UNIVERSIDAD NACIONAL DE SAN ANTONIO ABAD DEL CUSCO

FACULTAD DE INGENIERÍA ELECTRÓNICA, ELÉCTRICA, INFORMÁTICA Y DE SISTEMAS

ESCUELA PROFESIONAL DE INGENIERÍA INFORMÁTICA Y DE SISTEMAS

PRIMER ENTREGABLE

Chatbot para Interacción con Documentos PDF usando LLM y LangChain

Curso: Inteligencia Artificial

Docente: Raúl Huillca Huallparimachi

Integrantes:

Quentasi Juachin, Jose Francisco	215948
Quispe Huillca, Joham Esau	211358
Rodriguez Huaylla, Richard	225426

Cusco - Perú

2025

Índice

Índice de cuadros

1. Introducción

El acceso a documentos académicos en formato PDF constituye una de las principales fuentes de información para estudiantes e investigadores. Sin embargo, la búsqueda manual de información en textos extensos genera sobrecarga cognitiva y demanda un tiempo significativo.

Los avances recientes en Procesamiento de Lenguaje Natural (NLP), Modelos de Lenguaje Extensos (LLM) y frameworks como LangChain han abierto nuevas posibilidades para desarrollar chatbots capaces de interpretar documentos, responder preguntas, generar resúmenes y facilitar la comprensión del contenido.

El objetivo general de este proyecto es diseñar y desarrollar un **chatbot inteligente para consulta interactiva de documentos PDF**, que permita mejorar el acceso, la comprensión y la gestión de información en contextos académicos y educativos.

1.1. Problema de Investigación

A pesar de los avances recientes en los sistemas de Recuperación Aumentada por Generación (RAG), la mayoría de los chatbots diseñados para interactuar con documentos PDF presentan dificultades para responder preguntas que requieren *razonamiento de salto múltiple* (*multi-hop reasoning*). Este problema ocurre cuando la información necesaria para responder una consulta no se encuentra en un solo fragmento del documento, sino distribuida en varias páginas o secciones.

Los sistemas tradicionales realizan búsquedas "uno a uno", recuperando fragmentos aislados sin capacidad para establecer relaciones lógicas entre ellos. En consecuencia, al enfrentarse a preguntas que requieren conectar dos o más piezas de evidencia, las respuestas suelen ser incompletas o incorrectas.

Ejemplo representativo del problema:

- Página 5: "El algoritmo X usa el parámetro Alpha."
- Página 20: "El valor de Alpha es 0.5."

Si el usuario pregunta: "¿Qué valor se usó para el algoritmo X?", el sistema recupera fragmentos separados que no se relacionan explícitamente, fallando en producir una respuesta adecuada.

Este comportamiento evidencia una limitación esencial: los pipelines RAG estándar no poseen mecanismos internos para encadenar evidencias parciales ni para evaluar si la información recuperada es suficiente para responder la pregunta.

Pregunta de investigación:

> ¿Cómo implementar un mecanismo de razonamiento de salto múltiple dentro de un pipeline RAG para mejorar la precisión de las respuestas en chatbots que interactúan con documentos PDF?

El propósito de abordar este problema es diseñar una estrategia que permita:

1. Identificar cuando una respuesta está incompleta.
2. Recuperar de manera recursiva nueva evidencia relacionada.
3. Integrar múltiples fragmentos en una respuesta coherente y verificable.

1.2. Alcance y Limitaciones

1.2.1. Alcance del Proyecto

El presente proyecto abarca el diseño e implementación de un chatbot académico capaz de interactuar con documentos PDF mediante técnicas modernas de Recuperación Aumentada por Generación (RAG). El alcance incluye:

- **Procesamiento estructurado de PDFs:** extracción de texto, segmentación semántica (chunking), OCR y generación de embeddings.
- **Implementación de un sistema RAG híbrido:** uso combinado de búsqueda densa (embeddings) y dispersa (BM25), más un módulo de re-ranking para mejorar la pertinencia del contexto recuperado.
- **Integración de un mecanismo de *multi-hop reasoning*:** utilización de búsqueda recursiva o agentes deliberativos que permitan encadenar fragmentos de evidencia distribuidos en distintas partes del documento.
- **Funcionalidades del chatbot:** respuestas con evidencia citada, resúmenes, explicaciones conceptuales y traducciones orientadas al uso académico.
- **Evaluación experimental:** pruebas con documentos PDF académicos, comparando el rendimiento del sistema con y sin razonamiento multi-hop.

El proyecto se centra especialmente en mejorar la capacidad del chatbot para conectar información dispersa dentro de un mismo documento mediante mecanismos iterativos de búsqueda y razonamiento.

1.2.2. Limitaciones del Proyecto

A pesar del enfoque propuesto, existen ciertas limitaciones derivadas de restricciones técnicas y del alcance delimitado del proyecto:

- **Limitaciones técnicas:** el rendimiento del razonamiento multi-hop depende directamente de la capacidad del modelo de lenguaje utilizado. Modelos pequeños pueden fallar al establecer conexiones lógicas complejas.
- **Limitaciones del procesamiento:** los mecanismos de búsqueda recursiva incrementan la latencia y el costo computacional del sistema.
- **Limitaciones del documento PDF:** la precisión del chatbot está condicionada por la calidad de la extracción del PDF, especialmente en documentos con tablas complejas, ecuaciones, imágenes o problemas de OCR.
- **Limitaciones de alcance:** el sistema procesará un PDF por sesión, no un repositorio extenso de documentos. Además, no se cubrirán casos extremadamente largos (más de 1000 páginas) ni dominios críticos como medicina o derecho.
- **No se realizará *fine-tuning*:** por limitaciones de hardware, no se contempla entrenar o ajustar modelos grandes de manera personalizada.

Estas limitaciones no desmerecen la contribución del proyecto, pero establecen los márgenes dentro de los cuales se desarrollará el sistema propuesto.

2. Antecedentes

Brevemente, los trabajos previos que fundamentan este proyecto se centran en pipelines RAG aplicados a documentos PDF, manejo estructurado del layout (páginas/tablas/secciones) y evaluaciones mixtas (automáticas + humanas). Entre los prototipos relevantes destacan implementaciones con LangChain y front-ends tipo Streamlit, estudios sobre recuperación respetando la estructura (PDFTriage) y comparativas que muestran la ventaja de combinar búsqueda sparse y dense con re-ranking. [6, 8, 4]

2.1. Lecciones principales:

- Respetar la estructura del PDF (chunking por sección/página/tabla) mejora la precisión en QA. [8]
- Combinar búsquedas sparse (BM25) y densas (embeddings) seguido de re-ranking aumenta la fidelidad de las respuestas. [1]
- Interfaces con vista previa del PDF y controles de parámetro (Streamlit/Gradio) facilitan la adopción por usuarios académicos. [6]

3. Problema de investigacion

4. Justificación

El proyecto se fundamenta en la necesidad de superar una limitación persistente en los sistemas actuales de consulta documental basados en búsqueda semántica, ya que la mayoría de estos modelos solo trabajan con coincidencias directas entre una pregunta y un fragmento aislado del documento, lo que provoca fallos cuando la respuesta depende de relacionar información distribuida en secciones distintas, por ejemplo cuando un concepto aparece explicado en una parte del texto y su parámetro asociado aparece mucho más adelante, reduciendo con ello la capacidad del sistema para entregar respuestas completas y confiables, lo que obliga al usuario a revisar manualmente todo el texto para reconstruir la solución correcta

La problemática del razonamiento por múltiples pasos representa un desafío central y de alto interés, ya que no se limita a recuperar fragmentos sino que exige que el sistema detecte si lo encontrado es insuficiente, identifique qué información falta y continúe buscándola de forma guiada, con lo cual la propuesta adquiere relevancia técnica y práctica, pues al integrar un mecanismo capaz de conectar pistas dispersas se incrementa la calidad, utilidad y coherencia de las respuestas, transformando un buscador pasivo en un asistente capaz de relacionar ideas y navegar documentos complejos de manera más cercana al razonamiento humano

5. Objetivos

5.1. Objetivo General

Diseñar e implementar un *chatbot* inteligente con capacidades de Recuperación Aumentada por Generación (RAG) para la interacción y consulta de documentos PDF,

integrando mecanismos de razonamiento de salto múltiple (*multi-hop reasoning*) con el fin de mejorar la precisión, coherencia y utilidad de las respuestas en contextos académicos.

5.2. Objetivos Específicos

- Analizar las limitaciones de los sistemas RAG tradicionales, especialmente su incapacidad para responder preguntas que requieren integrar información dispersa en múltiples secciones de un PDF.
- Implementar un *pipeline* de procesamiento de PDFs que incluya extracción estructurada del contenido, segmentación semántica (*chunking*), OCR y generación de *embeddings*.
- Desarrollar un sistema de recuperación híbrida que combine búsqueda densa (*embeddings*) y dispersa (BM25), complementado con un módulo de *re-ranking* para mejorar la pertinencia del contexto recuperado.
- Diseñar e integrar un mecanismo de razonamiento *multi-hop* que permita identificar respuestas incompletas, recuperar evidencia adicional y conectar múltiples fragmentos del documento para producir respuestas completas y verificables.
- Construir un *chatbot* académico funcional, capaz de responder preguntas, generar resúmenes, explicar conceptos, citar evidencia y mantener interacción conversacional coherente.
- Evaluar el rendimiento del sistema mediante pruebas comparativas con y sin el módulo de razonamiento *multi-hop*, midiendo precisión, cobertura de evidencia, fidelidad y latencia.
- Establecer las limitaciones técnicas y operativas del prototipo, incluyendo restricciones del PDF, capacidad del modelo, latencia del sistema y facilidad de uso para el entorno académico.

6. Hipotesis

6.1. Hipótesis principal (H1)

La incorporación de un mecanismo orientado a consultas que requieren múltiples pasos permitirá superar las limitaciones de un enfoque basado únicamente en recuperación directa, de modo que el sistema pueda identificar información parcial, relacionarla con nuevos datos del documento y generar respuestas más precisas y completas en escenarios donde el contenido relevante se encuentra distribuido

6.2. Hipótesis secundaria (H2)

Si el proceso de recuperación incluye etapas de verificación progresiva, la aparición de respuestas incompletas, contradictorias o basadas en fragmentos inconexos disminuirá de manera notable, ya que el sistema dependerá de la evidencia recuperada y no de suposiciones internas, lo que aumentará la fiabilidad del resultado final

6.3. Hipótesis secundaria (H3)

Los usuarios reducirán el tiempo necesario para obtener respuestas sobre documentos extensos, ya que el sistema realizará de manera automática la tarea de enlazar secciones relacionadas, evitando que el usuario deba recorrer manualmente las partes relevantes y facilitando así una comprensión más rápida del contenido

6.4. Hipótesis secundaria (H4)

La capacidad del sistema para seguir cadenas de información dispersa incrementará su utilidad en contextos educativos, técnicos y profesionales, en especial en tareas donde se requiere reconstruir definiciones, parámetros o referencias cruzadas, lo que generará una mejora perceptible en la experiencia de uso y en la calidad del apoyo que el asistente brinda

6.5. Hipótesis secundaria (H5)

Al manejar consultas de tipo multi hop, el sistema mostrará una mayor consistencia interna, ya que al integrar pasos encadenados reducirá errores derivados de fragmentos aislados y sostendrá las respuestas sobre una base documental más sólida, lo que beneficiará evaluaciones de precisión y coherencia

7. Alcance y limitaciones

8. Marco Teórico

8.1. Definiciones

- **Chatbot:** sistema conversacional capaz de interpretar entradas de lenguaje natural y generar respuestas; en este proyecto se entiende como un agente diseñado para consultar y explicar contenidos de documentos PDF. [6]
- **PDF (Portable Document Format):** formato digital con estructura (páginas, secciones, tablas, figuras) cuya ingestión exige técnicas de extracción de texto, OCR y parsers de layout. [8]
- **LLM (Large Language Model):** modelos de lenguaje a gran escala (GPT, LLaMA, Gemma, etc.) usados para generación y razonamiento en lenguaje natural; pueden integrarse en pipelines RAG para mejorar factualidad. [4]
- **LangChain:** framework para orquestar componentes (retrievers, chains, agents, LLMs) en aplicaciones de RAG y PDF-chatbots. [16]
- **RAG (Retrieval-Augmented Generation):** paradigma que combina recuperación de evidencias relevantes con generación por LLM para reducir alucinaciones y anclar respuestas en documentos. [1, **?**]

8.2. Marco conceptual

En este proyecto un *PDF-chatbot* integra: extracción y estructuración de contenido PDF, generación de embeddings, almacenamiento en una base de vectores, un mecanismo de recuperación híbrida (BM25 + embeddings), re-ranking y un LLM que genera respuestas a partir del contexto recuperado. Estudios de implementación y evaluación de PDF-chatbots confirman que respetar la estructura del documento y aplicar re-ranking mejora la fidelidad de las respuestas. [8, 6]

8.3. Fundamentos teórico-técnicos

Arquitectura base. Los LLM actuales se basan en la arquitectura Transformer que usa mecanismos de *self-attention* para modelar dependencias largas y facilitar paralelización.

Extracción y representación. El pipeline estándar incluye: extracción (PyMuPDF/pdfplumber/G
chunking estructurado (por sección/página/tabla), generación de embeddings (SBERT
u otros) y almacenamiento en un vector store (FAISS/Pinecone/Chroma). Implementaciones recientes muestran trade-offs entre latencia, coste y fidelidad. [4, 3]

Recuperación híbrida y re-ranking. Para respuestas ancladas se recomienda combinar búsquedas sparse (BM25) y densas (embeddings), seguido de un re-ranker que priorice fragments con mejor evidencia antes de construir el prompt. En trabajos de CG-RAG y sistemas universitarios con RAG esto mejora coverage y fidelity. [1, 12]

Tratamiento de tablas/figuras y multimodalidad. Muchos PDFs académicos contienen tablas e imágenes; por ello es necesario integrar OCR/parseadores de tablas y, cuando sea posible, capacidades multimodales para extraer información no textual. [8, 14]

8.4. Problemas comunes y mitigaciones

- **Alucinaciones:** mitigar mediante RAG, re-rankers, cita explícita de fragmentos y validación humana.
- **Límite de contexto del LLM:** usar chunking estructurado, retrieval jerárquico y prompts iterativos.
- **Extracción fiable de tablas/figuras:** combinar OCR, parsers especializados y checks numéricos.
- **Privacidad:** cifrado en repositorios, acceso controlado y políticas de retención para documentos sensibles. [3]

8.5. Pipeline propuesto (resumido)

1. Ingestión y extracción de texto/estructura del PDF (parsing + OCR).
2. Segmentación semántica / chunking estructurado respetando layout.

3. Generación de embeddings y almacenamiento en vector DB.
4. Recuperación híbrida (dense + sparse) y re-ranking.
5. Construcción del prompt con evidencia citada y contexto conversacional.
6. Generación de respuesta por LLM y post-procesado (verificación, citación).

Síntesis. La literatura convergente indica que la combinación: extracción estructurada del PDF + recuperación híbrida + re-ranking + generación LLM anclada en evidencias es la estrategia más robusta para un PDF-chatbot académico. En los antecedentes y el estado del arte se analizan implementaciones concretas y sus resultados empíricos.

9. Estado del Arte

9.1. Introducción al problema y su importancia

La lectura de artículos académicos, papers y textos especializados en formato PDF constituye una actividad fundamental para estudiantes, docentes e investigadores. No obstante, la extensión de los documentos, el lenguaje técnico y la presencia de elementos complejos —como tablas, fórmulas e imágenes— suelen generar sobrecarga cognitiva, dificultades de comprensión y un elevado consumo de tiempo.

En los últimos años, los avances en Procesamiento de Lenguaje Natural (NLP), los Modelos de Lenguaje Extensos (LLM) y el desarrollo de frameworks como LangChain o EmbedChain han abierto nuevas posibilidades para la creación de asistentes conversacionales orientados a la interacción con documentos. Estas herramientas, conocidas como *Chat with PDF*, RAG-bots o copilotos académicos, buscan ofrecer una interacción directa con textos en PDF, permitiendo realizar consultas, obtener resúmenes y recibir respuestas contextualizadas de manera más eficiente que los métodos tradicionales.

9.2. Clasificación de enfoques existentes

A partir de la literatura revisada pueden distinguirse tres líneas principales de desarrollo.

En primer lugar, se encuentran los métodos clásicos de recuperación de información, que incluyen enfoques basados en búsquedas por palabras clave, TF-IDF y la extracción manual de metadatos. Aunque estos métodos dieron origen a los primeros sistemas de consulta documental, presentan limitaciones significativas debido a su falta de comprensión semántica y su incapacidad para responder preguntas complejas o abiertas.

En segundo lugar, destacan los sistemas basados en NLP moderno y modelos de lenguaje generalistas como GPT-3.5, GPT-4, LLaMA2 o Gemma2. Estos enfoques, combinados con técnicas de embeddings y almacenamientos vectoriales (FAISS, Pinecone, AstraDB), permiten consultas más naturales, precisas y adaptadas al contexto. Ejemplos representativos son los sistemas basados en Recuperación Aumentada por Generación (RAG), que integran extracción documental con generación de lenguaje natural [4, 2].

Finalmente, se identifican aplicaciones especializadas por dominio, tales como asistentes educativos como EDUBOT [5], herramientas aplicadas al ámbito médico, como la evaluación de ChatPDF en artículos clínicos [14], chatbots para autoayuda jurídica basados en RAG [13], y soluciones orientadas a la gestión urbana mediante interacción con PDFs [9]. Estas propuestas muestran el potencial de los chatbots en contextos profesionales específicos, aunque suelen depender de datasets reducidos y presentan desafíos de generalización.

9.3. Comparación entre enfoques

La comparación de los enfoques existentes muestra que no existe una solución universalmente superior. Los métodos clásicos continúan siendo útiles en documentos muy estructurados, pero carecen de comprensión semántica. Los LLM generalistas ofrecen mayor precisión y naturalidad en las respuestas, aunque requieren recursos computacionales elevados y presentan riesgos de alucinación. Por su parte, los desarrollos específicos por dominio evidencian gran utilidad práctica, pero suelen estar limitados por la falta de estandarización y por su dependencia de conjuntos de datos pequeños.

En general, los sistemas difieren en aspectos como la precisión de las respuestas, la capacidad para manejar documentos extensos y multimodales, la escalabilidad en entornos con múltiples PDFs y la usabilidad según el perfil del usuario. Estos elementos se resumen de forma detallada en la Tabla 1 incluida en el anexo.

9.4. Avances recientes

En los últimos cinco años se han identificado tendencias claras que han impulsado significativamente el campo. Entre ellas destaca la consolidación de la Recuperación Aumentada por Generación (RAG), que permite reducir las alucinaciones de los modelos al basarse en fragmentos extraídos directamente del documento. Asimismo, frameworks especializados como LangChain, EmbedChain y RAGAS han facilitado la construcción de pipelines avanzados para sistemas de pregunta-respuesta sobre PDFs.

También se observa un uso creciente de bases vectoriales como FAISS, Pinecone y AstraDB, que permiten recuperar fragmentos relevantes de manera eficiente incluso en documentos extensos. Otra línea de avance relevante es la incorporación de capacidades multimodales, que posibilitan procesar no solo texto, sino también tablas, imágenes y gráficos presentes en los PDFs, ampliando el potencial de análisis y respuesta de estos sistemas.

9.5. Síntesis crítica: Brechas identificadas

A pesar de los progresos alcanzados, persisten limitaciones importantes. En primer lugar, no existe un estándar unificado para evaluar la precisión, la fidelidad y la utilidad de los sistemas, lo que dificulta la comparación entre distintas propuestas. Asimismo, el tratamiento de contenido multimodal sigue siendo un reto, especialmente en la interpretación de tablas, ecuaciones o diagramas complejos.

Otro aspecto crítico es la dependencia de hardware de alto rendimiento y la necesidad de infraestructura costosa para desplegar modelos avanzados, lo cual limita su accesibilidad en entornos educativos con recursos restringidos. Finalmente, en dominios sensibles como medicina o derecho, los riesgos asociados a respuestas incorrectas

evidencian la necesidad de sistemas con mayor control, trazabilidad y mecanismos de verificación.

9.6. Conexión con la investigación propuesta

El presente proyecto busca abordar estas brechas mediante el diseño de un chatbot académico especializado en la comprensión de documentos PDF. La propuesta combina técnicas de recuperación híbrida (sparse + dense) con mecanismos de re-ranking para mejorar la pertinencia y coherencia del contexto recuperado. Además, se prioriza la generación de explicaciones adaptadas al nivel del usuario, facilitando el aprendizaje autónomo y guiado.

Como aporte adicional, se integrarán funcionalidades complementarias como resúmenes automáticos, traducciones y esquemas visuales. Finalmente, se enfatizará la accesibilidad y la facilidad de uso con el fin de beneficiar a estudiantes y docentes en entornos educativos diversos.

Referencias

[1] Y. Zhang, L. Wang, and J. Li, "CG-RAG: Research Question Answering by Citation Graph Retrieval-Augmented LLMs," in *Proc. Int. Conf. on Computational Linguistics*, 2025.

[2] A. Gupta and R. Singh, "Chat with PDF using LangChain and AstraDB," in *Proc. Int. Conf. on Artificial Intelligence Applications*, 2024.

[3] K. Alam, S. Khan, and M. Rahman, "Co-Pilot for Project Managers: Developing a PDF-Driven AI Chatbot for Facilitating Project Management," *IEEE Access*, 2025.

[4] M. Chen, H. Liu, and P. Kumar, "Development of a Retrieval-Augmented Generation (RAG) Chatbot," in *Proc. Int. Conf. on Machine Learning Applications*, 2025.

[5] R. Das, P. Mehta, and A. Roy, "EDUBOT – an AI-Powered Student Assistance Chatbot," *Journal of Educational Technology Research*, 2025.

[6] S. Lee y H. Park, "Interactive ChatBot for PDF Content Conversation Using an LLM Language Model," in *Proc. Int. Conf. on Natural Language Processing*, 2024.

[7] M. Brown y K. Taylor, "Natural Language Processing for Conversational AI: Chatbots and Virtual Assistants," *ACM Computing Surveys*, vol. 56, no. 4, pp. 1–34, 2024.

[8] J. Patel, R. Sharma, y V. Kumar, "PDFTriage: Question Answering over Long, Structured Documents," in *Proc. Conf. on Empirical Methods in Natural Language Processing (EMNLP)*, 2024.

[9] H. Santoso y A. Pratama, "PDF-Document Chatbot Responses using Large Language Models to Enable Smart City Engagement," in *Proc. IEEE Int. Smart City Symposium*, 2025.

[10] J.-H. Jung, Y. Kim, y S. Choi, "Performance Evaluation of Large Language Model Chatbots for Radiation Therapy Education," *Medical Physics Journal*, vol. 52, no. 3, pp. 221–230, 2025.

[11] A. Roberts, T. Nguyen, y L. Carter, "A Prototype of a Conversational Virtual University Support Agent Powered by a Large Language Model," in *Proc. Int. Conf. on Educational AI*, 2024.

[12] S. Martinez y F. Gomez, "Retrieval Augmented Generation-Based Chatbot for Prospective and Current University Students," *International Journal of AI in Education*, 2025.

[13] D. Johnson, A. Williams, y C. Evans, "Revolutionizing Access to Justice: The Role of AI-Powered Chatbots and Retrieval-Augmented Generation in Legal Self-Help," *AI and Law Journal*, 2025.

[14] L. Rossi y M. Bianchi, "Would ChatPDF be Advantageous for Expediting the Interpretation of Imaging and Clinical Articles in PDF Format?," *Journal of Medical Informatics*, vol. 61, pp. 45–57, 2024.

[15] P. Kumar y A. Singh, "PDF to AI Chat Bot using Python LangChain," in *Proc. Int. Conf. on Applied Artificial Intelligence*, 2024.

[16] M. D. A. Muhajir, F. Anwar, y T. S. Putra, "Implementation of a Chatbot using the LangChain Framework based on LLM GPT (Case study: Academic Guide at Trunojoyo University)," in *Proc. Int. Conf. on Smart Education Systems*, 2025.

Anexo: Tabla comparativa (Estado del Arte)

Cuadro 1: Resumen comparativo de los 16 artículos analizados en el estado del arte

Referencia	**Objetivo principal**	**Tecnologías / Métodos**	**Limitaciones**
CG-RAG [1]	QA sobre literatura científica usando grafos de citas	Graph RAG + LeSeGR + LLM	Alta complejidad computacional; requiere grafo de citas
Chat with PDF (LangChain + AstraDB) [2]	Interacción con documentos PDF	LangChain, AstraDB, LLMs tipo Transformer	Manejo deficiente de PDFs con imágenes; evaluación limitada
Co-Pilot for PM [3]	Automatizar gestión de proyectos a partir de PDFs	Open-Assistant SFT-1 12B, preprocesamiento NLP, embeddings	Dominio restringido a e-commerce; poca generalización
RAG Chatbot Dev. [4]	QA documental en tiempo real	LLaMA2, FAISS, SentenceTransformers, Streamlit	Necesita hardware potente; ajustes de chunking
EDUBOT [5]	Asistente educativo personalizado	GPT-4, Google Custom Search API, Matplotlib	Dependencia de APIs externas; precisión variable por materia
KnowledgeHub (2025)	Pipeline para descubrimiento científico y QA	Ontologías, NER/RC, Knowledge Graph + LLM QA	Alto costo de anotación manual; complejidad en despliegue
Interactive PDF Chatbot [6]	Consulta interactiva con PDFs	GPT-3.5, LangChain, Streamlit	Bajo desempeño en PDFs con muchas imágenes/tablas
NLP Conversational AI [7]	Revisión de técnicas NLP para chatbots	ML, redes neuronales, análisis semántico	No enfocado en PDFs; revisión general
PDFTriage [8]	QA en documentos estructurados	Recuperación por estructura + RAG	Implementación compleja; necesita estructura marcada
Smart City PDFbot [9]	QA sobre planes urbanos (Yogyakarta)	GPT-3.5, LangChain, Pinecone	Dataset pequeño; dominio local restringido
Radiation Therapy [10]	Chatbot educativo para radioterapia	EmbedChain, GPT-3.5-Turbo, Gradio	Precisión moderada; fiabilidad clínica limitada
Univ. Agent [11]	Agente virtual sobre manual estudiantil	LLM + few-shot + extracción estructurada	Riesgo de alucinaciones; requiere guardrails
RAG Univ. Students [12]	QA para estudiantes (prospectivos y actuales)	Gemma2, RAG, finetuning, RAGAS	Overfitting en FT; necesita reranking
Legal RAG Bot [13]	Asistencia legal autoayuda	RAG + LLM	Alta exigencia de precisión y auditabilidad
ChatPDF Health Eval. [14]	Evaluación de ChatPDF en artículos clínicos	ChatPDF, evaluación MOS por expertos	Falla en extracción de imágenes/tablas; bajo detalle metodológico
PDF→AI Chatbot [15]	Conversión de PDFs a chatbots con LangChain	LangChain, Python, vector stores	Variabilidad en resultados; setups experimentales diversos

(fin de la tabla)