Asistentes Virtuales de Enseñanza basados en Generación Aumentada por Recuperación

Juan Pablo Tessore [1 2], Claudia Russo [2 3], Hugo Ramón [2 3], Leonardo Esnaola [2], Tamara Ahmad [2]

Instituto de Investigación y Transferencia en Tecnología (ITT)
Centro asociado a la Comisión de Investigaciones Científicas (CIC)
Escuela de Tecnología (ET)
Universidad Nacional del Noroeste de la Provincia de Buenos Aires (UNNOBA)

Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET)

juanpablo.tessore@itt.unnoba.edu.ar, claudia.russo@itt.unnoba.edu.ar, hugo.ramon@itt.unnoba.edu.ar, leonardo.esnaola@itt.unnoba.edu.ar,, tamara.ahmad@itt.unnoba.edu.ar

Resumen

Los avances recientes en los Grandes Modelos de Lenguaje (LLM) los han consolidado como una tecnología transformadora en inteligencia artificial. Entrenados con grandes volúmenes de texto, aprenden patrones del lenguaje que les permiten generar contenido coherente, responder preguntas, resumir, traducir, programar y asistir en tareas cognitivas complejas. En educación, su potencial es especialmente relevante: pueden actuar como tutores virtuales, adaptar explicaciones al nivel del estudiante, proponer ejercicios, corregir actividades y acompañar procesos de aprendizaje personalizado, favoreciendo la accesibilidad y la retroalimentación inmediata.

No obstante, su aplicación efectiva requiere abordar el problema de las alucinaciones (es decir, la generación de información incorrecta o no verificada con apariencia plausible) así como integrar contenido institucional específico. En este sentido, la Generación Aumentada por Recuperación (RAG, Retrieval-Augmented Generation, por sus siglas en inglés) [1] combina un modelo generativo con un sistema externo de recuperación de información, lo que permite consultar documentos y repositorios verificados antes de producir una respuesta.

El presente trabajo propone medir empíricamente la mejora en la calidad de las respuestas de un LLM base frente a una variante con enfoque RAG en contextos educativos. Asimismo, busca elaborar una metodología para guiar la carga de contenido académico con dicha arquitectura.

***Palabras clave*:** grandes modelos del lenguaje, RAG, asistentes virtuales.

Contexto

Esta línea de investigación forma parte de los proyectos “Técnicas y algoritmos avanzados para la optimización de procesos: seguridad, sostenibilidad e innovación” e “Innovación y Desafíos en la Educación Digital” presentados en el marco de la convocatoria a Subsidios de Investigación Bianuales (SIB 2025) de la Secretaría de Investigación, Desarrollo y Transferencia de la Universidad Nacional del Noroeste de la Provincia de Buenos Aires (UNNOBA). A su vez, se enmarca en el contexto de un plan de trabajo aprobado por la Comisión Nacional de Investigaciones Científicas y Técnicas, en el marco de la

---

[1] Becario Postdoctoral CONICET. [2] Docente Investigador ITT. [3] Inv. Asociado Adjunto sin director CIC.

convocatoria “Becas Postdoctorales 2022 Temas Estratégicos”.

El proyecto se desarrolla en el Instituto de Investigación y Transferencia en Tecnología (ITT) dependiente de la mencionada Secretaría y se trabaja en conjunto con la Escuela de Tecnología de la UNNOBA.

El equipo está constituido por docentes e investigadores pertenecientes al ITT y a otros institutos de investigación, así como también estudiantes de las carreras de Informática de la Escuela de Tecnología de la UNNOBA.

Introducción

En los últimos años, los LLM han transformado de manera significativa la forma en que interactuamos con la información. Su desarrollo se basa en la arquitectura Transformer, introducida por Google [2], y posteriormente perfeccionada en modelos de gran escala como los desarrollados por OpenAI [3], Google [4] y Anthropic [5], entre otros. Dichos modelos han demostrado capacidades avanzadas para comprender y generar texto en lenguaje natural, responder preguntas, resumir contenidos, asistir en programación y apoyar procesos creativos. En el ámbito educativo, estas tecnologías permiten ofrecer tutorías personalizadas, generación automática de material didáctico, retroalimentación formativa inmediata y acompañamiento en procesos de aprendizaje autónomo. Su potencial radica en la capacidad de adaptarse al contexto del estudiante y responder en tiempo real a consultas diversas.

Sin embargo, pese a sus avances, los LLM presentan limitaciones relevantes. Una de las principales es el fenómeno de las alucinaciones. Dado que estos modelos se entrenan con grandes volúmenes de datos generales y no tienen acceso directo y actualizado a fuentes específicas, pueden producir respuestas inexactas o inventadas, especialmente cuando se les consulta sobre información institucional, reglamentos internos, programas académicos o contenidos específicos de una organización educativa. Esta limitación resulta crítica cuando se pretende implementar asistentes virtuales en contextos formales, donde la precisión y la confiabilidad de la información son fundamentales.

Ante esta problemática, existen distintas alternativas para implementar asistentes educativos. Una opción tradicional son los sistemas basados en reglas (rule-based systems), que operan mediante flujos predefinidos y árboles de decisión. Si bien ofrecen alto control sobre las respuestas y reducen el riesgo de errores factuales, presentan importantes desventajas: son rígidos, difíciles de escalar, costosos de mantener y poco flexibles ante consultas abiertas o imprevistas. Otra alternativa consiste en realizar fine-tuning de algún LLM con datos institucionales. Aunque esta estrategia permite adaptar el comportamiento del modelo a un dominio específico, implica costos computacionales elevados, requiere grandes volúmenes de datos curados, demanda procesos técnicos complejos y, además, no garantiza la actualización dinámica del conocimiento sin reentrenamientos periódicos.

En este contexto, el enfoque RAG emerge como una solución intermedia y eficiente. RAG combina la capacidad generativa de los LLM con un mecanismo de recuperación de información desde una base de conocimiento externa y controlada. Antes de generar una respuesta, el sistema consulta documentos relevantes (por ejemplo, reglamentos, programas de estudio o materiales institucionales) y utiliza ese contenido como contexto explícito para fundamentar la respuesta. De esta manera, se reduce significativamente la probabilidad de

alucinaciones y se asegura que la información provenga de fuentes verificadas y actualizadas, sin necesidad de reentrenar el modelo.

A continuación, se presentan diversas implementaciones basadas en RAG en el ámbito educativo, junto con los resultados obtenidos en términos de mejora en la calidad, precisión y pertinencia de las respuestas generadas.

En [6] se desarrolló un sistema de tutoría inteligente llamado *LPITutor* que integra un modelo de lenguaje con una arquitectura RAG y técnicas de prompt engineering para generar explicaciones adaptadas al nivel del estudiante. El sistema recupera información relevante desde fuentes externas antes de generar la respuesta, lo que permite ajustar la profundidad conceptual según la consulta. La evaluación reporta mejoras en la precisión y claridad de las respuestas, así como una reducción significativa de alucinaciones en comparación con un LLM sin recuperación, destacándose su capacidad para ofrecer explicaciones pedagógicamente alineadas con los contenidos académicos.

También en [7] se desarrolló un tutor basado en RAG llamado *SmartStudy*, en este caso para cursos universitarios de Machine Learning. El sistema procesa automáticamente los libros de texto del curso y construye una base de conocimiento que alimenta al modelo durante la generación de respuestas y la creación de exámenes de práctica. En una evaluación con estudiantes de ciencia de datos, se observaron mejoras en el desempeño en pruebas y efectos positivos en el aprendizaje percibido, lo que sugiere que la incorporación de recuperación contextual fortalece la comprensión conceptual y la retención de contenidos.

En [8] se implementó en el ámbito de la educación médica en la Geisel School of Medicine de Dartmouth, una arquitectura RAG alimentada con apuntes y bibliografía oficial de un curso de neuroanatomía, *NeuroBot*. El sistema registró una adopción cercana a un tercio de los estudiantes y fue utilizado especialmente en períodos previos a exámenes. Los estudiantes valoraron positivamente que las respuestas estuvieran fundamentadas en el material oficial, con niveles de utilidad percibida moderados pero consistentes. La integración de RAG permitió anclar las respuestas al contenido curricular, aumentando la confianza en el sistema.

Otro tutor [9], *TutorLLM*, fue desarrollado en la Universidad de Southampton y combinó RAG con técnicas de knowledge tracing para personalizar la tutoría según el estado de aprendizaje del estudiante. El sistema recuperaba fragmentos relevantes del curso y ajustaba la generación en función del progreso individual. En evaluaciones comparativas, el enfoque con RAG mostró mejoras aproximadas del 5% en resultados de evaluación y cerca de un 10% en satisfacción estudiantil frente a un LLM sin recuperación, evidenciando beneficios tanto en desempeño como en experiencia de uso.

Finalmente, en [10], también se implementaron chatbots restringidos exclusivamente al material de cada curso en dos universidades. Los resultados mostraron altos niveles de compromiso y motivación en los estudiantes, así como una tasa muy baja de respuestas incorrectas. La arquitectura RAG redujo de manera sustancial las alucinaciones respecto a versiones sin recuperación, demostrando que el acceso controlado a fuentes académicas específicas mejora la precisión y la pertinencia pedagógica de las respuestas en entornos universitarios.

De los trabajos analizados se evidencia que la arquitectura RAG no solo mejora el desempeño de los asistentes virtuales basados en LLM en términos de reducción de

alucinaciones, sino también en cuanto a la satisfacción de los usuarios. Resulta pertinente, en consecuencia, profundizar la investigación con esta arquitectura.

Líneas de Investigación, Desarrollo e Innovación

La presente investigación se encuadra dentro de los siguientes proyectos SIB 2025:

- *"Técnicas y algoritmos avanzados para la optimización de procesos: seguridad, sostenibilidad e innovación"*, cuyo objetivo general es *"Identificar procesos en actividades económicas susceptibles de ser mejorados y optimizados mediante la aplicación de tecnologías innovadoras, que podrían integrar las señales analizadas con un enfoque multimodal para potenciar aún más estos procesos"*.
- *"Innovación y Desafíos en la Educación Digital",* cuyo objetivo general es *"Investigar, diseñar y desarrollar soluciones informáticas innovadoras aplicadas a la educación digital, basadas en inteligencia artificial, tecnologías emergentes y analítica de datos, entre otras, con el propósito de transformar los procesos educativos en todas sus modalidades."*

En este sentido, la presente investigación sobre asistentes virtuales basados en RAG contribuye a dichos proyectos SIB al abordar los siguientes aspectos:

- Investigar arquitecturas, técnicas y herramientas que faciliten la implementación de asistentes virtuales basados en RAG, tales como bases de datos vectoriales, grandes modelos de lenguaje, estrategias de segmentación e indexación de contenido, etc.
- Definir métricas que permitan evaluar el impacto en el desempeño de los asistentes basados en RAG en comparación con una línea base sin RAG, así como medir su utilización e impacto educativo en cursos de nivel universitario.
- Realizar pruebas piloto en cursos universitarios con el objetivo de evaluar el desempeño de los asistentes implementados, utilizando las métricas definidas en el punto anterior.

Resultados y Objetivos

En el marco de este trabajo, se espera diseñar e implementar un asistente virtual de enseñanza basado en una arquitectura RAG, orientado a cursos universitarios. El sistema integraría un LLM de propósito general con una base de conocimiento compuesta por programas de estudio, reglamentos, apuntes de cátedra, trabajos prácticos y bibliografía oficial, procesados mediante estrategias de segmentación e indexación en una base de datos vectorial.

Como objetivo principal, se buscaría evaluar comparativamente el desempeño de un LLM base frente a su variante con RAG en escenarios educativos reales. Para ello, se definirían métricas para evaluar la calidad de las respuestas generadas por el asistente. Asimismo, se analizará la satisfacción estudiantil, la percepción de utilidad y el nivel de confianza en las respuestas generadas.

Entre los resultados esperados, se anticipa que la arquitectura RAG podría reducir significativamente la generación de respuestas incorrectas o no fundamentadas, al anclar la producción textual en documentos institucionales verificados. Se proyecta que el asistente basado en RAG mostraría mejoras en precisión y adecuación contextual respecto a la línea base, especialmente en consultas relacionadas con contenidos específicos del curso o normativa institucional.

Adicionalmente, se prevé que la implementación piloto permitiría obtener evidencia empírica sobre el impacto del asistente en un escenario educativo real.

Finalmente, el trabajo tendrá como objetivo elaborar una guía metodológica para la incorporación de contenido académico en arquitecturas RAG, contemplando criterios de curación, actualización y validación de la base de conocimiento. Dicha metodología podría servir como referencia para futuras implementaciones institucionales, contribuyendo a la escalabilidad de asistentes virtuales en el ámbito de la educación superior.

Formación de Recursos Humanos

En esta línea de I+D se ha obtenido y se encuentra desarrollando actualmente una beca postdoctoral del CONICET. Asimismo, se espera desarrollar una tesina de grado y/o PPS, dirigida por miembros de este proyecto.

Bibliografía

1. 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. (2020). *Retrieval-augmented generation for knowledge-intensive NLP tasks.* arXiv. https://arxiv.org/abs/2005.11401
2. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). *Attention is all you need.* In I. Guyon et al. (Eds.), *Advances in Neural Information Processing Systems* (Vol. 30). Curran Associates. http://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf
3. OpenAI. (2025). *ChatGPT* (versión 12 de mayo de 2025) [Modelo de lenguaje de IA]. https://chat.openai.com/
4. Anthropic. (2025). *Claude* [Modelo de lenguaje de IA]. https://www.anthropic.com/claude
5. Google AI. (2025). *Gemini* [Modelo de lenguaje de IA]. https://gemini.google.com/
6. Liu, Z., Agrawal, P., Singhal, S., Madaan, V., Kumar, M., & Verma, P. K. (2025). *LPITutor: An LLM based personalized intelligent tutoring system using retrieval-augmented generation and prompt engineering. PeerJ Computer Science, 11*, e2991. https://doi.org/10.7717/peerj-cs.2991
7. Vu-Minh, A., Nguyen-The, Q., Nguyen, N., Pham, X.-L., & Nguyen, A. (2025). *Designing a course-grounded AI tutor with retrieval-augmented generation: A DSR approach to technical education. HICSS-59 Proceedings.* https://doi.org/10.2139/ssrn.5607971
8. Thesen, T., & Park, S. H. (2025). *A generative AI teaching assistant for personalized learning in medical education. npj Digital Medicine, 8*, 627. https://doi.org/10.1038/s41746-025-02022-1
9. Li, Z., Yazdanpanah, V., Wang, J., Gu, W., Shi, L., Cristea, A. I., Kiden, S., & Stein, S. (2025). *TutorLLM: Customizing learning recommendations with knowledge tracing and retrieval-augmented generation* (Version 2502.15709v1) [Preprint]. arXiv. https://arxiv.org/abs/2502.15709
10. Németh, R., Tátrai, A., Szabó, M., Zaletnyik, P. T., & Tamási, Á. (2025). *Exploring the use of retrieval-augmented generation models in higher education: A pilot study on artificial intelligence-based tutoring. Social Sciences & Humanities Open, 12*, 101751. https://doi.org/10.1016/j.ssaho.2025.101751