UNIVERSIDAD NACIONAL DE INGENIERÍA

UNIDAD DE POSTGRADO FACULTAD DE INGENIERIA INDUSTRIAL Y DE SISTEMAS

PLAN DE TESIS

“Sistema de Consultas Multimodal RAG y LoRA Fine-Tuned 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 0: Índice General

DATOS GENERALES

1. **TITULO DEL PLAN DE TESIS:**
   Sistema de Consultas Multimodal RAG y LoRA Fine-Tuned para Manuales de Mantenimiento en Plantas Concentradoras de Cobre
2. **NOMBRE DE AUTOR:**
   Johan Manuel Callomamani Buendia
3. **NOMBRE DEL ASESOR O ASESORES:**
   «…»
4. **AREA INVOLUCRADA: UNIDAD DE POSGRADO FIIS:**
   UNIDAD DE POSTGRADO FACULTAD DE INGENIERIA INDUSTRIAL Y DE SISTEMAS
5. **LUGAR DONDE SE DESARROLLA EL PROYECTO:**
   Unidad Minera Mina Justa - San Juan de Marcona-Nazca
6. **DURACION ESTIMADA:**
   15 MESES (Setiembre 2025 - Noviembre 2026)

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 y no por equipos, procesos o clase de equipo, 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 un mismo proveedor, 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 por 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

¿Puede la implementación de un Sistema de Consultas Multimodal basada en Arquitectura Rag y fine tuning (LoRA) reducir el tiempo de busqueda de información técnica de manuales de de mantenimiento de equipos críticos de una planta concentradora de cobre?

¿De qué manera la implementación de un Sistema de Consultas Multimodal basado en Arquitectura RAG (Retrieval-Augmented Generation) reduce el tiempo de busqueda 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?

1.2.3 Justificación y Alcances

1.2.3.1 Justificación

La presenta investigación se justifica ante la creciente complejidad de la documentación técnica en entornos industriales y de ingeniería, donde la información crítica reside en manuales técnicos con información heterogénea (texto, imagenes, planos, etc.) esta información crítica escencial para toma de decisiones toma tiempo largo de extraer con metodos convencionales y los sistemas RAG(Retrieval-Augmented Generation) multimodales tienen la capacidad de entender y responder de forma rapida reduciendo esta brecha de tiempo entre la información y la toma de decisiones o acciones realies en la empresa industrial. Desde una perspectiva técnica y operativa, el proyecto contribuye en el tiempo de busqueda de información lo cual genera un incremento positivo en el % wrench time del equipo de mantenimiento como lo describe (Palmer, 2019) en su libro el wrench time(tiempo efectivo de mantenimiento) representa en el mejor de los casos del 55% del tiempo total disponible por el equipo técnico de mantenimiento, donde el 45% de tiempo restante corresponde a actividades de traslado, demoras entre sub procesos y **busqueda de información técnica** para ejecutar la actividad de mantenimiento, por lo que el proyecto va a contribuir en la eficiencia y utilización de las HHs del personal técnico de mantenimiento. Desde una perspectiva economica la reducción de tiempo de busqueda de información va a ayudar en una rapida y mejor toma de decisiones lo cual contribuye en una respuesta mas rapida ante emergencias contribuyendo positivamente en la disponibilidad global de las plantas concentradora y con ello la producción asociada.

1.2.3.2 Alcances

El proyecto pretende generar un sistema RAG multimodal y Fine-Tuning (LoRa) para manuales de mantenimiento en planta concentradora con manuales facilitados por el stakeholder el cual pretende tener un anonimato y que la información tenga un estandar de confidencialidad, por lo que el stakeholder va a brindar los manuales que se requieren para este proyecto:
Limitaciones:

- Solo se van a procesar PDF como tipo de información (ingesta de datos).
- Base de conocimiento (corpus) es estricta a los manuales facilidados por el stakehokder.
- El sistema no pretende generar nuevos diagramas, planos o imagenes.
- Base de datos y sistemas va a ser proporsionada por el stakeholder solo en la etapa de producción, la etapa de prototipo se va a realizar un proceso interno de validación de financiamiento.

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

2.1.1 Fundamentos de IA:

El desarrollo de LLMs en el mundo de creación de sistemas capaces de razonar y responder consultas ha ido en constante desarrollo, actualmente se tiene fundamentos donde se desarrollaron la capacidad de incrementar la información de LLMs con el fin de adaptar los estos a nuevos contextos en especifico y con el fin de este proyecto responder a consultas sobre docuemntos técnicos de mantenimiento para ellos surgen conceptos claves como son los siguientes:

1. **Grandes Modelos de Lenguaje (LLMs)**
   Los Grandes Modelos de Lenguaje (LLMs) representan un cambio de paradigma en el procesamiento del lenguaje natural, caracterizados por su capacidad de aprendizaje few-shot (pocos ejemplos) sin necesidad de actualizaciones de gradiente para tareas específicas. (Brown et al., 2020) demuestran que al escalar el tamaño del modelo y los datos de entrenamiento, los modelos desarrollan la capacidad de adaptarse a nuevas tareas simplemente a través de instrucciones textuales o demostraciones en el contexto (in-context learning). A pesar de su capacidad para almacenar conocimiento factual en sus parámetros, los LLMs enfrentan desafíos significativos con el conocimiento de «larga cola» (información que aparece raramente en los datos de entrenamiento). (Kandpal et al., 2023) establecen una relación causal y correlacional entre la capacidad de un modelo para responder preguntas factuales y el número de documentos relevantes

vistos durante el pre-entrenamiento, indicando que incluso modelos masivos luchan por retener información de baja frecuencia.

2. **Generación Aumentada por Recuperación (RAG)**
   La Generación Aumentada por Recuperación (RAG) combina una memoria paramétrica (el modelo seq2seq pre-entrenado) con una memoria no paramétrica (un índice vectorial denso de documentos, como Wikipedia) (Lewis et al., 2021). Este enfoque mitiga las alucinaciones y permite la actualización del conocimiento sin reentrenar el modelo base. (Gao et al., 2024) categorizan la evolución de RAG en tres paradigmas:

   - Naive RAG: El proceso tradicional de indexación, recuperación y generación.
   - Advanced RAG: Introduce estrategias de pre-recuperación y post-recuperación para mejorar la calidad de los documentos seleccionados.
   - Modular RAG: Ofrece mayor adaptabilidad mediante la incorporación de módulos especializados y la reconfiguración del flujo de trabajo.

   Además, (Cheng et al., 2025) destacan la importancia de comprender la dinámica interna entre la recuperación y la generación, proponiendo métodos para rastrear y diagnosticar errores en estos flujos de trabajo opacos.

3. **RAG Multimodal (mRAG)**
   El RAG Multimodal (mRAG) extiende el marco RAG para integrar datos multimodales (texto, imágenes, video) tanto en los procesos de recuperación como de generación (Mei et al., 2025) . A diferencia del RAG tradicional basado solo en texto, mRAG aborda la limitación de aprovechar información rica contenida en formatos no textuales. La evolución de mRAG desde sistemas que convierten datos multimodales a texto (pseudo-MRAG) hasta sistemas end-to-end que preservan los datos multimodales originales y utilizan Modelos de Lenguaje Grande Multimodales (MLLMs) para la generación.

4. **Low-Rank Adaptation (LoRA)**
   La Adaptación de Bajo Rango (LoRA) en el contexto de la eficiencia y el ajuste fino de modelos. LoRA se presenta como una técnica utilizada para la adaptación eficiente de modelos grandes, permitiendo el ajuste fino de LLMs para tareas específicas como la compresión de contextos o la codificación, reduciendo la carga computacional. También se emplea en arquitecturas como «mezcla de expertos» (Mixture of Experts) para mitigar conflictos de datos durante el ajuste de instrucciones en MLLMs. En esencia, permite adaptar modelos masivos a nuevas tareas o dominios sin la necesidad de re-entrenar todos sus parámetros, facilitando la eficiencia en la inferencia y el entrenamiento (Hu et al., 2021)

5. **Agentes basados en LLM**
   Los agentes basados en LLM representan una evolución hacia sistemas autónomos capaces de razonamiento y uso de herramientas. (Zhang et al., 2024) clasifican a los MLLMs en variantes de «Uso de Herramientas» (Tool-using), donde el LLM actúa como un controlador que invoca herramientas externas (como expertos en visión o API de búsqueda) para realizar tareas multimodales, en lugar de realizar todo el procesamiento end-to-end. (Gao et al., 2024) discuten el papel de los agentes dentro del paradigma

«Modular RAG», donde módulos funcionales como búsqueda, memoria y predicción son orquestados dinámicamente para resolver problemas complejos, permitiendo flujos de trabajo iterativos y adaptativos en lugar de lineales.

6. **Multimodal Document Parsing and Indexing**
   Este concepto se refiere al proceso fundamental de procesar documentos multimodales (como PDFs, HTMLs o diapositivas) para hacerlos buscables. El objetivo es analizar y estructurar elementos como texto, imágenes, tablas y videos provenientes de documentos no estructurados o semi-estructurados. Existen dos enfoques principales:
   1. **Basado en extracción:** Extrae información multimodal y la convierte en descripciones textuales o las procesa por separado (por ejemplo, usando OCR para texto y modelos de *captioning* para imágenes).
   2. **Basado en representación:** Utiliza capturas de pantalla de los documentos directamente para la indexación, preservando el diseño visual y la estructura original para evitar la pérdida de información durante la extracción. El resultado final es la creación de un índice vectorial que almacena representaciones de estos datos para su recuperación posterior.

7. **Multimodal Retrieval**
   La recuperación multimodal es el componente encargado de identificar y obtener documentos o fragmentos de información relevantes desde una base de conocimiento externa, utilizando consultas que pueden combinar diferentes modalidades. Este proceso supera la búsqueda de texto simple al permitir búsquedas cruzadas, como encontrar una imagen relevante usando una consulta de texto (text-to-image) o viceversa. La tecnología detrás de esto incluye el uso de «retrievers» (recuperadores) que codifican datos en espacios vectoriales compartidos para medir similitud, y componentes de «reranking» (re-clasificación) que refinan el orden de los resultados basándose en interacciones más profundas entre la consulta y los documentos multimodales recuperados (Mei et al., 2025).

8. **Multimodal Search Planning**
   La planificación de búsqueda multimodal se refiere a las estrategias inteligentes empleadas por los sistemas mRAG para gestionar consultas complejas que requieren información de múltiples fuentes o modalidades. En lugar de seguir una tubería fija, los sistemas avanzados utilizan una planificación adaptativa que descompone una consulta compleja (por ejemplo, una pregunta que requiere razonamiento visual y textual) en sub-tareas. Este módulo decide dinámicamente qué tipo de recuperación realizar (por ejemplo, si buscar una imagen o un texto) y puede reformular la consulta original para mejorar la precisión de la búsqueda, integrando pistas visuales y textuales. Su objetivo es optimizar la adquisición de información, minimizando búsquedas innecesarias y maximizando la relevancia del contenido recuperado (Mei et al., 2025).

9. **Multimodal Generation**
   La generación multimodal es la fase final donde el sistema sintetiza una respuesta coherente integrando la consulta del usuario y la información recuperada, abarcando múltiples modalidades. Gracias a los MLLMs, este proceso no solo produce texto, sino que puede generar respuestas mixtas que entrelazan texto, imágenes, audio y video de manera fluida. Esto permite escenarios donde «una imagen vale más que mil palabras»,

respondiendo directamente con datos visuales, o escenarios donde la inclusión de medios multimodales mejora la precisión y riqueza de una explicación textual (como en guías paso a paso). El sistema debe identificar inteligentemente dónde insertar estos elementos multimodales dentro de la narrativa para asegurar la coherencia y mejorar la experiencia del usuario (Mei et al., 2025).

10. **Métricas y evaluación:**

- **ROUGE-L(Recall-Oriented Understudy for Gisting Evaluation - Longest Common Subsequence)**
  ROUGE es un marco de métricas ampliamente utilizado para evaluar tareas de generación de texto y resumen automático. Aunque la familia ROUGE incluye variantes como ROUGE-N (que mide la superposición de N-gramas), ROUGE-L se utiliza específicamente para evaluar la calidad de la generación basándose en la subsecuencia común más larga entre el texto generado y la referencia. Este enfoque evalúa qué tan bien el texto generado captura el contenido esencial de la referencia, con un fuerte énfasis en la recuperación (recall) de la información (Mei et al., 2025).
- **BERTScore** Definición y Fundamento: BERTScore es una métrica de evaluación que utiliza incrustaciones contextuales (contextual embeddings) provenientes de modelos pre-entrenados como BERT para medir la similitud semántica entre el texto generado y el texto de referencia. A diferencia de las métricas basadas en n-gramas (como ROUGE o BLEU) que dependen de coincidencias exactas de palabras, BERTScore alinea los tokens basándose en sus representaciones vectoriales, lo que le permite capturar relaciones semánticas más profundas y matices de significado que no son evidentes en la superficie léxica (Mei et al., 2025).

2.1.2 Metodologías de investigación tecnológicas

1. **Agile+IA**

   La metodología Agile+IA no es una metodología rígida única, sino la adaptación de los principios del Manifiesto Ágil (comúnmente usando marcos como Scrum o Kanban) a la naturaleza experimental e incierta de los proyectos de Inteligencia Artificial. A diferencia del desarrollo de software tradicional (que es determinista), la IA es probabilística y requiere mucha experimentación, lo que obliga a adaptar el enfoque ágil de la siguiente manera:

   - Principio 1: Entrega Temprana de Valor En lugar de esperar meses para tener un «modelo perfecto», el objetivo es reducir el Time-to-Market mediante incrementos funcionales.
   - Principio 2: Feedback Continuo de Stakeholders La validación en IA es más compleja que en software tradicional porque los modelos pueden fallar de formas impredecibles (alucinaciones, sesgos).
   - Principio 3: Priorizar Desempeño y Calidad El Backlog del producto debe tratar los requisitos no funcionales (rendimiento del modelo) con la misma urgencia que las nuevas funcionalidades.

- Principio 4: Colaboración Multidisciplinaria Se incentiva la participación de varias tipos de diciplinas Data scientist, dev, DevOps y experto de dominio trabajan codo a codo en cada iteración.

2. **CRISP-DM** Es el estándar más utilizado en la industria para proyectos de minería de datos y ciencia de datos. A diferencia de Agile (que es una metodología de gestión de proyectos), CRISP-DM es un modelo de proceso específico para el ciclo de vida de los datos. Se estructura en 6 fases secuenciales pero iterativas, lo que significa que es común retroceder entre fases (por ejemplo, volver a la preparación de datos tras un modelado fallido). Las 6 Fases de CRISP-DM:

   1. Comprensión del Negocio (Business Understanding):
      Es la fase crítica donde se definen los objetivos del proyecto desde la perspectiva empresarial y se traducen en problemas técnicos de minería de datos. Se establecen los criterios de éxito.
   2. Comprensión de los Datos (Data Understanding):
      Implica la recolección inicial, descripción y exploración de los datos. Se busca identificar problemas de calidad (datos faltantes, errores) y descubrir primeros patrones o hipótesis.
   3. Preparación de los Datos (Data Preparation):
      Generalmente es la fase que consume más tiempo (60-80% del proyecto). Incluye la limpieza, selección, integración y transformación de datos (Feature Engineering) para crear el conjunto final que se usará en el modelado.
   4. Modelado (Modeling):
      Se seleccionan y aplican las técnicas de modelado (algoritmos de IA/ML). Se calibran los parámetros para optimizar resultados. A menudo se requiere volver a la fase de preparación para ajustar los datos a las necesidades del modelo específico.
   5. Evaluación (Evaluation):
      No solo se evalúa la precisión técnica del modelo, sino su eficacia para resolver el problema de negocio planteado en la fase 1. Se decide si el modelo es apto para ser desplegado o si requiere revisión.
   6. Despliegue / Implementación (Deployment):
      El conocimiento obtenido se presenta de forma útil para el usuario final. Puede ir desde generar un reporte simple hasta la implementación de un modelo predictivo en tiempo real integrado en una aplicación de software. Incluye planes de monitoreo y mantenimiento.

2.2 Definición de términos

1. Glosario de términos y de abreviaturas o siglas
   - 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.
- IOM (Manuales de Instalación, Operación y Mantenimiento): Es el documento técnico rector proporcionado por el fabricante (OEM) o desarrollado internamente, que contiene las especificaciones, procedimientos e intervalos necesarios para conservar la función de un activo.
- Búsqueda de Información Secuencial: Es el método de recuperación de información que sigue un orden lineal o cronológico preestablecido en el documento.
- Búsqueda de Información por Palabra Clave: Método de recuperación basado en la coincidencia léxica exacta (keyword matching) de términos específicos dentro de un índice o documento digital.
- Búsqueda por Tabla de Contenido: Método de navegación jerárquica que utiliza la estructura lógica del documento (Capítulos, Secciones, Subsecciones) para localizar información.
- Despiece de Partes de Equipos: Es una representación técnica gráfica (plano o diagrama) que muestra los componentes de un ensamblaje ligeramente separados por una distancia, indicando su orden de montaje y relación espacial.
- Procedimientos de Actividad de Mantenimiento de Equipo: Son documentos instruccionales estandarizados que describen paso a paso cómo ejecutar una tarea específica de mantenimiento para asegurar consistencia, seguridad y calidad.
- Documento Técnico: Término paraguas que engloba toda la información escrita o gráfica que describe la funcionalidad, arquitectura y manejo de un producto técnico o sistema.
- Sistema Pregunta-Respuesta (Question-Answering System): Es una aplicación de Inteligencia Artificial diseñada para responder automáticamente a preguntas formuladas por humanos en lenguaje natural.
- Maintenance Domain (Dominio de mantenimiento): Se refiere al campo de conocimiento especializado y al contexto operativo relacionado con la gestión, preservación

y restauración de activos físicos industriales. Se distingue por un vocabulario técnico altamente específico (ej. «cavitación», «holgura», «termografía»), una baja tolerancia al error (por riesgos de seguridad) y una gran dependencia de datos multimodales (sonidos de vibración, imágenes de desgaste, manuales PDF).

2. Principios de validación experimental y métricas clave

   - 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.

2.3 Estado del Arte

2.3.1 Taxonomía de métodos en IA aplicada

La Generación Aumentada por Recuperación (RAG) ha emergido como un paradigma fundamental para mitigar las alucinaciones y limitaciones de conocimiento en los Grandes Modelos de Lenguaje (LLMs). La investigación reciente ha trascendido las implementaciones «ingenuas» (Naive RAG), avanzando hacia arquitecturas especializadas que integran datos multimodales, estructuras de grafos de conocimiento y estrategias de recuperación híbrida para satisfacer las demandas de precisión en sectores industriales, automotrices y médicos.

La implementación de RAG en entornos industriales enfrenta desafíos únicos debido a la

complejidad de la documentación técnica y la necesidad de precisión operativa. En el sector automotriz, (Nam et al., 2025) desarrollaron un sistema RAG multimodal adaptado al dominio para manuales técnicos de vehículos (caso de estudio Hyundai Staria). Su enfoque utiliza el ajuste fino eficiente en parámetros (LoRA) sobre el modelo bLLossom-8B y embeddings BAAI-bge-m3, logrando mejoras significativas en métricas como BERTScore y ROUGE-L al integrar texto e imágenes para escenarios de resolución de problemas. Simultáneamente, (Knollmeyer et al., 2025a) abordaron la gestión de conocimientos en la planificación de producción (caso de estudio Audi). Identificaron que los modelos de embedding multilingües estándar tienen un rendimiento inferior en documentos técnicos en alemán en comparación con el inglés. Su solución propuesta es un enfoque de recuperación híbrida que combina búsqueda vectorial densa con búsqueda de texto completo, mejorando la precisión de recuperación en un 20% para documentos en alemán.

En el ámbito de la ingeniería de software y ferroviaria, (Ibtasham et al., 2025) propusieron «ReqRAG», un chatbot diseñado para la gestión de lanzamientos de software en Alstom. Este sistema utiliza documentos técnicos (notas de lanzamiento, arquitectura) para responder consultas sobre trazabilidad de requisitos, demostrando que el 70% de las respuestas generadas fueron consideradas adecuadas y útiles por expertos industriales.

En la minería, (Shu et al., 2024) presentaron un marco para la construcción de Grafos de Conocimiento Hiper-Relacionales destinados al análisis de fallas en montacargas de minas. Su metodología utiliza LLMs (GPT) para extraer entidades y relaciones complejas, optimizando los datos mediante predicción de enlaces para superar la escasez de datos en manuales de mantenimiento.

La literatura actual destaca que la recuperación puramente vectorial es insuficiente para capturar matices semánticos específicos o terminología exacta en dominios especializados.

(Santra et al., 2025) introdujeron el concepto de «LU-RAG», un enfoque híbrido que combina el aprendizaje en contexto (ICL) utilizando datos etiquetados con RAG basado en datos no etiquetados. Su metodología recalcula dinámicamente las puntuaciones de instancias etiquetadas y pasajes no etiquetados, demostrando que esta fusión supera a los enfoques aislados en tareas de verificación de hechos y clasificación de sentimientos.

En el dominio médico, (Wang et al., 2025) propusieron una optimización basada en RAG para la comprensión y razonamiento de conocimiento médico. Su enfoque innovador incluye una fusión adaptativa de recuperadores dispersos (TF-IDF) y densos (Transformer), junto con ingeniería de prompts y limpieza de datos rigurosa, para mitigar alucinaciones y mejorar la precisión en el conjunto de datos CCKS-TCMBench.

Esta tendencia hacia la hibridación también es respaldada por (Knollmeyer et al., 2025a), quienes validaron que un enfoque equilibrado (30/70) entre búsqueda vectorial y búsqueda de texto completo ofrece la recuperación más robusta para corpus técnicos bilingües, superando las limitaciones de los modelos de embedding en lenguajes específicos.

Para superar las limitaciones de razonamiento multi-salto (multi-hop) y fragmentación de contexto en RAG convencional, la integración de Grafos de Conocimiento (KGs) ha ganado tracción.

(Knollmeyer et al., 2025b) introdujeron «Document GraphRAG», un marco que estructura documentos técnicos en un Grafo de Conocimiento de Documentos (DKG). Este sistema preserva la estructura jerárquica (capítulos, secciones) y utiliza enlaces semánticos basados en palabras clave para mejorar la recuperación. La evaluación demostró que GraphRAG supera a las líneas base de RAG ingenuo, especialmente en preguntas que requieren razonamiento complejo a través de múltiples documentos. Similarmente, el trabajo de Shu et al. con grafos hiper-relacionales permite capturar relaciones multidimensionales en diagnósticos de fallas, lo cual es superior a las representaciones de relaciones binarias tradicionales.

La capacidad de procesar información más allá del texto es crucial para la aplicabilidad en el mundo real. Además del trabajo de Nam et al. con texto e imágenes, (Drushchak et al., 2025) propusieron un sistema «mRAG» unificado capaz de procesar texto, tablas, imágenes y video. Su tubería (pipeline) ingesta datos utilizando herramientas como Amazon Textract y Transcribe, y emplea LLMs multimodales (Claude 3.5 Sonnet) para el enriquecimiento semántico. Aunque el rendimiento en consultas de video e imagen aún presenta desafíos comparado con texto y tablas, su marco demuestra la viabilidad de pipelines unificados para datos diversos.

El estado del arte actual en RAG se caracteriza por un alejamiento de soluciones genéricas hacia arquitecturas altamente especializadas. La evidencia sugiere que la combinación de recuperación híbrida, la estructuración mediante grafos de conocimiento y la integración multimodal como se ve en la taxonomía de la Figura 1 son estrategias esenciales para desplegar sistemas de QA confiables en entornos críticos como el mantenimiento de equipos de mineria de plantas concentradoras.

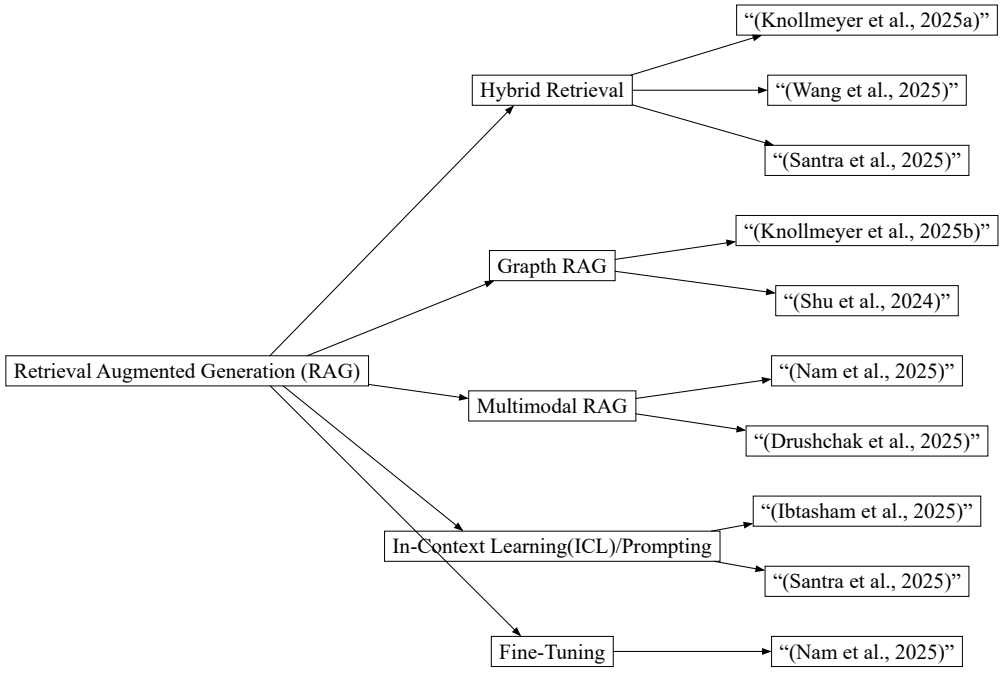

Figura 1: Taxonomía por tipo de RAGs de lecturas revisadas

2.3.2 Revisión comparativa: fortalezas y debilidades.

1. **Sistema RAG Multimodal con Ajuste Fino (LoRA)**
   (Nam et al., 2025) presentan una solución robusta para la industria automotriz, específi-

camente para la gestión de manuales técnicos del vehículo Hyundai Staria. La técnica empleada combina un sistema RAG multimodal con el ajuste fino eficiente de parámetros (PEFT) mediante LoRA (Low-Rank Adaptation) sobre el modelo de lenguaje bLLossom-8B y el modelo de embedding BAAI-bge-m32. El dominio es estrictamente técnico-automotriz, utilizando datos extraídos de manuales en PDF que contienen tanto texto como diagramas.

- **Resultados y Conclusiones:** El sistema logró mejoras notables frente a líneas base, con un incremento del 18.0% en ROUGE-L(27.12%-9.09%) y un 3.0% en similitud de coseno(78.11%-75.81%), destacando en la precisión de respuestas guiadas por imágenes. Los autores concluyen que la integración multimodal es esencial para la resolución de problemas técnicos complejos.
- **Fortaleza:** Su mayor virtud es la precisión procedimental, al integrar anotaciones de similitud a nivel de oración y pares texto-imagen, el sistema ofrece instrucciones paso a paso muy superiores a la búsqueda simple.
- **Debilidad:** La escalabilidad es su principal limitación. El mapeo imagen-texto se realizó manualmente (solo 200 pares), lo que representa un cuello de botella significativo para expandir el sistema a bibliotecas de manuales más extensas sin automatización avanzada.

2. **Recuperación Híbrida para Documentos en Alemán**

   (Knollmeyer et al., 2025a) abordan un problema lingüístico y técnico en la planificación de producción automotriz (caso Audi). La técnica empleada es una recuperación híbrida que fusiona la búsqueda vectorial densa (usando modelos multilingües como Cohere y Titan) con la búsqueda de texto completo dispersa (BM25F[1] /TF-IDF[2]). Los datos consisten en normas industriales (VDA) bilingües, permitiendo una comparación directa entre el rendimiento en inglés y alemán.

   - **Resultados y Conclusiones:** El estudio demuestra que los modelos de embedding multilingües tienen un rendimiento inferior en textos técnicos en alemán. El enfoque híbrido mejoró la precisión de recuperación en un 10% para documentos en alemán, igualando el rendimiento obtenido en documentos en inglés (English dataset Precisión: 78.9%, MMR: 0.64; German dataset Precisión: 79.3%, MMR: 0.64).
   - **Fortaleza:** Su pragmatismo y eficiencia destacan al resolver la deficiencia de los modelos de embedding en idiomas distintos al inglés mediante la reincorporación de algoritmos clásicos (BM25), logrando una solución robusta sin el alto costo computacional del ajuste fino.
   - **Debilidad:** El estudio depende de pares de preguntas y respuestas generados por LLMs para la evaluación, lo que introduce un sesgo potencial donde las preguntas podrían ser «demasiado fáciles» o artificiales, no reflejando completamente la complejidad de la consulta humana real.

3. **ReqRAG: Gestión de Lanzamientos de Software**

   (Ibtasham et al., 2025) (Fecha inferida: 2025) proponen «ReqRAG» para el sector ferroviario en Alstom. La técnica empleada es un pipeline RAG especializado que

[1](Rajaraman & Ullman, 2011)
[2](Robertson & Zaragoza, 2009)

utiliza modelos OCR (YOLOX, Detectron2) para la extracción de datos de PDFs complejos y compara varios LLMs (Phi-3, Llama-3.2) para la generación. El dominio es la gestión de requisitos y lanzamientos de software, utilizando datos de documentos de arquitectura y notas de lanzamiento.

- **Resultados y Conclusiones:** La evaluación humana indicó que el 70% de las respuestas cumplieron el críterio adequacy avg=3.69, usefulness avg=3.44, Relevance avg=3.32. Se concluye que la combinación de embeddings "stella_v5" con OCR Detectron2 ofrece una buena recuperación.
- **Fortaleza:** La validación industrial real es su punto fuerte; al incluir una evaluación cualitativa con expertos de dominio, el estudio trasciende las métricas sintéticas y prueba su utilidad en un entorno crítico cercano al real.
- **Debilidad:** El volumen de datos es bajo (solo 7 documentos técnicos), lo que plantea dudas sobre la generalización de los resultados a repositorios documentales masivos. Además, el uso de modelos de código abierto plantea interrogantes sobre la privacidad de datos propietarios.

4. **Grafos de Conocimiento Hiper-Relacionales en Minería**
   (Shu et al., 2024) se enfocan en el análisis de fallas para sistemas de izaje para minas. La técnica empleada es la construcción de Grafos de Conocimiento Hiper-Relacionales (HKG) utilizando LLMs (GPT) para la extracción de tuplas complejas y algoritmos de predicción de enlaces para completar datos faltantes. El dominio es el mantenimiento de sistema de izajes mineros, utilizando datos de informes de inspección y registros de mantenimiento.
   - **Resultados y Conclusiones:** El modelo optimizado (MHSD) logró una mejora de 0.008 en la métrica MRR (Mean Reciprocal Rank) sobre datos no optimizados y superó al modelo KICGPT. Se concluye que la representación hiper-relacional captura mejor la complejidad de las fallas mecánicas de sistemas de izaje mineros.
   - **Fortaleza:** La capacidad de modelado complejo es superior; al utilizar grafos hiper-relacionales, el sistema puede representar matices (condiciones, causas, consecuencias) que se pierden en los grafos de conocimiento binarios tradicionales.
   - **Debilidad:** Existe una fuerte dependencia de la ingeniería de prompts para la generación de datos y la extracción de relaciones, lo que puede introducir inconsistencias si el modelo generativo alucina o malinterpreta la terminología técnica sin una supervisión estricta.
5. **Optimización RAG para Conocimiento Médico**
   (Wang et al., 2025) proponen una optimización algorítmica para el sector salud. La técnica empleada incluye una fusión adaptativa de recuperadores dispersos (TF-IDF) y densos (Transformer), junto con una rigurosa limpieza de datos e ingeniería de prompts. El dominio es la medicina clínica y el razonamiento semántico, utilizando datos del benchmark CCKS-TCMBench (exámenes médicos y casos clínicos).
   - **Resultados y Conclusiones:** El modelo optimizado superó a las líneas base (incluyendo GPT-4 y ChatGLM3) en métricas de precisión y razonamiento, con un aumento promedio del 3.86% en métricas integrales. Se concluye que la fusión de recuperadores y la limpieza de datos son críticas para reducir alucinaciones médicas.

- **Fortaleza:** La robustez metodológica en el preprocesamiento y la fusión de recuperadores permite mitigar el riesgo de alucinaciones, un aspecto crítico y no negociable en aplicaciones médicas.
- **Debilidad:** La evaluación se basa principalmente en datos de competiciones, que aunque estandarizados, pueden no capturar la diversidad de los datos clínicos del mundo real (historias clínicas no estructuradas).

6. **Document GraphRAG en Manufactura** (Knollmeyer et al., 2025b) introducen «Document GraphRAG» para entornos de manufactura. La técnica empleada estructura documentos en un Grafo de Conocimiento de Documentos (DKG) que preserva la jerarquía (capítulos, secciones) y utiliza enlaces basados en palabras clave para la recuperación. El dominio es la manufactura automotriz, evaluado con datos públicos (SQUAD, HotpotQA) y un conjunto de datos interno de planificación de producción.
   - **Resultados y Conclusiones:** GraphRAG superó consistentemente a RAG ingenuo en métricas de recuperación y generación, mostrando beneficios notables en preguntas de razonamiento multi-salto (multi-hop). Se concluye que preservar la estructura del documento es vital para consultas complejas.
   - **Fortaleza:** Su capacidad para el razonamiento estructural; al mapear explícitamente la jerarquía del documento en el grafo, el sistema puede responder preguntas complejas que requieren entender el contexto de secciones completas, no solo fragmentos aislados.
   - **Debilidad:** El costo de latencia es alto; el tiempo de respuesta total fue aproximadamente 5 veces mayor que el de un sistema naive RAG (9.8s vs 1.7s), lo que podría afectar la experiencia de usuario en aplicaciones de tiempo real.
7. **LU-RAG: Fusión de Datos Etiquetados y No Etiquetados**
   (Santra et al., 2025) presentan un enfoque teórico-práctico denominado LU-RAG. La técnica empleada es un marco híbrido que combina el aprendizaje en contexto (ICL) usando ejemplos etiquetados con RAG basado en documentos no etiquetados, utilizando una combinación lineal de puntuaciones. El dominio es general (NLP), aplicado a tareas de verificación de hechos y clasificación de sentimientos, usando datos de FEVER y SST.
   - **Resultados y Conclusiones:** LU-RAG superó tanto a ICL puro como a RAG puro, logrando, por ejemplo, una mejora del 19.45% en F1-score para verificación de hechos frente a líneas base supervisadas. Se concluye que equilibrar datos etiquetados y no etiquetados ofrece «lo mejor de ambos mundos».
   - **Fortaleza:** La innovación algorítmica al fusionar paradigmas; demuestra que la combinación dinámica de ejemplos de entrenamiento (few-shot) con conocimiento externo (retrieval) es superior a usarlos por separado.
   - **Debilidad:** La sensibilidad de hiperparámetros; el rendimiento depende crucialmente del parámetro α (proporción de mezcla), que varía según la tarea y es difícil de generalizar sin un ajuste fino específico para cada nuevo conjunto de datos.
8. **RAG Multimodal Unificado (mRAG)**
   (Drushchak et al., 2025) proponen un sistema unificado para procesar múltiples tipos de medios. La técnica empleada es un pipeline «mRAG» que ingesta texto, tablas, imágenes y video utilizando servicios de AWS (Textract, Transcribe) y LLMs multi-

modales (Claude 3.5 Sonnet) para generar descripciones semánticas45. El dominio es el soporte técnico de servidores (manuales Dell), usando datos de PDFs y videos instruccionales.

- **Resultados y Conclusiones:** El sistema mejoró la relevancia contextual y redujo alucinaciones. Sin embargo, el rendimiento en consultas de video e imagen fue inferior al de texto y tablas. Se concluye que un pipeline unificado es viable pero requiere mejoras en la interpretación de datos visuales no estructurados.
- **Fortaleza:** La arquitectura unificada; es uno de los pocos enfoques que integra nativamente el video como una modalidad de recuperación junto con texto y tablas, abriendo la puerta a asistentes técnicos verdaderamente completos.
- **Debilidad:** El rendimiento desigual entre modalidades. La precisión en la recuperación de video e imagen sigue siendo baja comparada con el texto, lo que limita su fiabilidad en escenarios donde la información visual es crítica.

2.3.3 Vacíos y oportunidades de investigación.

En general el principal vacío de los papers revisados son información concreta de la latencia de sus propuestas, no se explican el tiempo que le toma a sus sistemas RAG en responder, a excepción de (Knollmeyer et al., 2025b) quien menciona esta limitante. Otro vacío es la falta de técnicas Parameter-Efficient Fine-Tuning (PEFT) el cual solo en el paper (Nam et al., 2025) se aplica y explica, el resto de paper no tocan esta técnica por lo que no se puede comparar o afirmar que es efectiva en mas tipos de data o dominio.

Tabla 1: Comparativa del Estado del Arte en Sistemas RAG y Multimodales

Trabajo	Método	Datos/Dominio	Métrica clave	Fortalezas	Debilidades
(Nam et al., 2025) Hyundai	mRAG+LoRA Fine Tuning (0.1%) Embeddings: BAAI-bge-m3 LLM: bLLossom-8B + LoRA	Manuales Hyundai Staria PDF. simple QA dataset, multi-turn QA dataset, RAG QA dataset	ROUGE-L: 27.12% BERT: 78.11% Encuesta: 4.4/5	• Integra texto e imágenes. • Similitud semántica en el dominio automotriz.	• Escalabilidad limitada por las anotaciones manuales. • Dominio restringido al mantenimiento automotriz. • No aplica técnicas de optimización.
(Knollmeyer et al., 2025a) (2025) hybrid	Hybrid Retrieval 30/70 (búsqueda vectorial Amazon Titan + búsqueda texto completo BM25F/ TF-IDF) Embeddings:	18 normas y estándares de VDA. Corpus idénticos en alemán e inglés. Data sintética QA (Claude Sonnet 3.5)	Precisión: 0.79 MMR: 0.64	• Dominio en idioma alemán. • Técnica escalable y simple. • Alta precisión.	• Se centra únicamente en texto. • Evaluación con datos sintéticos. • Falta de análisis de latencia.

Trabajo	Método	Datos/Dominio	Métrica clave	Fortalezas	Debilidades
	Cohere M-3.5 multilingual				
(Ibtasham et al., 2025) Software	ReqRAG OCR (YOLOX) Embeddings: mxbai-embeb--large-v1 stella_en_400M_v5	Gestión de Releases de Software Ferroviario (Alstom). 7 docs técnicos (TCMS), 27 queries reales.	R@3 (Recall): 0.90 Adequacy (Humana): 3.69/5	• Capacidad de procesar tablas y diagramas. • Alta precisión en el dominio industrial.	• Dataset de evaluación pequeño. • Dependencia crítica del OCR. • Ventana de contexto limitada para documentos extensos.
(Shu et al., 2024) Knowledge Graph	LLM-HKGCF (GPT o1 preview) Embeddings: UltraFastBERT	MHSD (Mine Hoist System Dataset) Open Dataset (JF17K, WD50K)	MRR: 0.715	• Alta automatización en extracción de conocimiento. • Representación de contextos multidimensionales. • Reducción de costos de construcción manual.	• Problemas de interpretabilidad de fallas. • Uso de LLMs costosos. • Limitado al dominio del dataset. • No soporta tablas/figuras.
(Wang et al., 2025) Medical	RAG optimizado (sparse TF-IDF y dense retrievers basado en Transformers)	Examen de licenciatura médica de China (9,788 preguntas examen, 5,473 práctica)	Comprensión: Precisión: 0.83 Rouge-L: 0.23 BertScore: 0.71 Razonamiento: Precisión: 0.91	• Reducción de alucinaciones. • Eficiencia de recursos (no fine-tuning). • **Trazabilidad:** permite rastrear la fuente original.	• Representatividad de los datos. • Falta de validación clínica en entornos reales. • Complejidad en resolución de ambigüedades.
(Knollmeyer et al., 2025b) Graph RAG	**Document GraphRAG** Estrategias búsqueda: 1. ICS 2. IKS 3. UKS	(AUDI AG) 17 documentos técnicos, +5,500 páginas. Planificación y estándares en alemán.	Recuperación: Recall@700: 0.79 MRR@1300: 0.63 Genera-	• Escalabilidad en creación de grafos. • Robustez a pérdida de rendimiento.	• Costoso (APIs externas). • Latencia elevada. • Susceptible a ruido por redundancia.

Trabajo	Método	Datos/Dominio	Métrica clave	Fortalezas	Debilidades
			ción: Faithfulness: 0.94	• Calidad de respuestas. • Pipeline modular.	• No soporta tablas/imágenes.
(Liu et al., 2025) text--image	**HM-RAG** (Vectorial, grafos y web)	ScienceQ CrisisMMD	Accuracy: 93.73	• Modularidad y escalabilidad. • Soporta datos heterogéneos. • Reducción de alucinaciones.	• Latencia en búsqueda web. • Dependencia APIs externas. • Complejidad de infraestructura. • Costo computacional alto.
(Drushchak et al., 2025) table, image	mRAG (Multimodal RAG)	36 documentos servidores Dell. 82 manuales video. 116 preguntas (texto, tablas, imágenes, videos)	Contextual Precision: 0.35 Contextual Recall: 0.69	• Procesa 4 modalidades en un pipeline. • Buena calidad en tablas y texto estructurado.	• Bajo rendimiento en video. • Datos no estructurados difíciles. • Dependencia de AWS.

Capítulo 3: Metodología de Investigación

3.1 Enfoque Metodológico

- El proyecto sigue una metodología **Agile+IA & CRISP-DM** combinada con el fin de poder adaptarse a las exigencias del stakeholder que solicita avances continuos y ser robusto en la etapa del procesamiento de la datos hereogeneos texto, imagenes, tablas de los Manuales tecnicos de mantenimiento. La metodología CRISP-DM (Cross-Industry Standard Process for Data Mining) contribuye en reforzar la parte de comprensión tanto del negocio como de los datos y su procesamiento, cabe mencionar que la metodología SEMMA no se considero ya que su etapa de muestreo puede ser contra producente ya que la data (pdfs) tienen estructuras diversas las cuales tienen que ser tomadas en su totalidad desde el principio y así tener comparaciones reales.

- **Agile+IA**

Como metodologia plantea los siguientes sprints para el desarrollo, implementación del proyecto. Ciclo de investigación:

- **Sprint 0 - Revisión de papers y baselane - Comprensión del negocio (CRISP-DM):** Como parte inicial se va a revisar literatura relacionada con papers relevantes y analizarlos de forma que podamos entender el problema y las propuestas de solución entendiendo tanto

las fortalizas, debilidades y oportunidades de mejora que se tienen en los papers, como principal entregable se tienen los Controles de Lectura que detallan el análisis realizado.

- Objetivo:
  - Desarrollar el estado del arte del proyecto.
  - Elaboración de Controles de Lecturas como apoyo para la elaboración del estado del arte.
  - Se va a comenzar a reunir con el stakeholer con el fin de entender y adecuar el estado del arte a la necesidad o problema a resolver.

- **Sprint 1 - Preparación de datos(CRISP-DM):** Se va a desarrollar un pipeline de ingesta de los manuales técnicos, estos van a ser procesados y analisados con una herramienta que tenga la capacidad de procesar texto, tablas e imagenes: Objetivos:
  - Extracción de texto: se va a emplear la libería de python PyMuPDF, como se uso en el paper (Nam et al., 2025) quien logro buenos resultados.
  - Extracción de tablas e imagenes, en este caso se tiene la propuesta de usar Gemini-2.5--flash siendo una variante a la propuesta en el paper (Drushchak et al., 2025), quien utiliza CLaude 3.5 Sonnet.
  - Embedding and Indexing: esta sub etapa consiste en vectorizar tanto el texto, tablas e imagenes y almacenarlo en un gestor de base de datos de embeddings, metadata y almacenamiento de documentos, como propuesta se tiene a Chroma como primera opción.
  - EDA Rápido: estadistivas y visualizaciones de la calidad de los documentos técnicos y las técnicas de extracción propuestas.

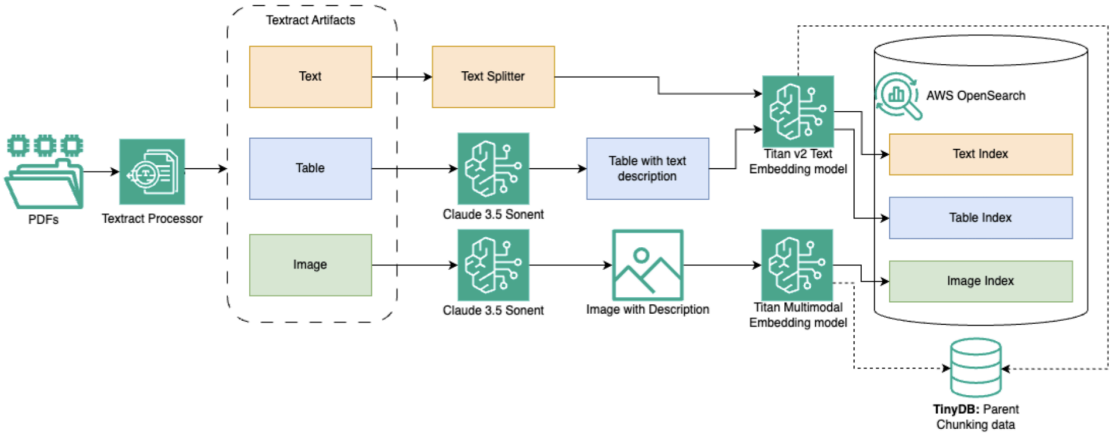

Figura 2: Pipeline de extracción de información de PDFs (Drushchak et al., 2025)

- **Sprint 2 - Modelado(CRISP-DM):** En esta etapa pretendemos implementar un naive RAG (RAG básico) como se propone en el estado del arte, con el cual se pretende tener un punto de partida con el cual se pueda demostrar la mejora en cada iteración, y poder realizar los primeros analisis y comenzar con la iteración de pruebas con indicadores o metricas propuestas como son ROUGE-L y BertScore. Objetivos:
  - naive RAG: crear el chat inteligente con RAG de manuales técnicos, se propone utilizar LangChain como plataforma para desarrollar el RAG.
  - EDA Rápido: metricas del naive RAG y registro de resultados.

- **Sprint 3 - Generación de Demo:** En esta etapa pretendemos demostrar el naive RAG y generar una plataforma donde el stakeholder y usuarios puedan usar el artefacto y brindar feedback. Objetivo:
  - Creación de plataforma de interacción 1 a 1 (1 usuario a la vez con el artefacto), se propone implementar con el paquete de python Gradio por sus ventajas y facilidad para el prototipado con fines de chats inteligentes.
  - Informe con feedback (EDA) de usuarios seleccionados por el stakeholder.
- **Sprint 4 - Integración multimodal del RAG:** Se va a tener un spring para la integración de los 3 tipos de datos o datasets (texto, tablas e imagenes), para ellos se va a usar la base de datos Chroma integrada con LangChain y poder cumplir con este objetivo..
  - Creación de Multimodal RAG he implementación en la plataforma Gradio.
  - Informe con feedback de usurios (EDA).
- **Sprint 5 - Fine tuning del modelo base (aplicación de LoRA):** En este spring se pretende realizar el fine tuning de modelo base y lograr un modelo con el dominio de mantenimiento de equipos de plantas concentradoras. Objetivo:
  - Creación de modelo fine-tuned: Para lograr esta implementación se apoya en la librearía PEFT de huggingface.
  - Documento Feedback del nuevo modelo y sus resultados.
  - Integración del feedback desde la plataforma Gradio.
- **Sprint 6 - Revisión de feedback y refinamiento del sistema:** Ene sta etapa se tiene como objetivo levantar observaciones respecto al sistema integrado y poder mejorar factores claves de las metricas elegidas. Obejtivos:
  - Generar un sistema mRAG Fined-tuned con metricas esperadas por el stake holder.

diseño, implementación, validación, comunicación.

3.2 Diseño Experimental

1. **Pipeline de datos**:
   - Fuente de datos: PDF de manuales de mantenimientos de equipos críticos
   - Ingesta: Se va a implementar 3 tipo de procesamiento y generación de 3 tipos de dataset:
     1. PDF a texto: Dataset de QA texto-texto, con PyMuPDF.
     2. PDF a imagen-texto: Dataset de texto-imagen, con un llm(Gemini) para generar tanto los bounding boxes y la descripción semantica de la imagen.
     3. PDF a tablas-texto: Dataset de tablas-tablas, con un llm(Gemini) para generar tanto los bounding boxes y la descripción semantica de la tabla, tambien se va a extraer información de la tabla con PyMuPDF/OCR (alternativa a evaluar).
   - Preprocesado y Feature engineering:
     1. Con la herramienta LangChain se va a realizar el enriquecimiento semantico y extracción de palabras claves mediante Gemini como se detalla en el punto de ingesta.
     2. Se va a generar la base de datos vectorial y de documentos con ChromaDB donde se va a almacenar los embeddings.

- Entrenamiento. Se va a implementar PEFT(Parameter-Efficient Fine-Tuning) para dotar o enriqueser el dominio en mantenimiento de equipos de planta concentradora mediante LoRA utilizando librerías como Hugging Face PEFT para adaptar LLMs.

2. **Validación**:
   - Se va a implementar el framwork RAGAS para poder realizar una evaluación automatizada como es el caso de (Knollmeyer et al., 2025a).

3. **Métricas**:
   - Se va a usar LangSmith libreria de langchain para poder aplicar las métricas ROUGE-L y BERTScore.

3.3 Interacción con Stakeholders

- Plan de consultas y retroalimentación a partir del spring 3, .

Capítulo 4: Administración del plan de tesis

4.1 Cronograma

		2025		2026											
Item	Actividades (Springs)	Nov	Dic	Ene	Feb	Mar	Abr	May	Jun	Jul	Ago	Set	Oct	Nov	Dic
0	Revisión de papers y baseline - Comprensión del negocio (CRISP-DM)	X	X												
1	Preparación de datos(CRISP-DM)			X	X	X	X								
2	Modelado(CRISP-DM)						X	X	X						
3	Generación de Demo								X	X					
4	Integración multimodal del RAG									X	X	X			
5	Fine tuning del modelo base (aplicación de LoRA)											X	X		
6	Revisión de feedback y refinamiento del sistema								X	X	X	X	X	X	X

Figura 3: Cronograma de implementación de proyecto

4.2 Presupuesto

Item	Descripción	Fórmula / Cálculo (Estimado)	Costo Total (USD)
1. Recursos Computacionales (Infraestructura)	**Instancia GPU para Entrenamiento (PEFT/LoRA)** Servidor con GPU (ej. NVIDIA A10G o L4) para realizar el ajuste fino eficiente (LoRA) de los modelos.	1 instancia x $1.50/hr x 300 horas (entrenamiento + pruebas)	$1,500.00
	Servidor CPU/RAM alta para alojar ChromaDB, ejecutar PyMuPDF Almacenamiento en Nube. Espacio para PDFs, imágenes extraídas, checkpoints de modelos y persistencia de ChromaDB.	Se va a implmentar en propia PC.	$0.00
2. Servicios y APIs	**Procesamiento Multimodal (Img/Tablas).** Uso de Gemini 1.5 Flash/Pro para describir imágenes y tablas (bounding boxes + semántica).	2,000 imágenes/tablas x $0.005 (promedio input/output por imagen)	$10.00
	Enriquecimiento Semántico y Keywords **Extracción de texto y generación de** metadatos con Gemini (Input + Output tokens).	5,000 páginas x ~350 tokens/pág x $0.50/1M tokens (aprox. Gemini Flash)	$875.00
	Evaluación con RAGAS (LLM-as-a-Judge) Uso de Gemini como "juez" para evaluar las respuestas generadas (métricas de RAGAS requieren llamadas intensivas al LLM).	2000 preguntas de test x 5 métricas x $0.01 por evaluación	$200.00
	LangChain y LangSmith (Monitorización) **Licencia Plus para trazabilidad avanzada y** cálculo de métricas ROUGE/BERTScore en producción/test.	Tarifa mensual por usuario/seat (Plan Developer/Plus) x 6 meses	$240.00
		TOTAL ESTIMADO	**$2,825.00**

Figura 4: Presupuesto de implementación de proyecto

4.3 Financiamiento

El stakeholder va a sustentar al área de finanzas en un proceso de caso de negocio para poder financiar el proyecto, sin embargo de cumplir con las espectativas de los usuarios el área de TI se compromete a crear un espacio para la implementación en el sistema de la compañia (Azure).

Capítulo 4.3: Anexos:

Sprint 1 - Modelo Baseline

Resumen Ejecutivo

EDA del sprint 1 donde se analizaron los documentos descargados del Aconex esta descarga consiste en **18.7Gb comprimido**(20.24 Gb descomprido) de documentos PDFs y otros comprimidos que no consideran unicamente manuales de mantenimiento, por lo que se realizo una revisión manual para identificar que documentos son efectivamente manuales, esto resulto en la reducción a **24 pdf que efectivamente son manuales de mantenimiento**, los cuales fueron procesados para la extracción de texto y posterior análisis.

Sprint Preparación de datos

Objetivo del Sprint: El principal objetivo del del sprint 01 es la extracción y analisis de los PDFs de manuales de mantenimiento de equipos críticos.

ID	Historia / Tarea	Responsable	Estado
0	Revisión de papers y baseline (CRISP-DM)	Johan Callomamani	Hecho
1	Preparación de datos (CRISP-DM)	Johan Callomamani	Hecho
2	Modelado (PDF/Imágenes)	Johan Callomamani	Por Empezar
3	Generación de Demo	Johan Callomamani	Por Empezar
4	Integración multimodal del RAG	Johan Callomamani	Por Empezar
5	Fine tuning del modelo base (Aplicación de LoRA)	Johan Callomamani	Por Empezar
6	Revisión de feedback y refinamiento	Johan Callomamani	Por Empezar

Data Pipeline Básico

1. **Descripción:**
   - Procesamiento de transmittals descargados del Aconex (18.7 Gb de documentos de todo tipo no solo manuales de mantenimiento).
   - Se identifica y clasifica los archivos que son manuales de mantenimiento.
   - Se realiza un análisis de los archivos generando una archivo csv con características de los archivos y graficas donde se analisan los pdf.
2. **Entregable:**
   - Lista de archivos transmittal `aconex_file.zip`
   - notebook de analisis de los pdfs `exploratory_analysis.ipynb`
   - `list_files.csv`
   - Graficas de análisis de los pdfs.

3. **Comentarios/Problemas:** Los documentos no son solo manuales, se tienen documentos adicionales, de no interés para el proyecto:
   - Información de ordenes de compra.
   - Solicitudes de transporte de componentes.
   - Documentos de aceptación de reportes y aceptación de proyectos.
   - Gantts de inspección/QA.
   - Invoices.
   - Monthly reports (avances de la etapa de proyecto).

EDA Rápido

1. **Hallazgos:**
   - Total de documentos 3607 archivos con un peso total 20.24 Gb (ver Figura 5).
   - Alta densidad de documentos PDF con un total 78.53% (ver Figura 5).
   - Se encontró que no todos los archivos son manuales, el 22% son pagos y solo el 60% corresponde a información técnica relevante.
   - Para una configuración de Size: 1000, Overlap=200 para realizar la partición de los documentos tenemos como resultado Total chunks (vectors): 56,625, avg chunks per document: 2359.4 (ver Figura 7).
   - Se tiene archivos mayoritariamente con paginas de 100 a 1800 paginas, sin embargo hay 2 archivos con cantidad superior a las 5000 paginas (ver Figura 6)
   - Se valida que se tiene una relación directa entre cantidad de paginas y el tamaño del archivo (ver Figura 6), lo cual indica que normalmente todas las paginas tiene contenido y no son paginas en blanco.
   - Considerar que el ruido en el texto (simbolos, caracteres especiales) es en su mayoria menor al 0.15 lo que indica que la calidad de los documentos es buena (ver Figura 8), sin embargo hauy 1 archivo que supera por poco y requiere una revisión adicional.
   - Considerar que los documentos de menor cantidad de palabras son los que tiene mayor diversidad de vocabulario (ver Figura 8), lo que indica que son documentos con mayor cantidad de paginas/palabras reduce su vocabulario diverso.
   - Los archivos son predominantemente manuales del año 2019 (ver Figura 8).
   - Se requiere revisar y retirar terminología que se repite constantemente en todas las hojas, textos de los foot headers, numeración de paginas, logos de empresas, etc (ver Figura 8), ya que al ser predominandtes en los pdfs puede afectar en la implementación del RAG.
2. **Visualizaciones:**.

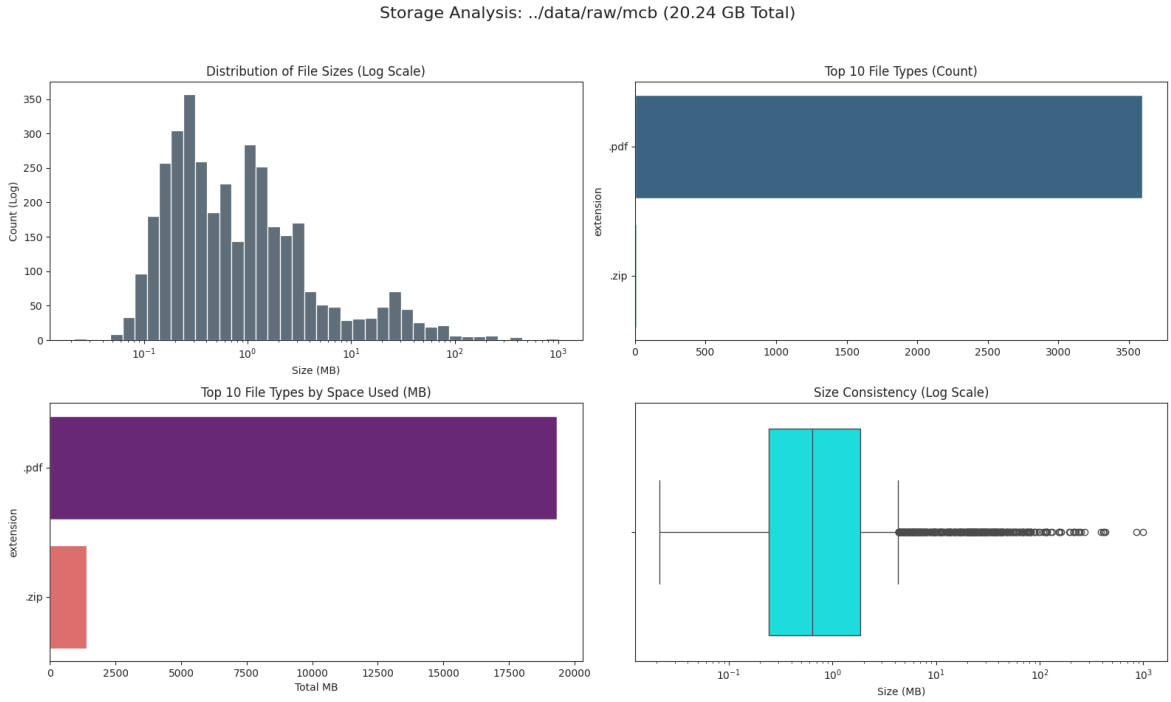

Figura 5: Análisis del total de archivos **sección 01** Distribusión del peso de los archivos por cantidad de estos, **sección 02** Cantidad de archivos por tipo de archivos, **sección 03** Peso de archivos por tipo de archivos, **sección 04** Boxplot del peso de los archivos

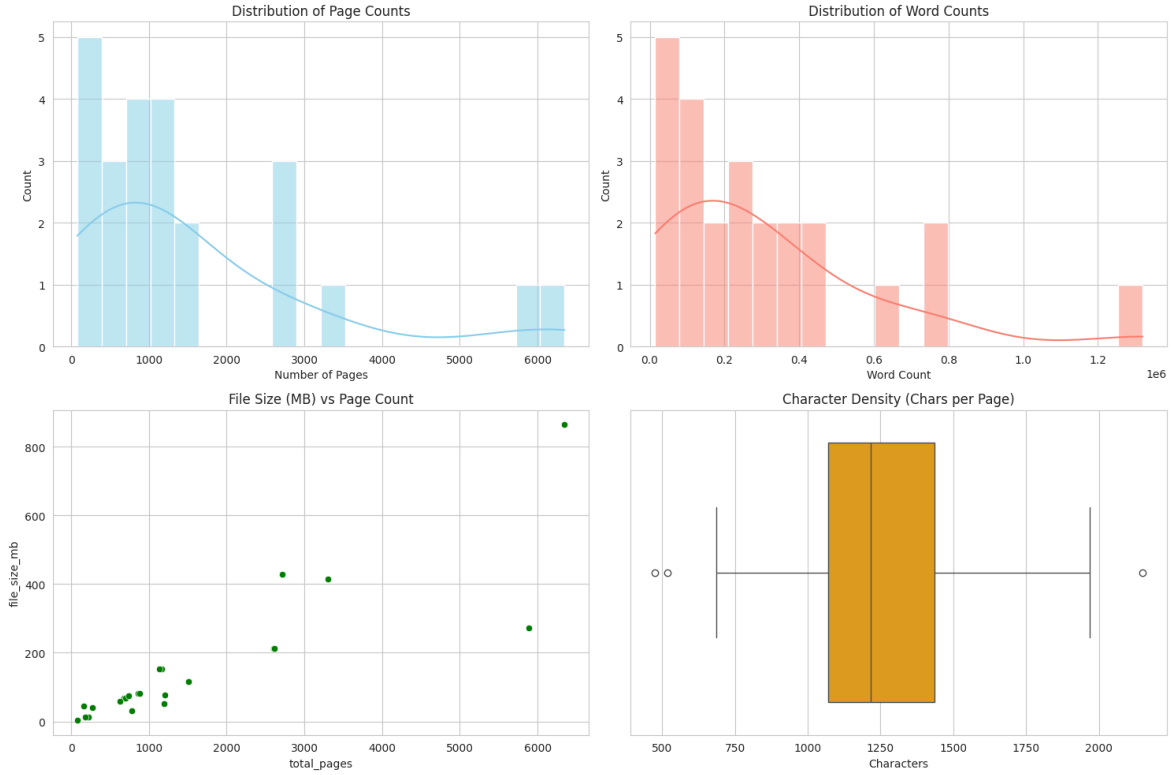

Figura 6: Análisis de 24 pdf Manuales de mantenimiento de equipos críticos **sección 01:** Cantidad de archivos por cantidad de paginas de los archivos, **sección 02:** Cantidad de archios por cantidad de palabras, **sección 03:** tamaño de los archivos por cantidad de paginas, **sección 04:** Boxplot de la cantidad de caracteres

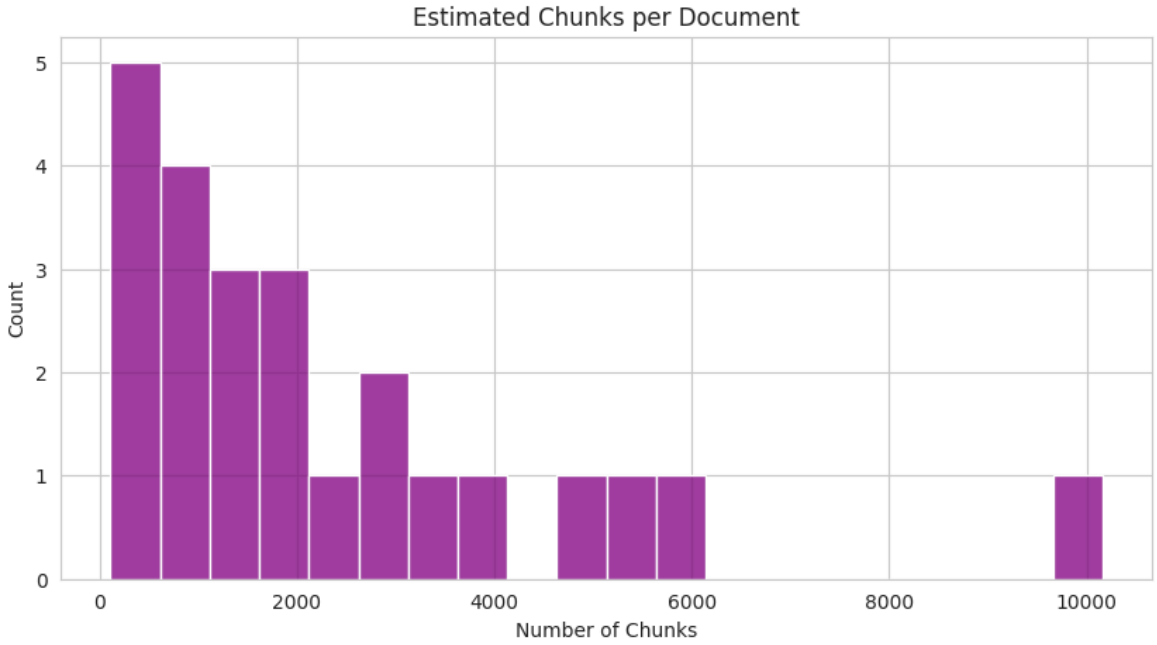

Figura 7: Análisis de estimación de Chunks por archivos

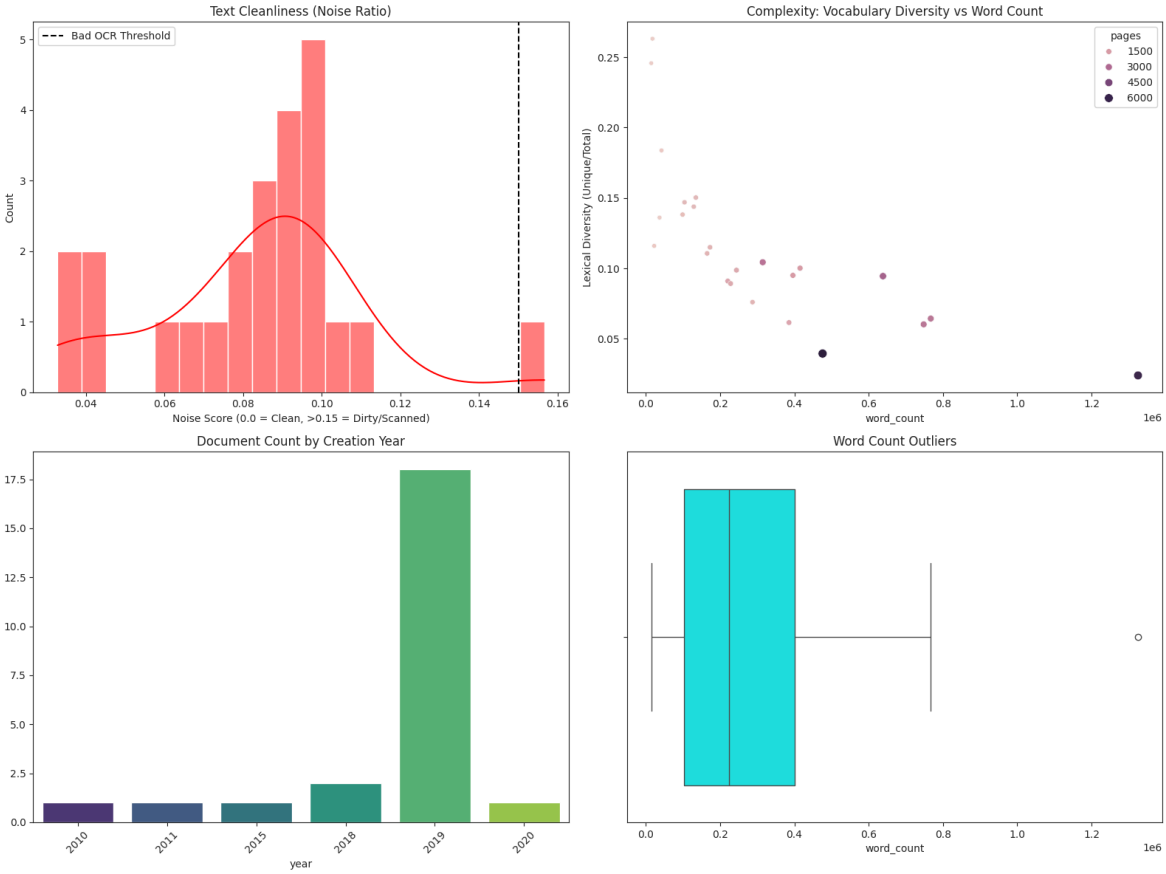

Figura 8: Análisis del contenido de los archivos de mantenimiento

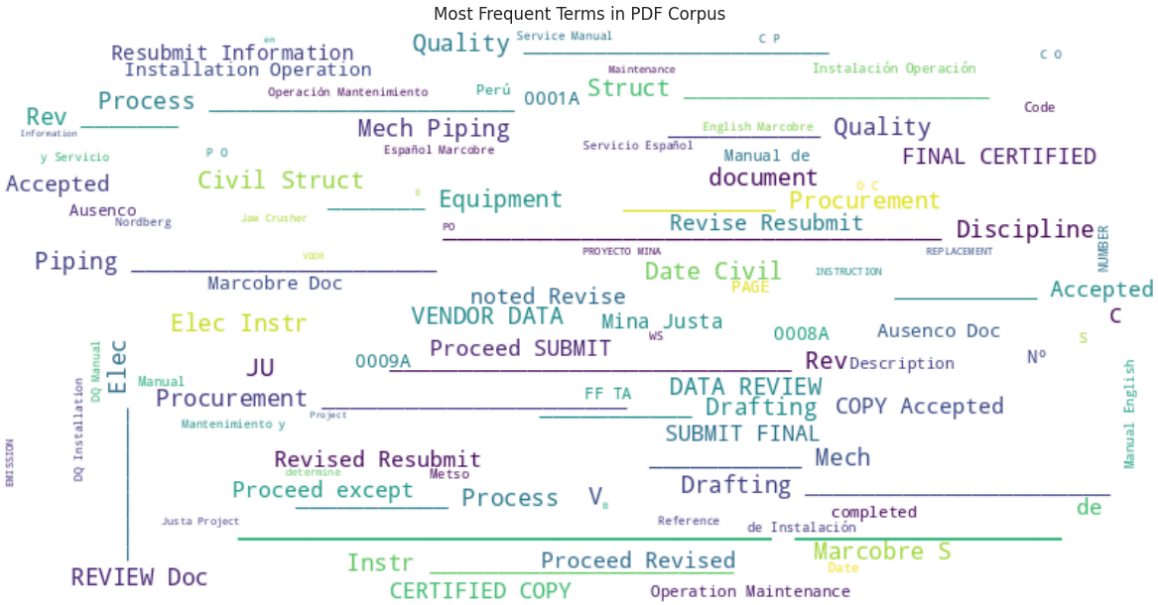

Figura 9: Mayor cantidad de terminas en los archivos de mantenimiento

- **Entregables:**
  - Informe EDA.
  - Piline de procesamiento de PDFs.
  - Graficas(Figura 5, Figura 6, Figura 7, Figura 8 y Figura 9).

Capítulo 5: Referencias Bibliográficas

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020, julio). *Language Models are Few-Shot Learners*. arXiv. https://doi.org/10.48550/arXiv.2005.14165

Cheng, S., Li, J., Wang, H., & Ma, Y. (2025). RAGTrace: Understanding and Refining Retrieval-Generation Dynamics in Retrieval-Augmented Generation. *Proceedings of the 38th Annual ACM Symposium on User Interface Software and Technology*, 1-20. https://doi.org/10.1145/3746059.3747741

Drushchak, N., Polyakovska, N., Bautina, M., Semenchenko, T., Koscielecki, J., Sykala, W., & Wegrzynowski, M. (2025). *Multimodal Retrieval-Augmented Generation: Unified Information Processing Across Text, Image, Table, and Video Modalities*.

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

Ibtasham, S., Bashir, S., Abbas, M., Haider, Z., Saadatmand, M., & Cicchetti, A. (2025). *Req-RAG: Enhancing Software Release Management through Retrieval-Augmented LLMs: An Industrial Study*.

Kandpal, N., Deng, H., Roberts, A., Wallace, E., & Raffel, C. (2023). Large Language Models Struggle to Learn Long-Tail Knowledge. *Proceedings of the 40th International Conference on Machine Learning*, 15696-15707. https://proceedings.mlr.press/v202/kandpal23a.html

Knollmeyer, S., Caymazer, O., & Grossmann, D. (2025b). 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

Knollmeyer, S., Pfaff, S., Akmal, M. U., Koval, L., Asif, S., Mathias, S. G., & Großmann, D. (2025a). Hybrid Retrieval for Retrieval Augmented Generation in the German Language Production Domain. *Journal of Advances in Information Technology*, *16*(6), 819-829. https://doi.org/10.12720/jait.16.6.819-829

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., Riedel, S., & Kiela, D. (2021, abril). *Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks*. arXiv. https://doi.org/10.48550/arXiv.2005.11401

Liu, P., Liu, X., Yao, R., Liu, J., Meng, S., Wang, D., & Ma, J. (2025). HM-RAG: Hierarchical Multi-Agent Multimodal Retrieval Augmented Generation. *Proceedings of the 33rd ACM International Conference on Multimedia*, 2781-2790. https://doi.org/10.1145/3746027.3754761

Mei, L., Mo, S., Yang, Z., & Chen, C. (2025, marzo). *A Survey of Multimodal Retrieval-Augmented Generation* (Número arXiv:2504.08748). arXiv. https://doi.org/10.48550/arXiv.2504.08748

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

Palmer, D. (2019). *Maintenance Planning and Scheduling Handbook, Fourth Edition* (4th edition.). McGraw-Hill Education,.

Rajaraman, A., & Ullman, J. D. (2011). *Mining of Massive Datasets*. Autoedicion. https://biblioteca.unisced.edu.mz/handle/123456789/2674

Robertson, S., & Zaragoza, H. (2009). The Probabilistic Relevance Framework: BM25 and Beyond. *Foundations and Trends® in Information Retrieval*, *3*(4), 333-389. https://doi.org/10.1561/1500000019

Santra, P., Ghosh, M., Ganguly, D., Basuchowdhuri, P., & Naskar, S. K. (2025). The “Curious Case of Contexts” in Retrieval-Augmented Generation With a Combination of Labeled and Unlabeled Data. *WIREs Data Mining and Knowledge Discovery*, *15*(2), e70021. https://doi.org/10.1002/widm.70021

Shu, X., Dang, X., Dong, X., & Li, F. (2024). Utilizing Large Language Models for Hyper Knowledge Graph Construction in Mine Hoist Fault Analysis. *Symmetry*, *16*(12), 1600. https://doi.org/10.3390/sym16121600

Wang, Y., Wan, Y., Lei, X., Chen, Q., & Hu, H. (2025). A retrieval augmented generation based optimization approach for medical knowledge understanding and reasoning in large language models. *Array*, *28*, 100504. https://doi.org/10.1016/j.array.2025.100504

Zhang, D., Yu, Y., Dong, J., Li, C., Su, D., Chu, C., & Yu, D. (2024, mayo). *MM-LLMs: Recent Advances in MultiModal Large Language Models*. arXiv. https://doi.org/10.48550/arXiv.2401.13601