing Surveys*, Vol. 57, No. 3, pp. 1–37.
53. **Wang, Y., Wang, F., Chen, W., Lv, B., Liu, M., Kong, X., Zhao, C., Pan, Z. (2025).** A large language model for multimodal identification of crop diseases and pests. *Scientific Reports*, Vol. 15, No. 1, pp. 21959.
54. **Williams, F., Hy HengChhun, H. H., Ratanakarn, W., Luechaikarm, C., Channoo, C., Win, K., Pham Thi Xuan, P. T. X., Tran Danh Suu, T. D. S., Khing SuLi, K. S., Muhammad Faheem, M. F., et al. (2018).** Plant clinics in Asia: reducing the use and risks of pesticides.
55. **Yang, S., Liu, Z., Mayer, W., Ding, N., Wang, Y., Huang, Y., Wu, P., Li, W., Li, L., Zhang, H.-Y., et al. (2024).** ShizishanGPT: An agricultural large language model integrating tools and resources. *International Conference on Web Information Systems Engineering*, Springer, pp. 284–298.
56. **Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., Artzi, Y. (2019).** Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.
57. **Zhang, Y., Fan, Q., Chen, X., Li, M., Zhao, Z., Li, F., Guo, L. (2025).** IPM-AgriGPT: A large language model for pest and disease management with a G-EA framework and agricultural contextual reasoning. *Mathematics* (2227-7390), Vol. 13, No. 4.
58. **Zhao, J., Xu, L., Ma, Z., Li, J., Wang, X., Liu, Y., Du, X. (2025).** A review of plant leaf disease identification by deep learning algorithms. *Frontiers in Plant Science*, Vol. 16, pp. 1637241.

Article received on 02/06/2026; accepted on 23/08/2026.
**Corresponding author is J. A. de Jesús Osuna-Coutiño.*