UNIVERSIDAD NACIONAL DE INGENIERÍA

UNIDAD DE POSTGRADO FACULTAD DE INGENIERIA INDUSTRIAL Y DE SISTEMAS

PLAN DE TESIS

“Sistema de Consultas Multimodal basado en RAG para Manuales de Mantenimiento en Plantas Concentradoras de Cobre”

PARA OBTENER EL GRADO ACADEMICO DE MAESTRO EN CIENCIAS CON MENCIÓN EN INTELIGENCIA ARTIFICIAL

ELABORADA POR:
JOHAN MANUEL CALLOMAMANI BUENDIA

ASESOR: DR. GLEN DARIO RODRIGUEZ RAFAEL

LIMA, PERU
2025

Capítulo 1: Planteamiento del Problema

En el contexto de la Industria 4.0 y la transformación digital, el volumen de información técnica disponible en las operaciones industriales ha crecido exponencialmente. En sectores críticos como la minería, y específicamente en las **plantas concentradoras de cobre**, la documentación técnica que incluye manuales de instalación, operación y mantenimiento (IOM), catálogos de partes, hojas de especificaciones, filosofías de control, arreglos generales y reportes de calidad— constituye la base fundamental para la toma de decisiones en mantenimiento. Estos activos de información son vitales para garantizar la continuidad operativa; sin embargo, suelen presentarse en formatos heterogéneos y no estructurados (PDFs escaneados, imágenes, planos y diagramas), lo que dificulta su gestión y accesibilidad inmediata.

Actualmente, la recuperación de información específica dentro de estos documentos presenta limitaciones significativas. Los técnicos, ingenieros y planificadores dependen de técnicas tradicionales como la lectura secuencial, el uso de índices estáticos (tablas de contenido), la búsqueda por palabras clave (*keyword search*) y búsqueda secuencial. Estos métodos resultan ineficientes ante la complejidad de los manuales modernos y antiguos, los cuales suelen caracterizarse por su gran extensión, terminología técnica ambigua y barreras idiomáticas. Más aún, existe una **desconexión semántica** en la búsqueda: la información crítica a menudo reside en formatos multimodales (como un plano de despiece o un diagrama de flujo) que los motores de búsqueda de texto tradicionales no pueden interpretar ni relacionar con las instrucciones escritas.

Esta ineficiencia en la gestión del conocimiento impacta directamente en la gestión del mantenimiento. Una planificación efectiva requiere identificar con precisión el procedimiento de cambio, los repuestos exactos (ubicados en listas o planos de partes), las herramientas especiales y los tiempos estimados. La demora o el error en la localización de estos datos no solo consume horas-hombre valiosas de ingeniería, sino que incrementa el riesgo de errores en la ejecución («mantenimiento incorrecto») o retrasos en la intervención. En una planta concentradora, donde la disponibilidad de los equipos es crítica, esta latencia en el acceso a la información puede traducirse en paradas de planta prolongadas y pérdidas significativas en la producción de cobre.

1.1 Diagnóstico

En el área de planificación y mantenimiento de las plantas concentradoras de cobre, se observa una gestión documental dispersa y poco funcional. Actualmente, los manuales de equipos críticos y no críticos (como molinos SAG, chancadoras, bombas y celdas de flotación) se almacenan en repositorios digitales masivos (SharePoint, servidores locales) sin una indexación semántica adecuada, el almacenamiento actual esta basado en como se adquirieron en la etapa de proyecto por procesos, flota o grupos de equipos (Ejemplo: Todas las bombas de la planta están agrupadas en una sola carpeta ya que todas las bombas de lodo fueron compradas a una misma empresa, caso similar con las celdas de flotación todas están en una sola carpeta y con un solo pdf por todos los tipos de celdas).

Se evidencia que, ante una falla o una parada programada, los planificadores/programadores de mantenimiento invierten un tiempo excesivo navegando manualmente entre carpetas y

archivos PDF extensos —algunos de los cuales son documentos escaneados («imágenes de texto»)— lo que impide el uso de herramientas de búsqueda convencionales (Ctrl+F). Un síntoma recurrente es la dificultad para correlacionar la información visual con la textual; por ejemplo, el técnico encuentra el procedimiento de desmontaje en la página 50, pero el plano de despiece con los códigos de repuestos está en un anexo al final del documento o en un archivo separado, obligando a una validación manual cruzada propensa a errores. Además, la existencia de manuales en inglés técnico complejo genera barreras de comprensión inmediata para parte del personal operativo, retrasando la ejecución de las órdenes de trabajo.

Se evidencia que los técnicos no realizan la búsqueda de especificaciones técnicas en manuales, generando que 6 de cada 10 condiciones de equipos reportados carecen de información suficiente para la gestión de planificación del mantenimiento, especialmente en la identificación de los repuestos a cambiar.

1.2 Identificación y Diagnóstico del Problema de Estudio

El problema central identificado no es la inexistencia de información, sino la incapacidad de los sistemas de búsqueda actuales para procesar y relacionar información multimodal (texto e imagen) contenida en documentos técnicos no estructurados.

Las técnicas de búsqueda tradicionales (basadas por palabras clave, lectura secuencial o búsqueda por tabla de contenidos) resultan insuficientes para interpretar consultas complejas de mantenimiento que requieren contexto, como «procedimiento de cambio de liner considerando el torque especificado en el plano A». Existe una brecha tecnológica entre la naturaleza heterogénea de los manuales IOM (que combinan diagramas, tablas de especificaciones y narrativas técnicas) y los mecanismos de recuperación de información disponibles, los cuales tratan el texto y la imagen como entidades desconectadas.

Esta limitación tecnológica deriva en una baja precisión y retrabajo en las consultas técnicas, lo que impacta negativamente en el tiempo medio de reparación (MTTR) y en la confiabilidad de la planificación de mantenimiento. Por lo tanto, el problema de estudio se define como la ineficiencia en la recuperación de información técnica contextualizada debido a la falta de integración semántica entre los datos textuales y visuales en los repositorios de mantenimiento de plantas mineras.

1.2.1 Antecedentes bibliográficos

1.2.2 Formulación del Problema

1.2.2.1 Formulación del Problema General

¿De qué manera la implementación de un Sistema de Consultas Multimodal basado en Arquitectura RAG (Retrieval-Augmented Generation) optimiza la recuperación de información técnica en los manuales de mantenimiento de plantas concentradoras de cobre, en comparación con los métodos de búsqueda tradicionales?

1.2.2.2 Formulación de los Problemas Específicos

1. ¿Cómo influye la integración semántica de datos multimodales (texto, planos, diagramas e imágenes) en la capacidad del sistema para responder consultas técnicas contextuales que requieren interpretación visual, a diferencia de la búsqueda puramente textual?

2. ¿Cuál es la mejora en la precisión y exhaustividad (recall) de la información recuperada al utilizar técnicas de embedding y re-ranking vectorial frente a la búsqueda léxica (palabras clave) en manuales con terminología heterogénea?
3. ¿En qué medida se reduce el tiempo de búsqueda de información crítica (procedimientos, especificaciones de repuestos y herramientas) para la planificación de mantenimiento al utilizar el asistente conversacional basado en RAG?

1.3 Objetivo General

Desarrollar un Sistema de Consultas Multimodal basado en arquitectura RAG (Retrieval-Augmented Generation) para optimizar la eficiencia y precisión en la recuperación de información técnica de los manuales de mantenimiento en plantas concentradoras de cobre.

1.4 Objetivo Especifico

Estos objetivos representan los pasos técnicos y validaciones necesarias para alcanzar el objetivo general. Están alineados 1 a 1 con tus problemas específicos:

1. Diseñar e implementar un pipeline de procesamiento de datos multimodal que permita la extracción, vectorización e indexación conjunta de texto no estructurado y esquemas visuales (planos de partes, diagramas procedimientos, imágenes) contenidos en los manuales de mantenimiento.
2. Evaluar el desempeño del motor de recuperación mediante métricas de relevancia (Precision y Recall) BERTscore y ROUGE L(Recall Oriented Understudy for Gisting Evaluation).
3. Validar la utilidad del sistema en un entorno operativo, cuantificando la reducción del tiempo empleado por los planificadores en la búsqueda de información crítica y atención de reportes de condición.

Capítulo 2: Marco Teórico y Estado del Arte

2.1 Bases Teóricas

1. **Taxonomía de Datos de Confiabilidad (ISO 14224)** Para que un Modelo de Lenguaje Grande (LLM) pueda razonar eficazmente sobre mantenimiento, debe «entender» la estructura jerárquica de los equipos industriales. La norma ISO 14224 proporciona el estándar de facto para la recolección e intercambio de datos de confiabilidad y mantenimiento. Aunque originada en la industria del petróleo y gas, su adopción en la minería es generalizada para estructurar los sistemas ERP (como SAP PM) y CMMS.

   La norma establece una jerarquía taxonómica que permite descomponer un activo complejo en unidades manejables. Esta estructura es vital para el diseño de la base de datos vectorial del sistema RAG, ya que permite la creación de metadatos precisos para el filtrado de búsquedas (Metadata Filtering).

Nivel Taxonómico (ISO 14224)	Ejemplo en Planta Concentradora	Aplicación en Sistema RAG
Nivel 3: Instalación	Planta Concentradora de Cobre	Contexto global del sistema.
Nivel 4: Sistema	Circuito de Molienda SAG	Delimitación del alcance de la consulta.
Nivel 5: Sub-sistema	Sistema de Lubricación de Chumaceras	Agrupación funcional de documentos.
Nivel 6: Equipo (Unit)	Unidad de Bombeo de Alta Presión	Entidad principal de consulta (Subject).
Nivel 7: Sub-unidad	Bomba de desplazamiento positivo	Componente específico de falla.
Nivel 8: Pieza (Part)	Sello Mecánico / Empaquetadura	Objeto de la instrucción de recambio.
Nivel 9: Ítem Mantenible	Anillo tórico (O-ring)	Detalle granular para repuestos.

2. **Procesos Críticos y Equipos Principales** La comprensión del dominio minero es esencial para evaluar la relevancia de las respuestas generadas. Los manuales técnicos en este sector cubren procesos con físicas y modos de falla distintos:

   Conminución (Chancado y Molienda): Es la etapa más intensiva en energía. Los manuales de Chancadores Giratorios y Cónicos (ej. Metso MP, FLSmidth) contienen procedimientos críticos de ajuste del setting (CSS) y cambio de revestimientos (mantles/bowls), tareas que involucran manipulación de componentes de varias toneladas. Los Molinos SAG y de Bolas requieren mantenimiento especializado en sus sistemas de transmisión (coronas, piñones) y lubricación hidrostática, donde un error en la interpretación de las tolerancias de presión de aceite puede fundir una chumacera.

   Flotación: Involucra Celdas de Flotación (mecánicas, neumáticas, columnas) donde el mantenimiento se centra en los mecanismos de agitación (rotores, estatores) y sistemas de instrumentación. La documentación técnica aquí es rica en diagramas de lazos de control y esquemas de distribución de aire.

   Gestión de Fluidos y Relaves: Los Espesadores y Bombas de Relaves son críticos para la continuidad hídrica. Los manuales detallan el mantenimiento de rastras hidráulicas y sistemas de accionamiento (drives) con altos torques. La interpretación correcta de las curvas de operación de las bombas es vital para evitar cavitación o arenamiento de líneas.

3. **Grandes Modelos de Lenguaje (LLMs) y sus Limitaciones** Los LLMs, fundamentados en la arquitectura Transformer, han revolucionado el Procesamiento de Lenguaje Natural

(NLP). Modelos como la serie GPT, Llama, o bLLossom (utilizado por Nam et al. en el estudio base) poseen una capacidad semántica profunda. Sin embargo, en el dominio de mantenimiento industrial, presentan limitaciones estructurales conocidas como la «tríada de la inviabilidad»:

- Alucinaciones: La tendencia a generar información plausible pero factualmente incorrecta. En minería, un LLM podría «inventar» un procedimiento de bloqueo de energía basándose en datos generales de internet, lo cual es inaceptable bajo normativas de seguridad.
- Obsolescencia del Conocimiento (Cut-off Date): El conocimiento paramétrico de un LLM es estático. No puede conocer las actualizaciones recientes de un manual de fabricante o los cambios en los PETS (Procedimientos Escritos de Trabajo Seguro) de la mina, que son documentos vivos.
- Falta de Acceso a Datos Propietarios: Los manuales detallados de plantas concentradoras (ej. planos as-built de Cerro Verde o Antamina) son propiedad intelectual privada y no forman parte de los corpus de entrenamiento públicos.

4. **Generación Aumentada por Recuperación (RAG)** RAG es un enfoque de generación que combina un motor de recuperación de información con un modelo de lenguaje para producir respuestas basadas en evidencia externa. El proceso se organiza en tres momentos: preparación del conocimiento, recuperación y generación. En la preparación, los documentos se limpian, se segmentan en fragmentos manejables, se enriquecen con metadatos y se indexan usando representaciones textuales que permiten buscarlos de manera eficiente. En la recuperación, ante una consulta, el sistema localiza los fragmentos más pertinentes mediante búsqueda léxica, búsqueda semántica o una mezcla de ambas, y opcionalmente reordena los resultados con modelos más precisos. En la generación, el modelo de lenguaje redacta una respuesta condicionada por la consulta y por los fragmentos recuperados, manteniendo la trazabilidad hacia las fuentes. RAG se entiende como una arquitectura donde el conocimiento principal se mantiene fuera del modelo y puede actualizarse sin re-entrenamiento, mientras el modelo actúa como redactor que integra y explica la evidencia encontrada.
   - Fase de Ingesta (Indexing): Descomposición de documentos PDF técnicos en fragmentos (chunks). A diferencia del texto plano, los manuales técnicos requieren estrategias de chunking que respeten la estructura del documento (encabezados, tablas), preservando el contexto semántico.29 Estos fragmentos se convierten en vectores densos (embeddings) mediante modelos como BAAI-bge-m3.4
   - Fase de Recuperación (Retrieval): Ante una consulta del usuario (ej. «¿Cuál es el torque de los pernos del revestimiento del Molino SAG?»), el sistema busca en la base de datos vectorial los k fragmentos más similares semánticamente utilizando métricas de distancia (similitud del coseno o producto punto).
   - Fase de Generación (Generation): El LLM recibe un prompt enriquecido que incluye la consulta del usuario y los fragmentos recuperados como «contexto de verdad». El modelo es instruido para responder basándose exclusivamente en este contexto, citando las fuentes.

5. **RAG multimodal** RAG multimodal extiende el principio anterior a fuentes de información heterogéneas como texto, imágenes, diagramas, tablas, audio transcrito y documentos con maquetación compleja. La definición abarca una cadena de procesamiento que inicia con la ingesta y normalización de los datos (incluyendo reconocimiento óptico de caracteres, extracción de tablas y detección de figuras), continúa con la creación de representaciones comparables entre modalidades para poder indexarlas y consultarlas en un espacio común, y culmina con la generación condicionada por evidencias de distinta naturaleza. En este marco, una consulta puede ser textual, visual o mixta; la respuesta puede incluir texto anclado a regiones de una imagen, referencias a celdas de una tabla o a secciones específicas de un documento. El sistema prioriza conservar el contexto estructural del origen, de modo que el usuario pueda verificar rápidamente la procedencia de cada afirmación.
   - Alineación Semántica Manual/Híbrida: (Nam et al., 2025). proponen un mapeo explícito donde las imágenes se vinculan a los párrafos de texto correspondientes antes de la vectorización. Esto asegura que, al recuperar un procedimiento textual, el sistema también recupere la imagen asociada para presentarla al usuario final.
6. **Low-Rank Adaptation (LoRA)** LoRA es una técnica de adaptación eficiente para modelos de lenguaje grandes que incorpora módulos adicionales de bajo costo computacional dentro de capas ya existentes del modelo. En lugar de modificar de manera completa los parámetros preentrenados, LoRA introduce pequeños componentes entrenables que actúan como correcciones y permiten especializar el comportamiento del modelo hacia un dominio, un estilo de respuesta o una tarea concreta. Estos componentes se insertan típicamente en proyecciones de atención y en capas internas del transformador y se entrenan manteniendo congelados los parámetros originales. La definición práctica de LoRA incluye la selección de dónde insertar los módulos, la configuración de su tamaño y la posibilidad de combinar varios adaptadores para distintas tareas sin interferencias, conservando compatibilidad con el flujo normal de inferencia (Hu et al., 2021).
   El vocabulario minero es altamente específico (ej. «chancado», «hidrociclón», «relave»). Los LLMs generalistas pueden no interpretar correctamente estos términos. El reentrenamiento completo (Full Fine-Tuning) es costoso y computacionalmente inviable para muchas operaciones. La técnica LoRA (Low-Rank Adaptation), empleada en el estudio base, congela los pesos del modelo pre-entrenado e introduce matrices de bajo rango entrenables en las capas de atención del Transformer. Esto permite adaptar el modelo al lenguaje técnico y al estilo de respuesta («instruccional») requerido en mantenimiento, modificando menos del 1% de los parámetros totales, lo que facilita su despliegue en infraestructura local (on-premise) típica de faenas mineras.
7. **Agentes basados en LLM** Un agente basado en un modelo de lenguaje es un sistema que percibe un estado del entorno, razona en lenguaje natural para planificar pasos, ejecuta acciones mediante herramientas externas y mantiene memoria a lo largo de múltiples interacciones. Se define por un ciclo continuo: interpretar la situación y el objetivo, decidir la siguiente acción, llamar a una herramienta si es necesario (por ejemplo, un buscador, una base de datos, un extractor de tablas o un ejecutor de código), integrar el resultado al contexto y actualizar su plan. La memoria del agente puede registrar episodios de conversación, hechos persistentes y procedimientos reutilizables. La orquestación suele

expresarse como flujos o grafos de estados que indican cuándo razonar, cuándo recuperar información, cuándo verificar y cuándo responder. De este modo, el agente no solo produce texto, sino que coordina recursos de información y cómputo para alcanzar metas definidas por el usuario.

8. **Métricas y evaluación: ROUGE-L** Es una métrica de comparación entre un texto candidato y una referencia que se basa en la subsecuencia común más larga para estimar cobertura y respeto del orden relativo de las palabras. La idea central es medir cuánto del contenido de la referencia aparece en el candidato manteniendo la secuencia de aparición, aunque no sea de forma contigua. A partir de la longitud de esa subsecuencia se obtienen medidas de cobertura sobre la referencia, precisión sobre el candidato y una combinación de ambas. ROUGE-L se usa ampliamente en tareas de resumen y en evaluación de generación porque captura de forma simple la presencia y el orden de unidades léxicas relevantes, y se puede agregar a nivel de documento o corpus manteniendo un procedimiento de cálculo transparente.

- **Metodologías: DSR, MLOps**.

2.2 Definición de términos

- Glosario de términos y de abreviaturas o siglas

1. Glosario Técnico:

2. Terminología de Inteligencia Artificial y RAG:
   - Embedding (Incrustación Vectorial): Representación matemática de datos (texto o imagen) como vectores en un espacio multidimensional continuo. La proximidad entre vectores indica similitud semántica.
   - Fine-Tuning (Ajuste Fino): Proceso de entrenamiento adicional de un modelo pre--entrenado (Foundation Model) con un conjunto de datos específico del dominio para especializar sus capacidades en una tarea concreta.
   - Hallucination (Alucinación): Fenómeno en el cual un modelo generativo produce contenido que es sintácticamente coherente y seguro, pero factualmente incorrecto o no fundamentado en los datos de entrada.
   - Knowledge Graph (Grafo de Conocimiento): Estructura de datos que representa entidades (nodos) y sus relaciones (aristas) de manera explícita. En RAG avanzado (GraphRAG), se utiliza para capturar la conectividad entre equipos (ej. Bomba A -> alimenta a -> Tanque B) que los embeddings vectoriales pueden perder.
   - LoRA (Low-Rank Adaptation): Técnica de PEFT (Parameter-Efficient Fine-Tuning) que permite adaptar LLMs gigantescos con recursos computacionales limitados, modificando solo matrices de bajo rango inyectadas en la red.
   - RAG (Retrieval-Augmented Generation): Paradigma arquitectónico que mejora la salida de un LLM al proporcionarle información externa recuperada en tiempo de ejecución, combinando la vastedad de conocimiento del modelo con la precisión de datos propietarios.
   - Vector Database (Base de Datos Vectorial): Sistema de almacenamiento optimizado para guardar y consultar vectores de alta dimensión (embeddings). Utiliza algoritmos

de búsqueda de vecinos más cercanos aproximados (ANN) como HNSW para una recuperación ultrarrápida.

3. Métricas de Validación Experimental

   - BERTScore: A diferencia de las métricas tradicionales basadas en n-gramas (que buscan coincidencia exacta de palabras), BERTScore evalúa la similitud semántica utilizando embeddings contextuales.Fundamento: Calcula la similitud del coseno entre los embeddings de cada token en la respuesta generada (x) y los tokens en la respuesta de referencia (y), utilizando una alineación voraz (greedy matching) para maximizar la puntuación.4Relevancia Minera: Es crucial porque en minería existen múltiples formas de referirse a un mismo concepto (ej. «Liner», «Revestimiento», «Blindaje»). Una métrica exacta penalizaría estas variaciones, mientras que BERTScore captura su equivalencia semántica. Nam et al. reportaron una mejora de 3.0 puntos porcentuales en esta métrica usando su arquitectura. Formula:

$$R_{\text{BERT}} = \frac{1}{|x|} \sum_{x_i \in x} \max_{y_j \in y} x_i^T y_j$$

   - ROUGE-L(Recall-Oriented Understudy for Gisting Evaluation):

   Se centra en la estructura y la secuencia, midiendo la subsecuencia común más larga (Longest Common Subsequence - LCS) entre la generación y la referencia.

   Fundamento: Evalúa la capacidad del modelo para preservar el orden de las palabras y la estructura de la oración.

   Fórmula:

$$F_{\text{lcs}} = \frac{(1+\beta^2) R_{\text{lcs}} P_{\text{lcs}}}{R_{\text{lcs}} + \beta^2 P_{\text{lcs}}}$$

   - Evaluación Cualitativa (Human-in-the-loop):

   Validación realizada por expertos del dominio (técnicos/ingenieros) mediante escalas Likert para medir satisfacción, claridad y utilidad.

- Principios de validación experimental y métricas clave.
- Marco normativo (leyes, normas técnicas, reglamentos)

2.3 Estado del Arte

1. Taxonomía de métodos en IA aplicada al Mantenimiento y Documentación La literatura actual (2023-2025) permite clasificar las soluciones de IA para gestión de conocimiento técnico en tres generaciones evolutivas (Gao et al., 2024):

   - Sistemas de Recuperación Basados en Palabras Clave (Lexical Search) :
   - Método: Utilizan algoritmos como TF-IDF o BM25.
   - Aplicación: Motores de búsqueda tradicionales en gestores documentales (DMS).
   - Limitación: No capturan el contexto semántico; fallan ante sinónimos o consultas naturales complejas.

2. Sistemas RAG Naive (Ingenuos) Unimodales (Gao et al., 2024):
    - Método: Indexación vectorial de texto plano + recuperación semántica + generación con LLM genérico.
    - Aplicación: Chatbots de primera generación para manuales simples.
    - Limitación: Sufren del problema de «Lost in the Middle», alucinaciones frecuentes y ceguera ante imágenes/tablas.
3. Sistemas RAG Modulares y Multimodales (MM-RAG) - Enfoque de la Tesis (Knollmeyer et al., 2025):
    - Método: Integran módulos de re-ranking, grafos de conocimiento (GraphRAG) y procesamiento de visión (Vision Encoders) para ingerir texto, imágenes y diagramas conjuntamente.
    - Tecnologías: Modelos de Embeddings densos (ej. BAAI-bge-m3), Bases de Datos Vectoriales, y adaptación de dominio vía PEFT/LoRA.
    - Aplicación: Estado del arte para documentación técnica compleja (automotriz, aeroespacial, y ahora propuesto para minería).

- Revisión comparativa: fortalezas y debilidades.
- Vacíos y oportunidades de investigación.

Capítulo 3: Metodología de Investigación

3.1 Enfoque Metodológico

- Investigación tecnológica aplicada con SDR y MLOps.
- Ciclo de investigación: diseño, implementación, validación, comunicación.

3.2 Diseño Experimental

- Pipeline de datos: fuente de datos, ingesta, preprocesado, feature engineering, entrenamiento.
- Validación: hold-out, K-Fold, bootstrapping, A/B Testing.
- Métricas: F1-score, AUC-ROC, MAE, MSE, métricas UX.

3.3 Interacción con Stakeholders

- Plan de consultas y retroalimentación.

Capítulo 4: Administración del plan de tesis

4.1 Cronograma

4.2 Presupuesto

4.3 Financiamiento

Capítulo 4.3: Anexos

Matriz de consistencia

Capítulo 5: Referencias Bibliográficas

Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, M., & Wang, H. (2024, marzo). *Retrieval-Augmented Generation for Large Language Models: A Survey*. arXiv. https://doi.org/10.48550/arXiv.2312.10997

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. (2021, octubre). *LoRA: Low-Rank Adaptation of Large Language Models*. arXiv. https://doi.org/10.48550/arXiv.2106.09685

Knollmeyer, S., Caymazer, O., & Grossmann, D. (2025). Document GraphRAG: Knowledge Graph Enhanced Retrieval Augmented Generation for Document Question Answering Within the Manufacturing Domain. *Electronics*, *14*(11), 2102. https://doi.org/10.3390/electronics14112102

Nam, Y., Choi, H., Choi, J., & Kwon, H. (2025). LoRA-Tuned Multimodal RAG System for Technical Manual QA: A Case Study on Hyundai Staria. *Applied Sciences*, *15*(15), 8387. https://doi.org/10.3390/app15158387