Evaluation of Local Language Models in a RAG System for Educational Assistance

Damián Barsotti[1], Luis Biedma[1], Manuel Díaz Ferreiro[2], Mariano Lescano[2], and Nahuel Seiler[1]

[1] Facultad de Matemática, Astronomía, Física y Computación
Universidad Nacional de Córdoba
Ciudad Universitaria, X5000HUA Córdoba, Argentina
https://www.famaf.unc.edu.ar
{damian.barsotti,lbiedma}@unc.edu.ar, nahuelseiler@gmail.com
[2] Procer Tecnologías S.A.S.
Belgrano 66 4°B, X5000AHM, Córdoba
https://www.procertecnologias.com/
{manu,mariano}@procertecnologias.com

Abstract. This paper presents the collaborative development between the Faculty of Mathematics, Astronomy, Physics and Computing (FAMAF) at the National University of Córdoba and the company Procer Tecnologías S.A.S. to build a Retrieval-Augmented Generation (RAG) system for educational assistance. The system is designed to run on a local computer to which a portable assistive device (developed by the company) connects, replacing the RAG system currently deployed in the cloud. Its objective is to improve autonomous access to school content for students with visual impairment, dyslexia, or other learning barriers.

The implemented architecture combines vector storage, hybrid search, a reranking stage, and response generation with language models executed locally via Ollama. Ten compact models were evaluated using a dataset of 396 Spanish question-answer pairs, built from real documents and users of the company's cloud service.

Results show that hybrid search with reranking consistently improves over pure vector search across all five metrics and all evaluated models. Notably, the evaluated models are compact enough to run on consumer-grade GPUs, yet still achieve competitive metrics relative to the company's current system.

Keywords: Retrieval-Augmented Generation , Local Language Models , Hybrid Search , Model Evaluation , Educational Inclusion

Evaluación de modelos de lenguaje locales en un sistema RAG para asistencia educativa

Resumen Este trabajo presenta el desarrollo colaborativo entre la Facultad de Matemática, Astronomía, Física y Computación (FAMAF) de la Universidad Nacional de Córdoba y la empresa Procer Tecnologías S.A.S. para construir un sistema de Retrieval-Augmented Generation (RAG) de asistencia educativa. El sistema está diseñado para ejecutarse en una computadora local a la que se conecta un dispositivo portátil de asistencia diseñado por la empresa, reemplazando el sistema RAG que tiene actualmente desplegado en la nube. Su objetivo es mejorar el acceso autónomo a contenidos escolares para estudiantes con discapacidad visual, dislexia u otras barreras de aprendizaje.

La arquitectura implementada combina almacenamiento vectorial, búsqueda híbrida, proceso de reranking y generación de respuestas con modelos de lenguaje ejecutados localmente vía Ollama. Se evaluaron 10 modelos reducidos utilizando un dataset de 396 pares pregunta-respuesta en español, construido a partir de documentos y usuarios reales del servicio en la nube de la empresa.

Los resultados muestran que la búsqueda híbrida con reranking mejora consistentemente a la búsqueda vectorial pura en las cinco métricas y en todos los modelos evaluados. Notablemente, los modelos evaluados son de tamaño reducido pudiéndose ejecutar en GPUs de uso hogareño, y aun así obtienen métricas competitivas tomando como referencia las respuestas del sistema que la empresa tiene actualmente.

Keywords: RAG , Modelos de lenguaje locales , Búsqueda híbrida , Evaluación de modelos , Inclusión educativa

1. Motivación y Problemática

El acceso equitativo a la educación representa un desafío crítico en Argentina y Latinoamérica. Según datos del INDEC, el 10,2 % de la población argentina de seis años o más presenta algún tipo de dificultad o discapacidad, lo que incluye a aproximadamente 70.000 niños con deficiencias visuales. A nivel regional, UNICEF estima que 19,1 millones de niños y adolescentes con discapacidad ven limitado su potencial debido a barreras en el entorno educativo. Estas brechas tienen consecuencias estructurales: los jóvenes con discapacidad en la región tienen una probabilidad un 23 % menor de completar la educación secundaria en comparación con sus pares.

Desde una perspectiva técnica, las herramientas actuales de asistencia (lectores de pantalla o aplicaciones convencionales) presentan limitaciones significativas al procesar contenidos multiformato complejos, como gráficos o tablas. Asimismo, la adopción de modelos de lenguaje comerciales suele carecer de mecanismos adecuados de trazabilidad, incrementando el riesgo de “alucinaciones”

que comprometen la calidad pedagógica. Ante este escenario, se identifica la necesidad urgente de desarrollar sistemas de IA que garanticen precisión, autonomía y confianza en el aula.

El presente trabajo surge de la colaboración estratégica entre la Facultad de Matemática, Astronomía, Física y Computación (FAMAF–UNC) y la empresa de base tecnológica Procer Tecnologías S.A.S. La FAMAF, con una trayectoria consolidada en investigación básica y aplicada en ciencias de la computación, canaliza su interacción con el sector productivo a través de su Secretaría de Innovación y Vinculación Tecnológica (SIVT), cuya misión es aplicar el conocimiento científico a la resolución de problemas regionales complejos. Por su parte, Prócer Tecnologías se especializa en el desarrollo de soluciones de Inteligencia Artificial (IA) para la asistencia educativa. Esta sinergia busca transformar el acceso a la información mediante la evolución de herramientas de asistencia tradicionales hacia sistemas interactivos de consulta y respuesta, optimizados para hardware específico en entornos de aprendizaje.

El trabajo se centra en el diseño e implementación de un motor de Generación Aumentada por Recuperación (RAG) textual. Si bien existen soluciones basadas en servicios de nube propietarios, la propuesta de co-creación con FAMAF busca un salto cualitativo hacia la soberanía tecnológica. De esta forma, el objetivo principal es el desarrollo de una arquitectura de IA controlada localmente, que permita la ingesta, indexación y consulta de materiales educativos cargados por el usuario. Este enfoque no solo asegura la propiedad intelectual de un activo tecnológico central, sino que permite la personalización avanzada del motor para hardware de bajos recursos y la implementación de mecanismos estrictos de trazabilidad y citas en las respuestas generadas.

2. Trabajos relacionados

Los sistemas de Recuperación Aumentada por Generación (RAG) fueron introducidos por Lewis et al. [12], quienes propusieron combinar un módulo de recuperación de documentos con un modelo de lenguaje generativo para producir respuestas fundamentadas en evidencia documental. Desde entonces, la arquitectura RAG se ha consolidado como el enfoque dominante para la construcción de sistemas de preguntas y respuestas sobre bases de conocimiento privadas [7]. Su aplicación en entornos educativos resulta especialmente atractiva: permite responder preguntas sobre los propios materiales del estudiante sin reentrenamiento del modelo, adaptándose a cualquier corpus de documentos de forma dinámica.

La calidad del retrieval es un factor determinante en el rendimiento de los sistemas RAG. Los *embeddings* son representaciones vectoriales densas de texto en un espacio de alta dimensión, donde la proximidad geométrica entre vectores refleja la similitud semántica entre los fragmentos que representan. La búsqueda vectorial mediante estos *embeddings* captura dicha similitud semántica, pero puede fallar ante consultas léxicamente específicas. Para superar esta limitación, los enfoques híbridos combinan búsqueda densa con búsqueda léxica (BM25 o full-text search), fusionando los rankings mediante técnicas como Reciprocal Rank

Fusion (RRF) [4], que mejora consistentemente la recuperación sin necesidad de parámetros adicionales de entrenamiento.

El reranking mediante cross-encoders representa una segunda etapa de refinamiento. A diferencia de los bi-encoders, que codifican consulta y documento de forma independiente, los cross-encoders como BAAI/bge-reranker-v2-m3 [3] evalúan cada par (consulta, fragmento) conjuntamente, produciendo scores de relevancia más precisos a costa de mayor tiempo de cómputo.

La evaluación de sistemas RAG presenta desafíos propios. El framework RAGAS [6] aborda este problema mediante métricas automatizadas basadas en LLM, complementadas con métricas de similitud textual como BLEU [19], ChrF [20] y ROUGE-L [13], que permiten comparar respuestas generadas con respuestas de referencia de forma reproducible. La evaluación en idiomas distintos al inglés introduce desafíos adicionales, ya que los modelos de embeddings y los LLMs exhiben diferente comportamiento en español; el presente trabajo aborda esta brecha con un dataset construido íntegramente en español a partir de documentos y usuarios reales.

En el contexto de despliegue local, la cuantización de modelos mediante técnicas como GGUF [9] ha permitido ejecutar LLMs en hardware *commodity* con pérdida mínima de calidad. Plataformas como Ollama [27] abstraen esta complejidad y habilitan sistemas de IA soberanos sin dependencia de servicios en la nube, abriendo la puerta a despliegues en entornos con conectividad limitada o requisitos estrictos de privacidad de datos.

3. Metodología del Proyecto

El sistema fue desarrollado en el marco de una colaboración de cuatro meses entre FAMAF-UNC y Procer Tecnologías S.A.S., formalizada mediante un convenio de colaboración técnica [25] y financiada parcialmente a través del programa Voucher de Innovación Colaborativa del Gobierno de la Provincia de Córdoba [26].

El equipo de FAMAF aportó experiencia en sistemas de recuperación de información, aprendizaje automático y procesamiento del lenguaje natural. El equipo de Procer contribuyó con el conocimiento del dominio educativo, los requerimientos de hardware del dispositivo Procer 4 [21] y el acceso al corpus de documentos y usuarios reales del servicio en la nube de la empresa. Ambos equipos se coordinaron mediante reuniones periódicas de seguimiento para validar avances y ajustar requerimientos.

El trabajo se estructuró en tres fases incrementales: (1) análisis de requerimientos y selección del modelo de *embeddings* para español; (2) implementación del pipeline de ingesta, construcción de la base vectorial e integración de la búsqueda híbrida con reranking; (3) pruebas de integración, evaluación del sistema y entrega del MVP funcional para usar con el dispositivo Procer 4.

4. Arquitectura del sistema

El sistema se diseñó bajo una premisa de soberanía tecnológica, permitiendo la ejecución de componentes críticos de forma local o controlada. La Figura 1 ilustra los dos flujos principales del sistema: la ingesta de documentos y el procesamiento de consultas.

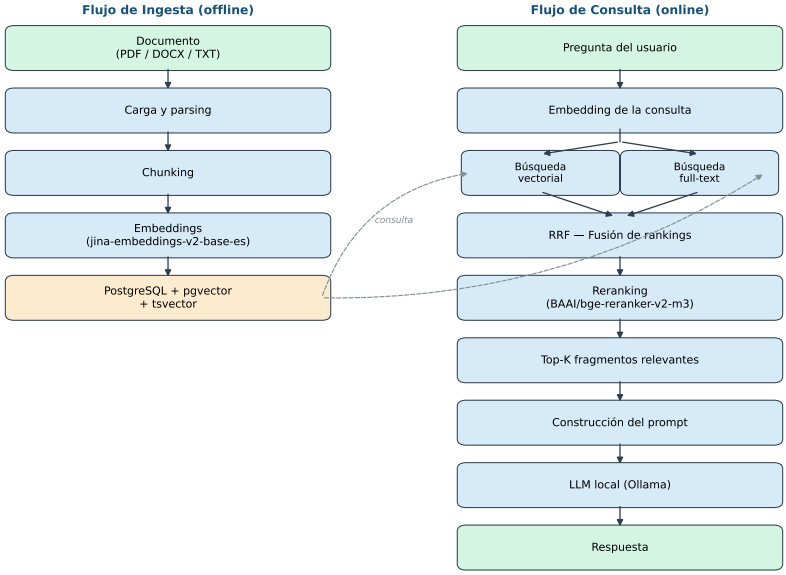

Figura 1. Pipeline del sistema Procer RAG: flujo de ingesta (izquierda) y flujo de consulta (derecha).

El sistema se estructura en siete componentes. La **API REST**, implementada con FastAPI [24], expone los endpoints para consultas, búsqueda e ingesta de documentos. La **orquestación RAG**, coordinada por LangChain [2], gestiona el pipeline completo de *retrieval* y generación. El **almacenamiento vectorial** se realiza sobre PostgreSQL + PGVector [11], que almacena y permite la búsqueda de *embeddings* de 768 dimensiones generados por el modelo Jina Spanish/English [16]. La **búsqueda full-text** utiliza el índice *tsvector* nativo de PostgreSQL para recuperar documentos por coincidencia de palabras clave. El **ranking** fusiona ambas búsquedas mediante RRF [4] y reordena los chunks por relevancia con el cross-encoder BAAI/bge-reranker-v2-m3 [3]. Finalmente, el **LLM**, ejecutado localmente vía Ollama [27] con modelo configurable en tiempo de despliegue, genera las respuestas al usuario.

La plataforma ofrece dos modos de búsqueda: vectorial y híbrida. En el modo vectorial, PGVector recupera los chunks semánticamente más similares a la consulta mediante similitud de coseno sobre los *embeddings*. En el modo híbrido, se combina esta búsqueda semántica con la búsqueda de texto completo de PostgreSQL (tsvector), que indexa el contenido de los chunks como vectores de palabras clave y recupera aquellos que contienen los términos exactos de la consulta. Ambos rankings se fusionan mediante RRF, seguido de reranking con BAAI/bge-reranker-v2-m3, mejorando la cobertura de recuperación respecto a cualquiera de los dos métodos por separado.

El sistema implementa una arquitectura *multi-tenant*: cada usuario dispone de una colección de documentos propia y aislada en las peticiones a la API. De este modo, un usuario no puede acceder a los documentos de otro, garantizando la privacidad y el aislamiento de la información entre los distintos usuarios de la plataforma.

5. Diseño de evaluación

La evaluación del sistema se realizó mediante el framework RAGAS [6], diseñado específicamente para medir la calidad de las respuestas generadas en términos de fidelidad factual, relevancia y similitud con respecto al contexto recuperado. Se utilizó como corpus los documentos y usuarios reales que Procer emplea en su sistema en la nube, garantizando que la evaluación refleje condiciones de uso reales.

El dataset de evaluación fue construido en tres etapas: (1) generación sintética de preguntas mediante GPT-5.2 [18] a partir de cada documento; (2) generación de la respuesta de referencia (*golden answer*) por el sistema en la nube de Procer; y (3) revisión manual por parte de los usuarios para validar pertinencia y corrección. Cada documento aporta entre 5 y 6 pares pregunta-respuesta. Las preguntas son de tipo comprensivo y analítico en español: requieren relacionar información distribuida en el documento y van más allá de la recuperación literal de fragmentos. Los documentos abarcan materiales educativos heterogéneos: apuntes de clase, textos académicos, guías de estudio y material de lectura complementaria.

El dataset resultante se compone de:

- **Usuarios:** 31 perfiles activos.
- **Documentos:** 76 archivos en español (apuntes, textos académicos y guías escolares).
- **Dataset de evaluación:** 396 pares de pregunta-respuesta generados y revisados manualmente.

Se evaluaron 10 modelos de lenguaje ejecutados localmente vía Ollama con cuantización Q4_K_M, ordenados por cantidad de parámetros en la Tabla 1.

Métricas utilizadas

Se emplearon dos categorías de métricas de evaluación:

Tabla 1. Modelos evaluados con cuantización Q4_K_M.

Modelo	Params	VRAM aprox.
Gemma 3 [8]	4B	3.3 GB
Mistral [10]	7B	4.4 GB
Llama 3 [14]	8B	4.7 GB
DeepSeek R1 [5]	8B	5.2 GB
Qwen 3 [22]	8B	5.2 GB
Mistral Nemo [15]	12B	7.1 GB
Gemma 3 [8]	12B	8.1 GB
DeepSeek R1 [5]	14B	9.0 GB
Phi-4 [1]	14B	9.1 GB
Qwen 3 [22]	14B	9.3 GB

Métricas basadas en LLM (via RAGAS):

- **Factual Correctness (F1) [23]:** mide la precisión factual comparando hechos extraídos de la respuesta generada contra la respuesta de referencia mediante un LLM evaluador.
- **Answer Accuracy [17]:** evalúa la calidad general de la respuesta generada respecto a la referencia.

Métricas de similitud textual:

- **BLEU [19]:** precisión de n-gramas respecto a la referencia.
- **ChrF [20]:** F-score a nivel de caracteres entre respuesta y referencia.
- **ROUGE-L [13]:** subsecuencia común más larga entre respuesta y referencia.

6. Resultados

Se evaluaron 10 modelos de lenguaje de distintos tamaños ejecutados localmente con Ollama, bajo dos modalidades de búsqueda. Los valores en **negrita** indican el mejor resultado por métrica.

Búsqueda vectorial

La Tabla 2 muestra los 10 modelos evaluados ordenados por cantidad de parámetros.

Con búsqueda vectorial pura, los 10 modelos se agrupan en un rango estrecho de Factual Correctness (0.491–0.531), con Qwen 3 8B liderando en esa métrica. La diferencia entre el mejor y el peor modelo es de apenas 0.040 puntos, lo que indica que el cuello de botella no es el LLM sino la calidad del retrieval.

Búsqueda híbrida + reranking

La Tabla 3 muestra los 10 modelos evaluados con búsqueda híbrida (vector + full-text con RRF) seguida de reranking con BAAI/bge-reranker-v2-m3, ordenados por cantidad de parámetros.

Tabla 2. Modelos evaluados – Búsqueda vectorial.

Modelo	F. Correct.	Ans. Acc.	BLEU	ChrF	ROUGE-L
Gemma 3 (4B)	0.4913	0.5587	0.1887	0.4651	0.3715
Mistral (7B)	0.5194	0.5865	0.1767	**0.4764**	0.3226
DeepSeek R1 (8B)	0.5027	0.5726	0.1372	0.4280	0.3102
Llama 3 (8B)	0.5255	0.5720	0.1460	0.4611	0.3257
Qwen 3 (8B)	**0.5313**	**0.6496**	0.1570	0.4489	0.2904
Gemma 3 (12B)	0.5113	0.5448	**0.1935**	0.4758	**0.3881**
Mistral Nemo (12B)	0.5087	0.5612	0.1738	0.4581	0.3315
DeepSeek R1 (14B)	0.5019	0.5997	0.1431	0.4405	0.2869
Phi-4 (14B)	0.5198	0.6004	0.1701	0.4613	0.2924
Qwen 3 (14B)	0.5198	0.6162	0.1609	0.4554	0.3053

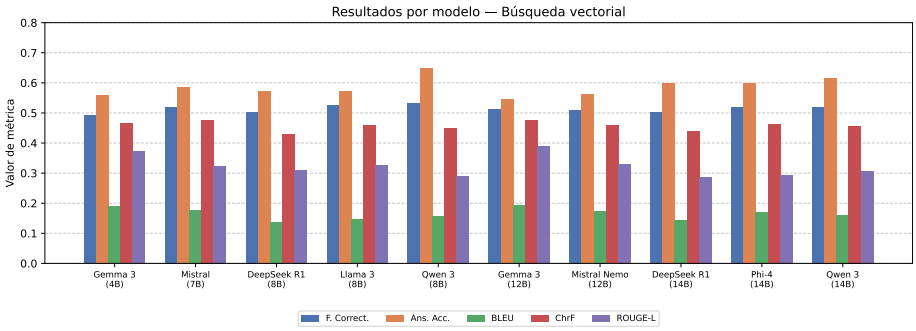

Figura 2. Resultados por modelo – Búsqueda vectorial.

Mejora híbrida + reranking vs. vectorial

La Tabla 4 muestra la mejora porcentual de la búsqueda híbrida + reranking respecto a la búsqueda vectorial ($\Delta\,\% = \frac{\text{híbrida}-\text{vectorial}}{\text{vectorial}} \times 100$). Los valores en **negrita** indican la mayor mejora por métrica.

Al incorporar búsqueda híbrida (vector + full-text con RRF) y reranking con cross-encoder (BAAI/bge-reranker-v2-m3), **los 10 modelos mejoran en las 5 métricas evaluadas sin excepción**. Los resultados clave son:

Factual Correctness (F1): mejora promedio del **19.3 % (+0.0988 puntos)**. El mayor incremento se observa en Gemma 3 12B (+0.1216); el menor en Qwen 3 8B (+0.0768).

Answer Accuracy: mejora promedio del **34.1 % (+0.1998 puntos)**. El mayor salto corresponde a Qwen 3 14B (+0.2370), que pasa de 0.6162 a 0.8532.

Métricas textuales: mejoras sistemáticas en BLEU, ChrF y ROUGE-L en todos los modelos, confirmando que las respuestas generadas se acercan más en contenido y formulación a las respuestas de referencia.

La mejora en estos resultados se debe principalmente a la forma en que esta técnica trabaja en capas:

Tabla 3. Modelos evaluados – Búsqueda híbrida + reranking.

Modelo	F. Correct.	Ans. Acc.	BLEU	ChrF	ROUGE-L
Gemma 3 (4B)	0.5867	0.7342	0.2138	0.5232	0.4169
Mistral (7B)	0.5976	0.7418	0.2103	0.5085	0.3530
DeepSeek R1 (8B)	0.6215	0.7854	0.1815	0.5073	0.3684
Llama 3 (8B)	0.6032	0.7595	0.1857	0.5097	0.3660
Qwen 3 (8B)	0.6081	**0.8572**	0.1741	0.4795	0.3138
Gemma 3 (12B)	**0.6329**	0.7576	**0.2289**	**0.5417**	**0.4474**
Mistral Nemo (12B)	0.6183	0.7607	0.2084	0.5246	0.3772
DeepSeek R1 (14B)	0.6183	0.7784	0.1637	0.4748	0.3086
Phi-4 (14B)	0.6102	0.8314	0.1920	0.4951	0.3154
Qwen 3 (14B)	0.6232	0.8532	0.1789	0.5004	0.3365

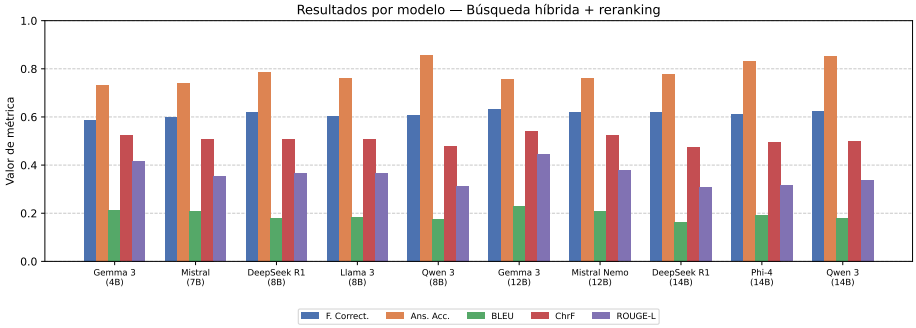

Figura 3. Resultados por modelo – Búsqueda híbrida + reranking.

1. **Búsqueda híbrida (vector + full-text con RRF):** recupera documentos que coinciden semánticamente y por palabras clave exactas, reduciendo los falsos positivos de la búsqueda vectorial pura.
2. **Reranking con cross-encoder:** el modelo BAAI/bge-reranker-v2-m3 reevalúa los candidatos recuperados como pares (pregunta, fragmento), produciendo un ranking más preciso que el score de similitud coseno.
3. **Efecto en el LLM:** al recibir un contexto más relevante, el modelo genera respuestas más fieles a la evidencia documental, independientemente de su tamaño o familia.

Análisis adicional

Los siguientes análisis, salvo indicación contraria, se basan en los resultados de búsqueda híbrida + reranking (Tabla 3).

Eficiencia por consumo de VRAM. Al relacionar la Factual Correctness en búsqueda híbrida con la VRAM requerida, Gemma 3 (4B) resulta el modelo más eficiente: alcanza FC = 0.587 consumiendo apenas 3.3 GB (0.178 FC/GB).

Tabla 4. Mejora porcentual: búsqueda híbrida + reranking vs. vectorial.

Modelo	F.C.	Ans. Acc.	BLEU	ChrF	ROUGE-L
Gemma 3 (4B)	+19.42 %	+31.41 %	+13.30 %	+12.49 %	+12.22 %
Mistral (7B)	+15.06 %	+26.48 %	+19.02 %	+6.74 %	+9.42 %
DeepSeek R1 (8B)	+23.63 %	+37.16 %	**+32.29 %**	**+18.53 %**	**+18.76 %**
Llama 3 (8B)	+14.78 %	+32.78 %	+27.19 %	+10.54 %	+12.37 %
Qwen 3 (8B)	+14.45 %	+31.96 %	+10.89 %	+6.82 %	+8.06 %
Gemma 3 (12B)	**+23.78 %**	**+39.06 %**	+18.29 %	+13.85 %	+15.28 %
Mistral Nemo (12B)	+21.54 %	+35.55 %	+19.91 %	+14.52 %	+13.78 %
DeepSeek R1 (14B)	+23.19 %	+29.80 %	+14.40 %	+7.79 %	+7.56 %
Phi-4 (14B)	+17.39 %	+38.47 %	+12.87 %	+7.33 %	+7.87 %
Qwen 3 (14B)	+19.89 %	+38.46 %	+11.19 %	+9.88 %	+10.22 %

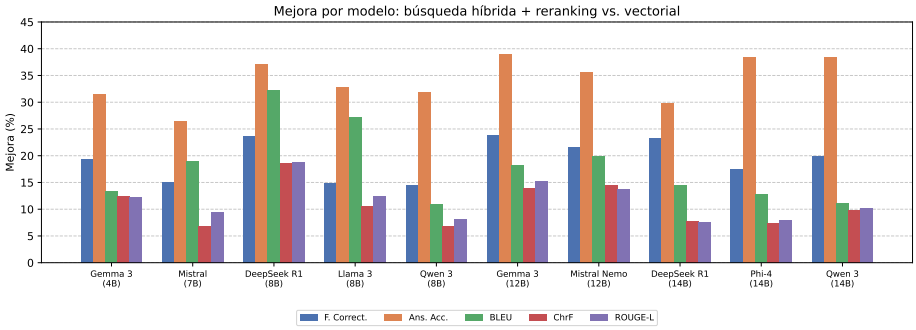

Figura 4. Mejora porcentual por modelo y métrica: híbrida + reranking vs. vectorial.

En contraste, los modelos de 14B alcanzan FC entre 0.618 y 0.623 con 9.0–9.3 GB (0.067–0.069 FC/GB). Esta diferencia es relevante para despliegues en hardware con VRAM limitada, como la computadora local a la que se conecta el dispositivo portatil Procer 4 en reemplazo del servicio en la nube.

Correlación entre cantidad de parámetros y rendimiento. Los resultados no muestran una correlación positiva entre número de parámetros y rendimiento. Qwen 3 (8B) obtiene la mayor Answer Accuracy en modalidad híbrida (0.857), superando a todos los modelos de 12B y 14B en esa métrica. DeepSeek R1 (8B) iguala la Factual Correctness de los modelos de 14B (0.622). Esto sugiere que la arquitectura y el ajuste fino inciden más que el tamaño del modelo en este dominio.

Divergencia entre métricas LLM y métricas textuales. Se observa una disociación entre las métricas basadas en LLM (Factual Correctness, Answer Accuracy) y las de similitud textual (BLEU, ChrF, ROUGE-L). Qwen 3 (8B) lidera en Answer Accuracy (0.857, híbrido) pero obtiene el ROUGE-L más bajo del conjunto (0.314). En contraste, Gemma 3 (12B) lidera en BLEU (0.229), ChrF (0.542) y

ROUGE-L (0.447) con una Answer Accuracy intermedia (0.758). Esta divergencia indica que algunos modelos generan respuestas factualmente correctas con formulación distinta a la referencia, mientras otros producen texto más similar al original pero no necesariamente más preciso en los hechos.

Mejor modelo con recursos limitados (<6 GB VRAM). Considerando únicamente los modelos que requieren menos de 6 GB de VRAM (modelos de 4B y 8B), Gemma 3 (4B) presenta el mayor promedio global en búsqueda híbrida (0.495), seguido por DeepSeek R1 (8B) con 0.493. Para entornos con hardware restringido, ambos representan la mejor relación entre rendimiento y consumo de recursos.

Ranking consolidado. La Tabla 5 presenta el promedio aritmético de las cinco métricas para cada modelo en ambas modalidades, ordenado por rendimiento en búsqueda híbrida.

Tabla 5. Ranking consolidado: promedio de las 5 métricas por modelo.

Modelo	**Params**	**Prom. vec.**	**Prom. híb.**
Gemma 3	12B	**0.4227**	**0.5217**
Qwen 3	14B	0.4115	0.4984
Mistral Nemo	12B	0.4067	0.4978
Gemma 3	4B	0.4151	0.4950
DeepSeek R1	8B	0.3901	0.4928
Phi-4	14B	0.4088	0.4888
Qwen 3	8B	0.4154	0.4865
Llama 3	8B	0.4061	0.4848
Mistral	7B	0.4163	0.4822
DeepSeek R1	14B	0.3944	0.4688

7. Conclusión y perspectiva a futuro

Se presentó un sistema RAG de código abierto que opera íntegramente de forma local, evaluado sobre un corpus real de documentos educativos con 396 pares pregunta-respuesta. La arquitectura implementada combina búsqueda vectorial densa, búsqueda full-text y reranking neuronal, sin depender de servicios externos en la nube.

Los resultados de la evaluación muestran una mejora consistente de la búsqueda híbrida con reranking respecto a la búsqueda vectorial pura en las cinco métricas evaluadas y en los diez modelos analizados. Estos resultados sugieren que invertir en la etapa de recuperación —mediante búsqueda híbrida y reranking neuronal— produce mayores ganancias que escalar el tamaño del modelo generador. Esta conclusión tiene implicaciones prácticas importantes: permite

obtener respuestas de mayor calidad sin aumentar los requerimientos de hardware, lo que resulta especialmente relevante para despliegues en dispositivos con recursos limitados.

El sistema propuesto opera íntegramente sobre infraestructura local, eliminando la dependencia de servicios externos y garantizando privacidad y soberanía sobre los datos educativos. Este enfoque demuestra que es viable construir soluciones de IA de calidad competitiva sobre hardware de consumo, abriendo camino hacia su adopción en entornos con conectividad limitada o requisitos estrictos de privacidad.

Como trabajo futuro se propone incorporar las métricas Context Precision y Context Recall al framework de evaluación, con el objetivo de medir la calidad del retrieval de forma independiente al LLM: estas métricas permiten determinar si el sistema recupera los fragmentos relevantes y si los fragmentos recuperados son efectivamente los necesarios para responder cada pregunta.

Otra mejora que se pretende realizar a futuro es agregar soporte para consultas multimodales, permitiendo que el alumno adjunte imágenes como parte del contexto de su pregunta. Esta capacidad resultaría especialmente valiosa para los perfiles de usuario a los que va destinada la herramienta: estudiantes con discapacidad visual que utilizan el dispositivo como herramienta de asistencia auditiva, con dislexia o con otras barreras de aprendizaje que dificultan la formulación escrita de preguntas. Incorporar imágenes como entrada ampliaría significativamente la usabilidad del sistema y reforzaría su propósito de reducir la brecha de desigualdad educativa.

Agradecimientos

Este trabajo utilizó recursos computacionales de UNC Supercómputo (CCAD) de la Universidad Nacional de Córdoba (https://supercomputo.unc.edu.ar), que forman parte del Sistema Nacional de Computación de Alto Desempeño (SNCAD) de la República Argentina.

Referencias

1. Abdin, M., et al.: Phi-4 technical report. arXiv preprint arXiv:2412.08905 (2024)
2. Chase, H., et al.: LangChain. https://github.com/langchain-ai/langchain (2022)
3. Chen, J., Xiao, S., Zhang, P., Luo, K., Lian, D., Liu, Z.: M3-Embedding: Multilinguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. arXiv preprint arXiv:2402.03216 (2024)
4. Cormack, G.V., Clarke, C.L., Buettcher, S.: Reciprocal rank fusion outperforms Condorcet and individual rank learning methods. In: Proceedings of the 32nd Annual ACM SIGIR Conference. pp. 758–759. ACM (2009)
5. DeepSeek-AI: DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. arXiv preprint arXiv:2501.12948 (2025)
6. Es, S., James, J., Espinosa-Anke, L., Schockaert, S.: RAGAS: Automated evaluation of retrieval augmented generation. arXiv preprint arXiv:2309.15217 (2023)

7. Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, M., Wang, H.: Retrieval-augmented generation for large language models: A survey. arXiv preprint arXiv:2312.10997 (2023)
8. Gemma Team, Google DeepMind: Gemma 3 technical report. arXiv preprint arXiv:2503.19786 (2025)
9. Gerganov, G., et al.: llama.cpp: LLM inference in C/C++. https://github.com/ggerganov/llama.cpp (2023)
10. Jiang, A.Q., et al.: Mistral 7B. arXiv preprint arXiv:2310.06825 (2023)
11. Kane, A.: pgvector: Open-source vector similarity search for Postgres. https://github.com/pgvector/pgvector (2021)
12. Lewis, P., et al.: Retrieval-augmented generation for knowledge-intensive NLP tasks. In: Advances in Neural Information Processing Systems. vol. 33, pp. 9459–9474 (2020)
13. Lin, C.Y.: ROUGE: A package for automatic evaluation of summaries. In: Text Summarization Branches Out. pp. 74–81. ACL (2004)
14. Llama Team, Meta AI: The Llama 3 herd of models. arXiv preprint arXiv:2407.21783 (2024)
15. Mistral AI: Mistral NeMo. https://mistral.ai/news/mistral-nemo/ (2024)
16. Mohr, I., Krimmel, M., Sturua, S., et al.: Multi-task contrastive learning for 8192-token bilingual text embeddings. arXiv preprint arXiv:2402.17016 (2024)
17. NVIDIA Corporation: Answer Accuracy – RAGAS documentation. https://docs.ragas.io/en/latest/concepts/metrics/available_metrics/nvidia_metrics/#answer-accuracy (2025)
18. OpenAI: Introducing GPT-5.2. https://openai.com/es-419/index/introducing-gpt-5-2/ (2025)
19. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: BLEU: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of the ACL. pp. 311–318 (2002)
20. Popović, M.: chrF: character n-gram F-score for automatic MT evaluation. In: Proceedings of the Tenth Workshop on Statistical Machine Translation. pp. 392–395. ACL (2015)
21. Procer Tecnologías S.A.S.: Procer 4: Lector auditivo portátil con inteligencia artificial. https://www.procertecnologias.com/products/procer4_es (2024)
22. Qwen Team: Qwen3 technical report. arXiv preprint arXiv:2505.09388 (2025)
23. Factual Correctness – RAGAS documentation. https://docs.ragas.io/en/latest/concepts/metrics/available_metrics/factual_correctness/ (2025)
24. Ramírez, S.: FastAPI. https://fastapi.tiangolo.com (2018)
25. Resolución decanal 758/2025: Servicio de desarrollo de la Facultad de Matemática, Astronomía, Física y Computación (FAMAF) de la Universidad Nacional de Córdoba (UNC) para la empresa Procer Tecnologías SAS. Digesto Electrónico de la UNC https://digesto.unc.edu.ar/handle/123456789/598620 (2025)
26. Voucher de innovación colaborativa (prensa agencia innovar y emprender córdoba). https://prensa.cba.gov.ar/informacion-general/innovacion-abierta-y-financiamiento-lanzan-dos-programas-clave-para-empresas-y-startups/ (2025)
27. Ollama. https://ollama.com (2023)