Reporte Técnico ChatBot de Inteligencia Artificial con Arquitectura RAG - RAGvengers

Jerónimo Toro Calvo y Juan Pablo Valencia Chaves

Facultad de Ingeniería, Universidad de Caldas
Sistemas Inteligentes I - grupo 1

Profesor: Luis Fernando Castillo Ossa

Diciembre 2025

TABLA DE CONTENIDO

1. Introducción y Objetivos **3**
2. Corpus de Conocimientos **4**
3. Corpus de Conocimientos **6**
4. Protocolo de Evaluación **6**
6. Discusión **10**
Anexo: Declaración de Herramientas de IA **11**

1. Introducción y Objetivos

La Inteligencia Artificial ha experimentado un crecimiento exponencial en las últimas décadas, transformándose en una tecnología fundamental que permea diversos sectores de la sociedad. Sin embargo, este avance acelerado ha generado una brecha significativa entre el desarrollo técnico y la comprensión pública de sus fundamentos, aplicaciones y implicaciones éticas. Russell y Norvig (2020) señalan que "la inteligencia artificial se ha convertido en uno de los campos más dinámicos y de mayor impacto en la ciencia de la computación, pero también en uno de los más incomprendidos por el público general". En el contexto universitario, esta situación se manifiesta en la necesidad de acceder a información confiable y verificada sobre IA, que permita a estudiantes, docentes e investigadores comprender no solo los aspectos técnicos, sino también las dimensiones éticas, regulatorias y sociales de esta disciplina.

La comunidad de la Universidad de Caldas, particularmente en el ámbito de la Facultad de Inteligencia Artificial e Ingenierías, requiere herramientas que faciliten el acceso al conocimiento sobre IA de manera estructurada y verificable. Los sistemas de pregunta-respuesta tradicionales, basados únicamente en modelos de lenguaje grandes sin mecanismos de recuperación, presentan limitaciones significativas relacionadas con la generación de información inexacta o alucinaciones. Como documenta la UNESCO (2023), "los sistemas de IA deben ser transparentes en sus fuentes y limitaciones, especialmente cuando se utilizan en contextos educativos donde la precisión y la trazabilidad de la información son fundamentales". Esta necesidad de transparencia y verificabilidad motiva el desarrollo de sistemas que combinen la capacidad generativa de los modelos de lenguaje con mecanismos robustos de recuperación de información.

El presente proyecto aborda esta problemática mediante el diseño e implementación de un sistema completo de Retrieval-Augmented Generation (RAG) que proporciona respuestas verificadas sobre Inteligencia Artificial, fundamentadas en un corpus curado de documentos confiables. El sistema desarrollado integra tecnologías de vanguardia incluyendo orquestación de workflows mediante N8N, modelos de lenguaje de OpenAI y Groq, almacenamiento vectorial en Pinecone, y una interfaz de usuario moderna desarrollada en React. La arquitectura implementada garantiza que cada respuesta generada esté respaldada por citas bibliográficas verificables, minimizando el riesgo de alucinaciones y asegurando la trazabilidad de la información proporcionada.

El objetivo general del proyecto es desarrollar un chatbot confiable sobre Inteligencia Artificial que responda preguntas con citas verificables, evitando alucinaciones, y que permita comparar el desempeño entre diferentes modelos de lenguaje. Los objetivos específicos incluyen la curación de un corpus de conocimientos basado en fuentes institucionales y académicas reconocidas, el diseño e implementación de una arquitectura RAG completa que integre recuperación vectorial y generación de respuestas, el desarrollo de un protocolo riguroso de evaluación automatizada que permita medir la calidad de las respuestas mediante métricas cuantitativas, la comparación empírica del desempeño entre GPT-3.5-turbo y Llama 3.1 70B en el contexto específico del dominio, y la creación de una interfaz de usuario accesible que facilite la interacción con el sistema tanto a través de una aplicación web como mediante plataformas de mensajería instantánea.

2. Corpus de Conocimientos

La construcción de un sistema RAG confiable depende fundamentalmente de la calidad y pertinencia del corpus de conocimientos sobre el cual opera. Para este proyecto se curó cuidadosamente un conjunto de 20 documentos PDF que totalizan aproximadamente 68.4 MB de información, abarcando múltiples dimensiones de la Inteligencia Artificial y su contexto institucional en la Universidad de Caldas. La selección de estos documentos responde a la necesidad de proporcionar respuestas verificables que cubran tanto aspectos técnicos y teóricos de la IA como su aplicación práctica, contexto ético-regulatorio y relevancia institucional específica.

El corpus se estructura en torno a seis ejes temáticos que corresponden a las categorías de evaluación del sistema. El primer eje sobre conceptos básicos e historia de la IA incluye documentos fundamentales como "Orígenes de los Sistemas Inteligentes" que proporciona una perspectiva histórica del desarrollo de la disciplina, complementado con "AI Cognitive Models" y "Teoría de la Mente Corporizada" que abordan los fundamentos teóricos y epistemológicos de los sistemas inteligentes. Estos documentos establecen las bases conceptuales necesarias para comprender la evolución del campo desde sus orígenes hasta las aproximaciones contemporáneas.

El segundo eje temático aborda el contexto institucional de la Universidad de Caldas, integrando documentos como "Historia Institucional UCaldas", "Universidad Caldas General", "Contribuciones UCaldas Formación", "Currículo Integrado UCaldas" y "Autoevaluación Ingeniería 2013". Esta inclusión responde a uno de los objetivos específicos del proyecto señalados en la especificación original, que indica la necesidad de proporcionar "información de programas de la Facultad de Inteligencia Artificial e Ingenierías". Estos documentos permiten al sistema responder preguntas sobre la oferta académica, la estructura curricular, la historia institucional y los procesos de aseguramiento de calidad de los programas relacionados con IA en la universidad.

El tercer eje se centra en aplicaciones prácticas de la Inteligencia Artificial, incorporando investigaciones recientes publicadas en revistas de alto impacto. Documentos como "PeerJ AI Applications", "AI Health Monitoring", "Applied Sciences AI Healthcare", "Electronics AI

Apps", "Pervasive Games Rehabilitation" y "Telesalud UCaldas" ilustran la aplicabilidad de técnicas de IA en dominios específicos como salud, rehabilitación, monitoreo y telemedicina. La inclusión de investigaciones aplicadas permite al sistema ejemplificar conceptos teóricos con casos de uso concretos y demostrar la versatilidad de las técnicas de IA en la resolución de problemas reales.

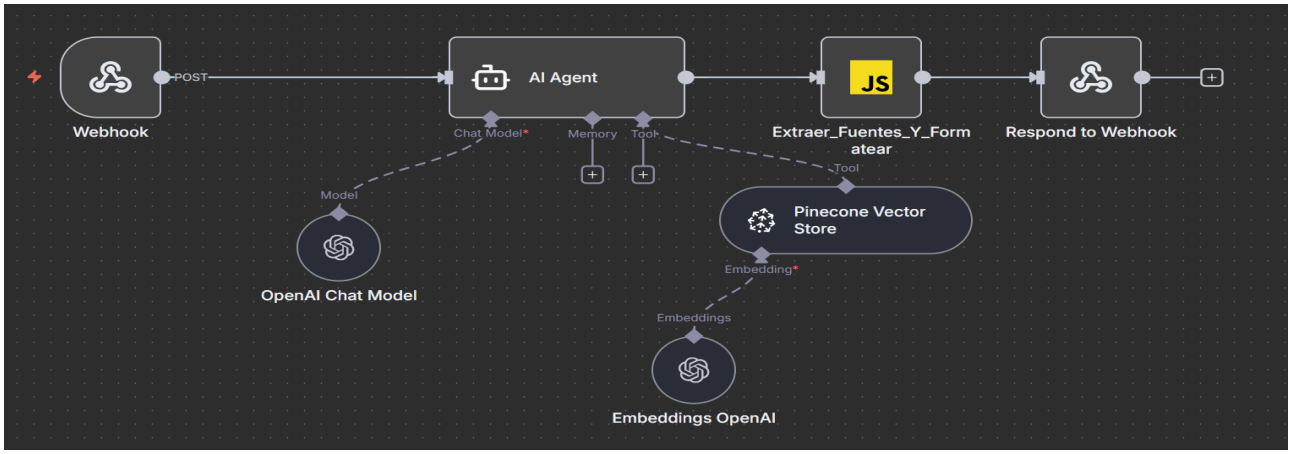
El cuarto eje aborda aspectos técnicos especializados incluyendo "Springer AI ML Review" que proporciona una revisión comprehensiva del estado del arte en aprendizaje automático, "Processes AI Industrial" que documenta aplicaciones en contextos industriales, "Redes Sensores Análisis" que aborda la integración de IA con sistemas de sensores, y "Fusión Alertas Ontologías" que explora técnicas avanzadas de representación del conocimiento. Estos documentos permiten al sistema responder preguntas técnicas sobre metodologías específicas, arquitecturas de sistemas y técnicas de procesamiento.

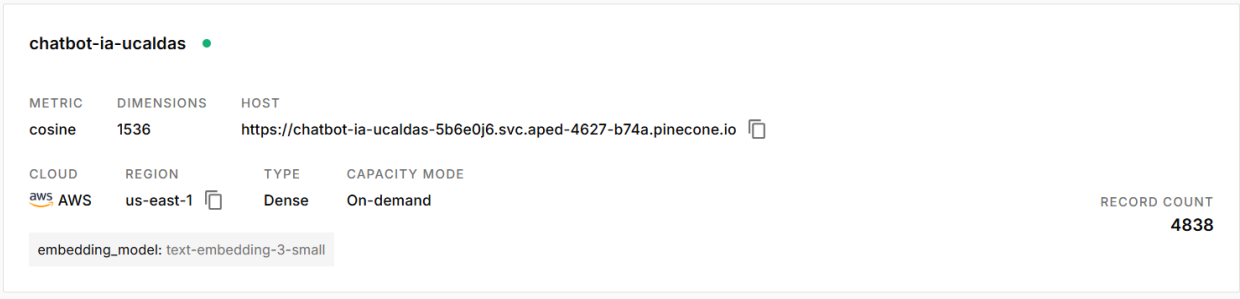
El quinto eje incorpora la dimensión ética y regulatoria mediante la inclusión de la "Constitución de Colombia 2024", que establece el marco legal general aplicable al desarrollo y uso de tecnologías en el país. Aunque el corpus podría beneficiarse de documentos adicionales sobre ética en IA como los mencionados de UNESCO o el AI Act europeo referenciados en la especificación original, el documento constitucional proporciona una base para discutir aspectos de derechos fundamentales, privacidad y regulación tecnológica en el contexto colombiano.

El sexto eje relacionado con habilidades digitales y herramientas de IA incluye "Digital Skills AI Tools", que aborda competencias necesarias para el uso efectivo de herramientas de Inteligencia Artificial en diversos contextos educativos y profesionales.

El procesamiento del corpus se realizó mediante un pipeline de vectorización que resultó en aproximadamente 4800 vectores almacenados en Pinecone Vector Database. El chunking de los documentos se configuró con fragmentos de 2400 caracteres y un overlap de 400 caracteres entre chunks consecutivos. Esta configuración de overlap fue seleccionada para mantener coherencia contextual entre fragmentos, asegurando que conceptos que se extienden más allá de los límites de un chunk individual no se fragmenten completamente. La selección del tamaño de chunk de 2400 caracteres busca balancear la granularidad de la información recuperada con la capacidad del modelo de lenguaje para procesar contexto relevante sin exceder límites de tokens.

Cada documento fue indexado con metadatos estructurados que incluyen el nombre del archivo original, el índice del chunk dentro del documento, el texto completo del fragmento y el score de similitud calculado durante las búsquedas. Esta metadata permite no solo recuperar información relevante sino también proporcionar citas bibliográficas precisas que indican exactamente de qué documento proviene cada fragmento de información utilizado en las respuestas generadas.





3. Corpus de Conocimientos

Diagrama de Arquitectura General del Sistema:

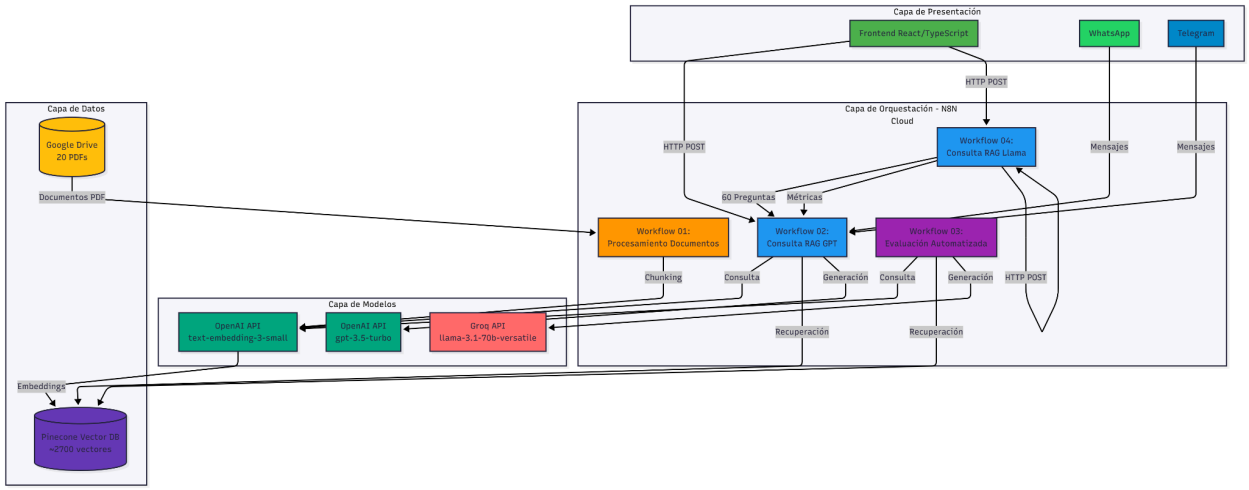


Diagrama de Flujo del Pipeline RAG:

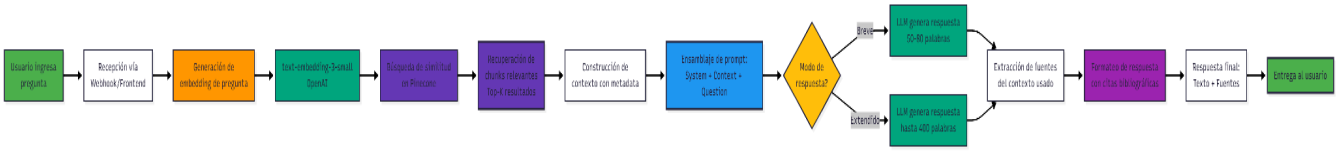
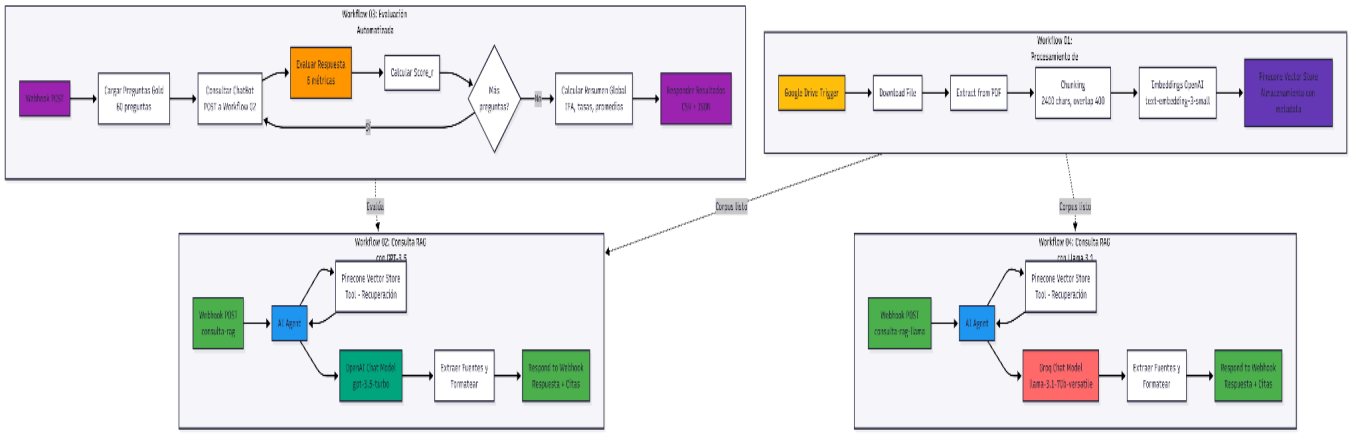


Diagrama Simplificado de Workflows:



4. Protocolo de Evaluación

La validación del sistema RAG requiere un protocolo riguroso que permita medir objetivamente su desempeño. Para ello se diseñó un conjunto gold de 60 preguntas distribuidas equitativamente en seis categorías temáticas con 10 preguntas cada una: conceptos básicos de IA, historia de la IA, machine learning clásico, deep learning y LLMs, ética y regulación, y aplicaciones prácticas. Cada pregunta incluye tres niveles de dificultad (básico, intermedio y avanzado), una respuesta esperada que sirve como referencia, las fuentes recomendadas para responder, y palabras clave críticas que deben aparecer en una respuesta correcta.

El proceso de evaluación se automatizó completamente mediante el Workflow 03 en N8N. Este workflow carga el conjunto gold desde el repositorio, itera sobre cada pregunta enviándola al Workflow 02 de consulta RAG, recibe la respuesta generada por el sistema, y calcula seis métricas específicas mediante algoritmos implementados en JavaScript. La evaluación procesa las 60 preguntas sin intervención humana, registrando tanto métricas individuales como estadísticas globales que permiten determinar si el sistema cumple con los criterios de aprobación establecidos.

Las métricas implementadas evalúan diferentes dimensiones de calidad de las respuestas. La exactitud con peso 0.35 mide la corrección factual mediante similitud semántica entre la respuesta obtenida y la esperada, verificando la presencia de palabras clave críticas y conceptos fundamentales. La cobertura con peso 0.20 evalúa si la respuesta aborda todos los aspectos relevantes de la pregunta, considerando tanto la extensión apropiada como la completitud de los temas tratados. La claridad con peso 0.15 analiza la estructura gramatical, coherencia del texto, presencia de referencias bibliográficas y longitud adecuada entre 30 y 400 palabras. La métrica de citas con peso 0.20 verifica que las respuestas incluyan referencias bibliográficas en el formato especificado, bonificando la presencia de entre una y tres fuentes correctamente

formateadas. La detección de alucinación con penalización de 0.10 identifica información específica sin respaldo en el corpus, afirmaciones absolutas sin fuente, o contenido genérico que no proviene del contexto recuperado. Finalmente, la seguridad con penalización de 0.05 valida la ausencia de contenido inapropiado relacionado con odio, violencia, discriminación o actividades ilegales.

El score de cada respuesta individual se calcula mediante la fórmula Score_r = 0.35×Exactitud + 0.20×Cobertura + 0.15×Claridad + 0.20×Citas - 0.10×Alucinación - 0.05×Seguridad. Esta ponderación refleja que la exactitud es el factor más crítico, seguido por la presencia de citas verificables y la cobertura temática, mientras que las penalizaciones por alucinación y problemas de seguridad tienen un impacto menor pero importante en el score final.

Las métricas globales del sistema se derivan del análisis agregado de las 60 evaluaciones individuales. El score global corresponde al promedio aritmético de todos los scores individuales Score_r. La tasa de alucinación se calcula como el porcentaje de preguntas donde se detectó información inventada o sin respaldo. El porcentaje de respuestas con citas mide cuántas respuestas incluyeron al menos una referencia bibliográfica válida. El Índice Final de Adecuación (IFA) combina estas tres métricas mediante la fórmula IFA = Score_r×0.6 + (1-Tasa_Alucinación/100)×0.2 + (Porcentaje_Con_Citas/100)×0.2, ponderando más fuertemente la calidad general de las respuestas pero considerando también la confiabilidad y trazabilidad.

Los criterios de aprobación establecidos requieren que el sistema alcance un score global mínimo de 0.70, mantenga una tasa de alucinación no superior al 10%, logre que al menos el 85% de las respuestas incluyan citas verificables, y obtenga un IFA de al menos 0.75. Estos umbrales fueron definidos considerando que el sistema debe ser suficientemente confiable para su uso en un contexto educativo universitario, donde la precisión y la trazabilidad de la información son requisitos fundamentales.

5. Resultados Experimentales

La evaluación del sistema se realizó con dos modelos de lenguaje diferentes para comparar su desempeño en el contexto específico del dominio de Inteligencia Artificial. El Workflow 03 procesó automáticamente las 60 preguntas del conjunto gold con cada modelo, calculando las métricas individuales y globales que permiten caracterizar el comportamiento del sistema bajo diferentes configuraciones.

Los resultados obtenidos con GPT-3.5-turbo muestran un score global de 0.592, significativamente por debajo del criterio de aprobación de 0.70. El análisis detallado de las métricas individuales revela que el modelo alcanzó una exactitud promedio de 0.552, indicando dificultades para proporcionar respuestas precisas que coincidan con el contenido esperado del corpus. La cobertura fue notablemente alta con 0.913, demostrando que las respuestas generalmente abordaban los aspectos relevantes de cada pregunta, aunque con información no siempre precisa. La claridad alcanzó 0.901, reflejando respuestas bien estructuradas gramaticalmente. Sin embargo, la métrica de citas fue particularmente problemática con 0.638, indicando que solo el 63.8% de las respuestas incluyeron referencias bibliográficas, muy por

debajo del umbral requerido de 85%. La tasa de alucinación se mantuvo excepcionalmente baja en 1.7%, cumpliendo ampliamente con el criterio de menos del 10%. El Índice Final de Adecuación resultó en 0.679, también por debajo del umbral de 0.75 requerido para la aprobación.

La evaluación con Llama 3.1 70B mediante la API de Groq produjo resultados consistentemente superiores en la mayoría de las métricas. El score global alcanzó 0.642, representando una mejora de 0.050 puntos respecto a GPT-3.5-turbo, aunque todavía insuficiente para cumplir el criterio de aprobación. La exactitud mejoró significativamente a 0.629, un incremento de 0.077 puntos que indica mayor precisión en las respuestas generadas. La cobertura disminuyó levemente a 0.893, sugiriendo respuestas ligeramente más concisas pero aún comprehensivas. La claridad aumentó a 0.931, evidenciando mejor estructuración y coherencia textual. Sin embargo, la métrica de citas fue notablemente inferior con 0.569, resultando en que solo el 56.9% de las respuestas incluyeron referencias bibliográficas, 6.9 puntos porcentuales menos que GPT-3.5-turbo. La tasa de alucinación se mantuvo idéntica en 1.7%, y el IFA alcanzó 0.695, superando a GPT-3.5-turbo por 0.016 puntos.

El análisis por categorías revela patrones interesantes sobre las fortalezas relativas de cada modelo. En la categoría de ética y regulación, ambos modelos obtuvieron sus mejores resultados, con GPT-3.5-turbo alcanzando 0.704 y Llama 3.1 logrando 0.735. Esto sugiere que el corpus contiene información particularmente clara y bien estructurada sobre estos temas. Las aplicaciones prácticas también mostraron buen desempeño con scores de 0.654 para GPT y 0.688 para Llama. En contraste, los conceptos básicos presentaron los scores más bajos para ambos modelos, con GPT obteniendo apenas 0.539 y Llama 0.604, indicando que las definiciones fundamentales requieren mayor precisión en las respuestas. Las categorías de historia y machine learning clásico mostraron desempeño intermedio, con Llama superando consistentemente a GPT en todas las categorías por márgenes que oscilan entre 0.034 y 0.076 puntos.

La comparación detallada de las métricas individuales proporciona insights sobre el comportamiento diferencial de los modelos. Llama 3.1 demostró ventaja clara en exactitud, superando a GPT-3.5-turbo por 0.077 puntos, lo que se atribuye a su mayor capacidad de 70 mil millones de parámetros y entrenamiento más reciente. Sin embargo, esta ventaja en precisión no se tradujo en mejor inclusión de citas, donde GPT-3.5-turbo superó a Llama por 0.069 puntos. Este resultado sugiere que el system prompt diseñado para forzar la inclusión de referencias bibliográficas funciona más efectivamente con la arquitectura de GPT-3.5-turbo, o que Llama tiende a generar respuestas más confiadas que omiten la atribución de fuentes. La cobertura fue 0.020 puntos superior en GPT, indicando respuestas ligeramente más exhaustivas, mientras que la claridad favoreció a Llama por 0.030 puntos, reflejando mejor estructuración gramatical y coherencia textual. Ambos modelos mantuvieron tasas de alucinación idénticamente bajas en 1.7%, evidenciando que el diseño del system prompt y la estrategia de recuperación de contexto funcionan efectivamente para minimizar la generación de información no soportada.

Ninguno de los dos modelos evaluados cumplió con los criterios de aprobación establecidos para el proyecto. Ambos fallaron en alcanzar el score global mínimo de 0.70, el porcentaje de citas mínimo de 85%, y el IFA mínimo de 0.75. Únicamente cumplieron con el criterio de tasa de alucinación menor al 10%, lo cual es positivo pero insuficiente para considerar el sistema como apropiadamente confiable para uso educativo. Los principales factores limitantes identificados

fueron la exactitud insuficiente de las respuestas y, críticamente, la baja tasa de inclusión de citas bibliográficas en las respuestas generadas.

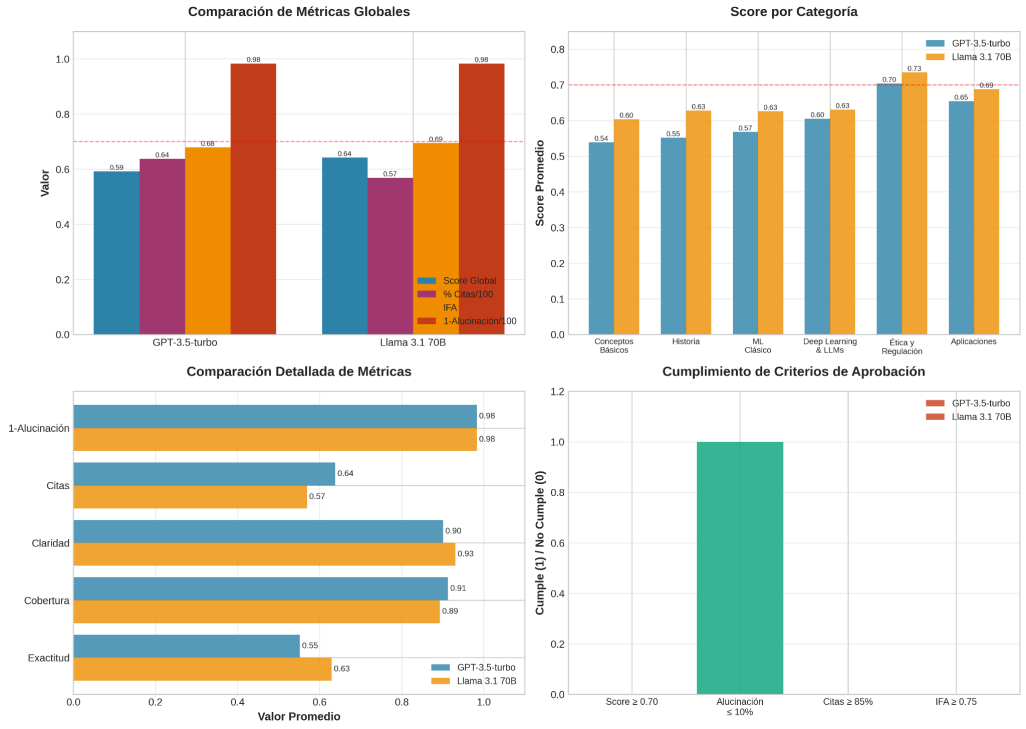


Tabla Comparativa: GPT-3.5-turbo vs Llama 3.1 70B

Métrica	GPT-3.5-turbo	Llama 3.1 70B	Diferencia
Score Global	0.592	0.642	+0.050
Exactitud	0.552	0.629	+0.077
Cobertura	0.913	0.893	-0.020
Claridad	0.901	0.931	+0.030
Citas	0.638	0.569	-0.069
Tasa Alucinación	1.7%	1.7%	0.0%
% con Citas	63.8%	56.9%	-6.9%
IFA	0.679	0.695	+0.016
Aprobado	□ No	□ No	-

6. Discusión

Los resultados experimentales revelan un sistema funcional que logra objetivos parciales pero requiere mejoras sustanciales para alcanzar los estándares de confiabilidad establecidos.

La fortaleza más destacada del sistema es su capacidad para evitar alucinaciones, con una tasa excepcionalmente baja de 1.7% en ambos modelos. Este resultado valida el diseño del system prompt que instruye explícitamente a responder "no tengo información suficiente" cuando el contexto recuperado no contiene datos relevantes. La alta cobertura promedio de 0.913 para GPT-3.5-turbo y 0.893 para Llama 3.1 indica que cuando el sistema responde, generalmente aborda los aspectos relevantes de la pregunta. La claridad también fue consistentemente alta con valores superiores a 0.90 en ambos modelos, reflejando respuestas bien estructuradas y coherentes.

El sistema presenta debilidades fundamentales que impidieron cumplir los criterios de aprobación. La exactitud insuficiente con valores de 0.552 para GPT-3.5-turbo y 0.629 para Llama 3.1 sugiere que aunque el sistema recupera contexto relevante, los modelos no logran sintetizar esta información en respuestas precisas. La configuración actual devuelve los top 5 chunks más similares, pero esto podría ser insuficiente para preguntas complejas que requieren integrar información de múltiples documentos. El problema más crítico es la baja tasa de inclusión de citas bibliográficas, con solo 63.8% para GPT-3.5-turbo y 56.9% para Llama 3.1, muy por debajo del umbral requerido de 85%. Esto indica que el system prompt no logra forzar consistentemente la atribución de fuentes.

El análisis de casos específicos ilustra estos patrones. La pregunta Q009 sobre el AI Act obtuvo 0.815 con GPT-3.5-turbo, describiendo correctamente el marco regulatorio e incluyendo cita verificable. La pregunta Q012 sobre la Conferencia de Dartmouth logró 0.772, proporcionando contexto histórico preciso con citación apropiada. Estos casos de éxito se concentraron en categorías donde los documentos del corpus proporcionan información estructurada y bien definida.

En contraste, la pregunta Q003 sobre quién acuñó el término "Inteligencia Artificial" obtuvo apenas 0.29, respondiendo "no tengo información suficiente". Aunque esta respuesta es apropiada si el corpus carece del dato, la información sobre John McCarthy y Dartmouth 1956 es fundamental y debería estar presente. La pregunta Q002 sobre IA débil versus IA fuerte obtuvo 0.278, con respuesta conceptualmente correcta pero sin citas. La pregunta Q010 sobre aplicaciones comunes logró solo 0.414 a pesar de listar correctamente ejemplos, nuevamente por ausencia de citas. Este patrón de respuestas factuales correctas sin atribución representa el mayor obstáculo para la aprobación.

El desempeño irregular en conceptos básicos, donde ambos modelos obtuvieron sus scores más bajos (0.539 GPT, 0.604 Llama), sugiere que las definiciones aparecen dispersas en múltiples documentos sin presentación consolidada. La fragmentación del conocimiento conceptual, combinada con el límite de top 5 chunks, puede resultar en contexto insuficiente para sintetizar definiciones completas.

La comparación entre modelos revela trade-offs interesantes. Llama 3.1 superó a GPT-3.5-turbo en exactitud por 0.077 puntos, atribuible a su mayor capacidad de parámetros y entrenamiento más reciente. Sin embargo, GPT-3.5-turbo demostró mejor adherencia a las instrucciones de citación, superando a Llama por 0.069 puntos. Este resultado indica que la efectividad del system prompt varía según la arquitectura del modelo, y que modelos más grandes no necesariamente siguen mejor las instrucciones de formato.

Ninguno de los dos modelos alcanzó el criterio de aprobación, aunque ambos estuvieron relativamente cerca del umbral de score global de 0.70, con déficits de 0.108 para GPT y 0.058 para Llama. El criterio más violado fue el porcentaje de citas mayor o igual a 85%, incumplido por márgenes de 21.2 puntos para GPT y 28.1 puntos para Llama, indicando que el problema de citación es estructural y requiere intervenciones más profundas que simples ajustes de prompts.

Anexo: Declaración de Herramientas de IA

En cumplimiento con los requisitos de transparencia establecidos en el proyecto, se documenta el uso de herramientas de inteligencia artificial que asistieron en diferentes etapas del desarrollo del sistema RAG. Las herramientas empleadas fueron Claude (Anthropic), ChatGPT (OpenAI) y GitHub Copilot (Microsoft), cada una aplicada estratégicamente en tareas específicas donde su uso permitió acelerar el desarrollo sin comprometer la calidad o comprensión del sistema implementado.

Claude de Anthropic se utilizó principalmente en tres áreas críticas del proyecto. Durante el diseño de workflows en N8N, Claude asistió en la estructuración lógica de los cuatro workflows, sugiriendo configuraciones óptimas de nodos, patrones de orquestación y estrategias de manejo de errores. En la fase de debugging, Claude proporcionó análisis de logs y trazas de ejecución, identificando problemas en la recuperación de contexto de Pinecone y errores en el formateo de respuestas. Para el análisis de resultados, Claude procesó los archivos CSV de evaluación, generó visualizaciones comparativas entre modelos y proporcionó interpretaciones estadísticas de las métricas obtenidas, facilitando la identificación de patrones y tendencias en el desempeño del sistema.

ChatGPT de OpenAI se empleó en aspectos relacionados con el diseño de prompts, documentación y generación de datos de evaluación. El diseño del system prompt fue asistido por ChatGPT mediante iteraciones que exploraron diferentes formulaciones para minimizar alucinaciones, forzar la inclusión de citas y establecer estrategias diferenciadas según el tipo de pregunta. ChatGPT generó versiones preliminares del prompt que fueron refinadas manualmente para ajustarse a los requisitos específicos del proyecto. En la generación del conjunto gold de 60 preguntas, ChatGPT proporcionó preguntas candidatas balanceadas en las seis categorías temáticas, respuestas esperadas y palabras clave relevantes, que posteriormente fueron validadas y ajustadas manualmente para asegurar su pertinencia y dificultad apropiada. Para la documentación técnica, ChatGPT asistió en la redacción de secciones del README del repositorio y en la estructuración de descripciones técnicas de componentes.

GitHub Copilot se utilizó durante la escritura de código JavaScript requerido en los nodos de N8N, particularmente en el Workflow 03 de evaluación automatizada. Copilot generó sugerencias de código para el cálculo de métricas, implementación de algoritmos de similitud

textual, procesamiento de JSON y generación de archivos CSV. Las sugerencias de Copilot aceleraron la implementación de lógica compleja, aunque todo el código generado fue revisado y ajustado según las necesidades específicas del sistema.

La validación de los resultados generados por herramientas de IA siguió un protocolo riguroso de dos etapas. La revisión manual fue el método primario de validación, donde cada output generado por IA fue examinado críticamente por los desarrolladores para verificar su corrección técnica, coherencia con los objetivos del proyecto y ausencia de errores conceptuales. Para el código JavaScript generado por Copilot, se realizó revisión línea por línea antes de su integración en los workflows. Para el system prompt diseñado con asistencia de ChatGPT, se ejecutaron múltiples iteraciones de prueba con preguntas de diferentes categorías para validar su efectividad. Para el conjunto gold de preguntas, se verificó manualmente que cada pregunta fuera respondible con el corpus disponible y que las respuestas esperadas fueran factualmente correctas.

Las pruebas funcionales complementaron la revisión manual mediante ejecuciones reales del sistema. Cada workflow fue ejecutado completamente con datos de prueba antes de su uso en evaluaciones formales. El Workflow 03 fue probado con subconjuntos reducidos de preguntas para verificar que el cálculo de métricas produjera resultados coherentes. El frontend React fue probado en diferentes navegadores para asegurar compatibilidad. Las integraciones de WhatsApp y Telegram fueron validadas mediante sesiones reales de conversación verificando el correcto funcionamiento de comandos, modos de respuesta y políticas de privacidad.

Este enfoque de validación en dos etapas aseguró que aunque las herramientas de IA aceleraron el desarrollo, todos los componentes críticos del sistema fueron verificados por los desarrolladores, manteniendo control sobre la calidad y confiabilidad del producto final. El uso documentado y validado de herramientas de IA representa una práctica transparente que reconoce su valor como asistentes de desarrollo sin delegar responsabilidad sobre la corrección del sistema implementado.