06

MITIGACIÓN DE FUGAS

DE DATOS E INYECCIÓN DE PROMPTS EN CHATBOTS UNIVERSITARIOS MEDIANTE ARQUITECTURA PERIMETRAL DESACOPLADA

© 2026 Los autores. Este artículo se distribuye bajo la licencia Creative Commons Atribución-NoComercial-CompartirIgual 4.0 Internacional (CC BY-NC-SA 4.0). https://creativecommons.org/licenses/by-nc-sa/4.0/

ISSN: 2631-2662

MITIGACIÓN DE FUGAS

DE DATOS E INYECCIÓN DE PROMPTS EN CHATBOTS UNIVERSITARIOS MEDIANTE ARQUITECTURA PERIMETRAL DESACOPLADA

MITIGATING DATA LEAKS AND PROMPT INJECTION IN UNIVERSITY CHATBOTS USING A SECURE DECOUPLED PERIMETER-BASED ARCHITECTURE

Jorge Wilson Saa-Saltos[1]
E-mail: jsaa@uteq.edu.ec
ORCID: https://orcid.org/0000-0002-8368-4131
Nancy Magali Rodríguez-Gavilanes[1]
E-mail: nrodriguez@uteq.edu.ec
ORCID: https://orcid.org/0000-0002-0861-4352
[1] Universidad Técnica Estatal de Quevedo. Ecuador.

Cita sugerida (APA, séptima edición)

Saa-Saltos, J. W., & Rodríguez-Gavilanes, N. M. (2026). Mitigación de fugas de datos e inyección de prompts en chatbots universitarios mediante arquitectura perimetral desacoplada. *Revista Metropolitana de Ciencias Aplicadas, 9*(6), 56-64.

Fecha de presentación: 21/07/2026 **Fecha de aceptación:** 29/09/2026 **Fecha de publicación:** 01/11/2026

RESUMEN

La educación superior está adoptando rápidamente herramientas de Inteligencia Artificial Generativa, situación que ha expuesto a las universidades a vulnerabilidades críticas de ciberseguridad, principalmente la inyección de prompts y la fuga de datos. El presente estudio diseña y evalúa una arquitectura de un agente conversacional descentralizado que funciona como un asistente virtual de asignatura, considerando las recomendaciones de ciberseguridad del cuerpo docente. El enfoque metodológico es teórico experimental, la investigación se desarrolla en tres fases: un diagnóstico de conocimiento de ciberseguridad de los docentes mediante un cuestionario digital, el diseño de un middleware perimetral que mitigue la fuga de datos y sea eficaz ante la inyección de prompts, y la evaluación experimental del diseño. Los resultados experimentales demuestran que el diseño de la arquitectura propuesta reduce significativamente la exposición de datos y la vulnerabilidad ante inyecciones de prompts. La investigación aporta una solución práctica y replicable que fortalece la privacidad, la seguridad de la información y la adopción responsable de la IA en el entorno universitario.

Palabras clave:

Modelos de Gran Lenguaje, inyección de prompts, fuga de datos, inteligencia artificial, arquitectura descentralizada, educación superior y ciberseguridad.

ABSTRACT

Higher education is rapidly adopting Generative Artificial Intelligence tools, a situation that has exposed universities to critical cybersecurity vulnerabilities, mainly prompt injection and data leakage. This study designs and evaluates an architecture for a decentralized conversational agent that functions as a virtual course assistant, taking into account the cybersecurity recommendations of the faculty. The methodological approach is theoretical-experimental, and the research is conducted in three phases: an assessment of faculty members' cybersecurity knowledge via a digital questionnaire; the design of perimeter middleware that mitigates data leakage and is effective against prompt injection; and the experimental evaluation of the design. The experimental results demonstrate that the proposed architectural design significantly reduces data exposure and vulnerability to prompt injection attacks. The research provides a practical and replicable solution that strengthens privacy, information security, and the responsible adoption of AI in the university setting.

Keywords:

Large Language Models, Prompt Injection, data leakage, IA, higher education y cybersecurity competencies.

INTRODUCCIÓN

La transformación digital de la educación superior ha incrementado la utilización de plataformas tecnológicas para la gestión de los procesos de enseñanza y aprendizaje. En este escenario, los docentes deben atender consultas de los estudiantes relacionadas con contenidos académicos, materiales didácticos, tareas, sílabos, planificación y criterios de evaluación, entre otros aspectos, incluso fuera de sus horarios habituales de trabajo. Los Sistemas de Gestión del Aprendizaje (LMS, por sus siglas en inglés), como Moodle o Canvas, permiten almacenar y organizar grandes volúmenes de información académica y constituyen repositorios fundamentales para la actividad docente.

Sin embargo, su función se concentra principalmente en la gestión y disponibilidad de recursos, por lo que la incorporación de interfaces conversacionales basadas en Inteligencia Artificial Generativa (IAG) representa una alternativa para facilitar el acceso a información académica de manera inmediata, intuitiva y personalizada. Esta incorporación también plantea interrogantes sobre la integridad académica, debido a que las capacidades generativas de herramientas como ChatGPT modifican las formas tradicionales de producir, evaluar y acreditar el trabajo académico, lo que exige revisar las prácticas de evaluación y las estrategias institucionales para promover un uso responsable de estas tecnologías (Cotton et al., 2024).

La incorporación de agentes conversacionales en estos entornos introduce, al mismo tiempo, desafíos relacionados con la confiabilidad de las respuestas, la privacidad de la información y la seguridad de las interacciones. La vulnerabilidad no depende únicamente de las características técnicas del modelo, sino también del contexto en el que se produce la interacción y de la información que los usuarios pueden revelar durante sus conversaciones con estos sistemas (De Freitas & Cohen, 2025). Entre los riesgos tecnológicos se encuentra la fuga de datos (***data leakage***), mediante la cual información confidencial puede quedar expuesta involuntariamente o ser accesible para usuarios no autorizados (Li et al., 2025). A ello se suma la inyección de prompts (***prompt injection***), que ocurre cuando se introducen instrucciones en lenguaje natural diseñadas para alterar, evadir o anular las directrices y restricciones establecidas para el funcionamiento del sistema. La exposición de los LLM a entradas adversariales constituye un problema relevante incluso en aplicaciones educativas especializadas, por lo que se han investigado estrategias de entrenamiento y robustecimiento destinadas a mantener un comportamiento estable frente a intentos de manipulación (Anghel et al., 2026). Ante estas amenazas, también se han explorado mecanismos de protección mediante marcos de seguridad o ***guardrails***, así como estrategias basadas en aprendizaje federado para fortalecer la protección, privacidad y robustez de los sistemas que incorporan modelos de lenguaje de gran escala (Luo & Ji, 2025).

Estos problemas adquieren especial relevancia en el contexto universitario, donde los agentes conversacionales pueden interactuar con materiales docentes y otros contenidos académicos que requieren mecanismos específicos de protección. Incluso los modelos desarrollados o ajustados con criterios de seguridad continúan siendo susceptibles a manipulaciones sofisticadas, lo que justifica la incorporación de capas externas de contención y validación en el flujo de trabajo del LLM (Dzhaliuk et al., 2026).

Desde una perspectiva arquitectónica, la distribución de funciones entre componentes especializados constituye un antecedente relevante para el diseño de sistemas de ciberseguridad, pues las arquitecturas multiagente permiten separar tareas de detección, procesamiento y enriquecimiento de información en componentes coordinados (Trinks et al., 2023). Asimismo, la literatura evidencia la necesidad de ampliar la validación empírica, la evaluación comparativa de mecanismos defensivos y la documentación de experiencias de implementación segura en contextos reales (Gundu, 2023). En particular, resulta necesario profundizar en soluciones educativas de código abierto, ligeras y de bajo costo que incorporen mecanismos para validar las entradas y salidas del sistema y reducir los riesgos de exposición o manipulación de los materiales utilizados por el docente.

Esta problemática también se manifiesta en el contexto específico analizado. El diagnóstico empírico realizado a 105 docentes durante el Primer Periodo Académico 2026-2027 de una universidad pública del Ecuador mostró que el 84,80 % identificó la privacidad de los datos institucionales como una preocupación relevante para la adopción de IAG, mientras que el 52,40 % señaló la necesidad de reducir el riesgo de respuestas inexactas o alucinaciones. Estos resultados evidencian que la incorporación de asistentes conversacionales en la educación superior no depende únicamente de sus capacidades funcionales, sino también de la existencia de mecanismos que permitan proteger la información, controlar las interacciones con el modelo y fundamentar las respuestas en contenidos académicos previamente establecidos.

Frente a estas limitaciones, la presente investigación aborda la brecha identificada mediante el diseño de una arquitectura de software descentralizada, desacoplada y orientada a la seguridad para un agente conversacional universitario. La propuesta integra capas de contención perimetral en el flujo de trabajo del LLM como mecanismos externos de protección frente a entradas potencialmente maliciosas y posibles exposiciones de información. Desde esta perspectiva, el aporte teórico se centra en la articulación de estos mecanismos dentro de una arquitectura destinada al contexto educativo, mientras que el aporte tecnológico radica en una solución basada

en herramientas de código abierto y componentes desacoplados que favorecen su replicabilidad y reducen los requerimientos de infraestructura institucional.

El objetivo de este estudio es diseñar un chatbot inteligente basado en una arquitectura descentralizada y desacoplada, orientado a mitigar la fuga de datos y los ataques por inyección de prompts, en correspondencia con los requerimientos ético-pedagógicos de la educación superior. Las principales contribuciones de la investigación son: 1) la caracterización cuantitativa de las percepciones y competencias de los docentes relacionadas con la ciberseguridad y el uso de la IAG; y 2) el diseño de una arquitectura descentralizada y desacoplada para un chatbot con funciones de asistente virtual seguro, de código abierto y replicable.

MATERIALES Y MÉTODOS

La investigación se desarrolló desde un enfoque cuantitativo, con alcance descriptivo y orientación aplicada, articulando el diagnóstico de las percepciones y competencias docentes en materia de ciberseguridad con la especificación técnica de una arquitectura segura para un agente conversacional universitario. El estudio consideró dos ejes complementarios: el factor humano asociado con el uso de herramientas de inteligencia artificial en el contexto educativo y los requerimientos tecnológicos necesarios para reducir los riesgos de seguridad derivados de su utilización. A partir de estos ejes, la propuesta se estructuró en tres dimensiones: diseño de una arquitectura desacoplada, incorporación de controles perimetrales de seguridad y valoración de la viabilidad técnica y la correspondencia ético-pedagógica de la solución con las necesidades del entorno universitario.

La población estuvo constituida por la totalidad del cuerpo docente de una universidad pública, integrado por N = 537 profesores. La recopilación de información se realizó durante el Primer Periodo Académico 2026-2027 mediante la distribución de un cuestionario digital a la totalidad de los docentes. Se empleó un muestreo no probabilístico por respuesta voluntaria y la muestra efectiva quedó conformada por n = 105 participantes que respondieron el instrumento de manera libre e informada, lo que representó una tasa de respuesta del 19,55 % de la población. El diagnóstico se orientó a obtener información sobre las prácticas, percepciones y competencias de los docentes frente al uso de herramientas de inteligencia artificial, así como sobre los requerimientos considerados necesarios para la implementación de un asistente conversacional en el contexto universitario.

Como instrumento de recopilación de información se utilizó un cuestionario digital compuesto por ocho preguntas, estructuradas en tres dimensiones: diagnóstico del factor humano y educativo, competencias de ciberseguridad de los docentes y requerimientos del chatbot inteligente en su función de asistente virtual. El instrumento combinó preguntas valoradas mediante una escala de Likert de cinco puntos con ítems de selección múltiple. La información obtenida permitió caracterizar las percepciones y competencias de los participantes e identificar requerimientos funcionales y de seguridad que fueron considerados posteriormente en la especificación tecnológica de la propuesta.

Para el desarrollo de la solución se adoptó un paradigma de código abierto y una arquitectura distribuida en tres capas funcionalmente desacopladas: presentación, gestión del flujo de datos e inteligencia del sistema, y ciberseguridad. La capa de presentación se desarrolló en Python mediante Streamlit (2026) y se desplegó en Streamlit Cloud, con integración continua y despliegue continuo mediante un repositorio privado de GitHub. Esta capa proporciona acceso mediante HTTPS y permite a docentes y estudiantes cargar material bibliográfico en formato PDF e interactuar con el asistente conversacional desde una interfaz web.

La gestión, almacenamiento y recuperación vectorial de la información se realizó mediante ChromaDB (Chroma Core Inc, 2024), implementada en modo embebido dentro de la aplicación desplegada en Streamlit Cloud. El motor de inferencia de inteligencia artificial se estableció en una instancia independiente de Ollama (2023), desacoplada de la interfaz y comunicada remotamente mediante solicitudes a una API REST cifrada. Esta separación funcional se concibió para evitar la conexión directa del agente conversacional con las bases de datos institucionales y limitar la exposición de los recursos críticos de la universidad, manteniendo el procesamiento asociado con el modelo de lenguaje separado de la infraestructura institucional sensible.

La capa de ciberseguridad se diseñó mediante la incorporación de controles situados antes y después de la interacción con el modelo de lenguaje. Para la validación de las entradas se empleó Pydantic, integrado mediante un script destinado a verificar la estructura de los datos y aplicar reglas de filtrado definidas para el sistema. La mitigación de intentos de inyección de prompts se implementó mediante Rebuff.ai (2023) utilizado como mecanismo perimetral para analizar las entradas antes de su procesamiento por el modelo de lenguaje. Este componente combina procedimientos heurísticos y filtrado vectorial para identificar patrones potencialmente asociados con intentos de manipulación de las instrucciones del sistema.

Para disminuir el riesgo de exposición de información sensible en las respuestas se incorporó Microsoft (2023) como mecanismo de análisis de las salidas generadas por el modelo. Este componente examina el texto antes de su presentación en la interfaz del usuario con el propósito de identificar información sensible, entre ella nombres, direcciones de correo electrónico, números de identificación y otros datos susceptibles de protección. De esta manera, la propuesta establece controles perimetrales

independientes del modelo de lenguaje tanto para las entradas como para las salidas, de modo que las solicitudes potencialmente maliciosas puedan ser interceptadas antes de alcanzar el motor de inferencia y las respuestas sean examinadas antes de ser entregadas al usuario.

El flujo operacional se estructuró de manera secuencial. El usuario accede inicialmente a la interfaz web, incorpora el material documental permitido y formula una consulta. La entrada atraviesa los mecanismos de validación y filtrado establecidos en la capa perimetral de seguridad y, una vez superados estos controles, se procesa la información correspondiente y se realiza, cuando procede, la recuperación de contenido mediante ChromaDB. La solicitud validada se transmite posteriormente al motor de inferencia desacoplado a través de la API REST cifrada. La respuesta generada retorna a la capa de ciberseguridad, donde es examinada mediante los mecanismos establecidos para identificar información sensible, antes de ser presentada en la interfaz. El procedimiento configura así un perímetro de protección alrededor de la interacción con el modelo y mantiene desacoplados los componentes que intervienen en la gestión, procesamiento y presentación de la información.

Finalmente, los datos obtenidos mediante el cuestionario se procesaron mediante estadística descriptiva. Las respuestas se organizaron de acuerdo con las tres dimensiones establecidas en el instrumento y se calcularon frecuencias y porcentajes para caracterizar las tendencias observadas en la muestra. Los resultados del diagnóstico se utilizaron como insumo para establecer los requerimientos considerados en la especificación de la arquitectura del agente conversacional, articulando las necesidades identificadas en el contexto docente con los componentes técnicos y controles de ciberseguridad incorporados en la propuesta.

RESULTADOS Y DISCUSIÓN

Los resultados obtenidos a partir de los 105 docentes universitarios evidencian una elevada incorporación de herramientas de inteligencia artificial (IA) en las prácticas académicas, acompañada, sin embargo, de brechas importantes en el conocimiento de los riesgos asociados con su utilización. La Tabla 1 sintetiza los principales indicadores mediante frecuencias absolutas (n) y relativas (%). El 76,20 % de los participantes declara utilizar herramientas de IA de manera frecuente o muy frecuente, lo que confirma su creciente integración en las actividades docentes. Este nivel de adopción contrasta con el conocimiento sobre las amenazas de ciberseguridad, pues el 84,80 % manifiesta un desconocimiento total o parcial acerca de la inyección de prompts. La diferencia entre ambos indicadores evidencia que la incorporación de estas tecnologías avanza con mayor rapidez que la preparación de los usuarios para reconocer determinados riesgos derivados de su empleo.

Esta situación adquiere especial relevancia al analizar las prácticas relacionadas con el tratamiento de información académica. El 25,71 % de los docentes reconoce conductas que pueden comprometer la privacidad de los datos, entre ellas la incorporación de tareas, calificaciones o información personal en plataformas de IA. La coexistencia de una elevada frecuencia de uso con un conocimiento limitado de determinadas amenazas demuestra que la protección de la información no debería depender exclusivamente del comportamiento individual del usuario. Aunque la formación y las buenas prácticas constituyen componentes necesarios, los resultados respaldan la incorporación de mecanismos tecnológicos capaces de detectar, filtrar o bloquear automáticamente interacciones que puedan representar un riesgo para la confidencialidad de la información.

Los requerimientos funcionales identificados refuerzan esta necesidad. El 61,90 % de los participantes prioriza la incorporación de mecanismos de ciberseguridad activa en el agente conversacional, mientras que el 52,40 % considera relevante implementar estrategias para reducir las alucinaciones mediante técnicas de Generación Aumentada por Recuperación (RAG). A su vez, el 49,50 % señala como prioridad que el asistente virtual pueda explicar conceptos complejos de manera sencilla. Estos resultados muestran que las expectativas docentes no se concentran únicamente en la capacidad generativa del sistema, sino que integran tres componentes fundamentales: seguridad de la interacción, confiabilidad de las respuestas y utilidad pedagógica.

Los hallazgos revelan una brecha entre la rápida adopción de herramientas de IA y el nivel de preparación para afrontar determinados riesgos asociados con su utilización. Esta relación sustenta la necesidad de incorporar la ciberseguridad desde el diseño del agente conversacional, en lugar de trasladar exclusivamente al usuario la responsabilidad de proteger la información. Desde esta perspectiva, los resultados respaldan la propuesta de una arquitectura institucional descentralizada y desacoplada, complementada con controles perimetrales para la validación de entradas y salidas, mecanismos orientados a mitigar la fuga de datos y la inyección de prompts, y técnicas RAG destinadas a reducir la generación de respuestas no sustentadas. De este modo, los requerimientos identificados en el diagnóstico docente se vinculan directamente con las decisiones arquitectónicas y los mecanismos de protección incorporados en la propuesta tecnológica.

Tabla 1. Indicadores clave del diagnóstico docente y ciberseguridad.

Dimensión	Indicador Evaluado	Criterio Dominante / Hallazgo	Frecuencia (n)	Porcentaje (%)
Uso y adopción de IA generativa por los docentes.	Frecuencia de uso de IA en docencia.	Frecuente o muy frecuente	80	76.20%
	Actividades docentes con IA.	Retroalimentar trabajos o resumir textos.	57	54.3%
	Información subida a la IA.	Calificaciones, tareas y datos personales.	27	25.71%
Competencias de ciberseguridad y conciencia de privacidad.	Destino de la información.	Son usados por la empresa de IA	37	35.20%
		Pueden filtrarse o quedar expuesto a otros usuarios.	27	25.70%
	Privacidad de los datos.	Poco preocupado	27	24.80%
		Moderadamente preocupado	28	26.70%
	Amenaza por inyección de prompt.	Desconocimiento total.	52	48.60%
		Desconocimiento parcial no sabe cómo prevenirlo.	38	36.20%
Requerimientos funcionales del chatbot con el rol de asistente virtual.	Función prioritaria del chatbot.	Explicar conceptos complejos de forma sencilla.	52	49.50%
	Fidelidad pedagógica	Mitigación de alucinaciones.	55	52.40%
		Filtros anti-manipulación (Guardrails)	65	61.90%

La arquitectura propuesta, sintetizada en la Tabla 2, responde a las preocupaciones de ciberseguridad y privacidad identificadas en el diagnóstico docente mediante la integración de controles especializados en diferentes puntos del flujo de información. Rebuff.ai y Pydantic conforman una primera barrera perimetral destinada a validar las entradas y detectar intentos de inyección de prompts o manipulación lógica antes de que las solicitudes alcancen el motor de inferencia. Esta configuración desplaza parte de la protección desde la conducta individual del usuario hacia mecanismos incorporados en la propia arquitectura y contribuye a establecer un entorno de asistencia académica con mayores controles de seguridad y privacidad.

La protección de los datos se complementa mediante Microsoft Presidio, incorporado para identificar información personal o sensible en el contenido procesado antes de su presentación en la interfaz. Este mecanismo permite detectar y restringir la exposición de elementos como nombres, direcciones de correo electrónico, números de identificación u otros datos susceptibles de protección. De este modo, la arquitectura establece controles tanto sobre las entradas dirigidas al modelo como sobre las salidas generadas, configurando un esquema de protección perimetral orientado a reducir el riesgo de fuga de información.

La propuesta también responde a los requerimientos ético-pedagógicos identificados en el diagnóstico, particularmente a la necesidad de reducir las alucinaciones, señalada por el 52,40 % de los docentes. Para ello, ChromaDB funciona como base de datos vectorial para la recuperación de información procedente del material bibliográfico incorporado al sistema, proporcionando contexto documental al modelo mediante un esquema de Generación Aumentada por Recuperación (RAG). Esta estrategia busca fundamentar las respuestas en las fuentes disponibles y disminuir la generación de contenidos no sustentados, sin asumir que el riesgo de alucinación pueda eliminarse completamente.

Desde el punto de vista de la implementación, la interfaz y la base de datos vectorial embebida en ChromaDB se ejecutan sobre Streamlit Cloud, mientras que la inferencia generativa se delega a una instancia remota y desacoplada de Ollama. Esta separación permite distribuir los componentes funcionales y evitar que el agente conversacional dependa de una conexión directa con las bases de datos institucionales. Asimismo, reduce la necesidad de desplegar infraestructura computacional especializada dentro de la universidad y favorece una implementación ligera basada predominantemente en herramientas de código abierto y servicios externos. La Tabla 2 resume la correspondencia entre los componentes de la arquitectura, los riesgos que atienden y los requerimientos ético-pedagógicos identificados en el diagnóstico.

Tabla 2. Especificación conceptual de la arquitectura y su alineación ético-pedagógica.

Capa	Componente	Función del diseño	Alineación ética / pedagógica
Interfaz	Streamlit Cloud	Interfaz de usuario	Aislamiento seguro de sesiones de usuario.
Flujo de datos e inferencia generativa	ChromaDB (Embebida)	Indexación de documentos institucionales.	Anti-alucinación: Las respuestas son en base a documentos cargados en el chatbot. Requerimiento del (52.38% del diagnóstico).
	Ollama / Llama-3	Motor de lenguaje desacoplado.	Privacidad de datos al no enviar telemetría a APIs comerciales.
Ciberseguridad	Microsoft Presidio	Intercepción de prompts previo al procesamiento	Información de identificación personal se sanitiza los datos sensibles. Requerimiento del (84.76% del diagnóstico).
	Rebuff.ai + Pydantic	Validación de peticiones y respuestas.	Detecta ataques de inyección de prompt y fuerza el formato estricto de las respuestas.

El diseño de la arquitectura descentralizada y desacoplada para el agente conversacional se corresponde con la literatura reciente sobre la integración de la Inteligencia Artificial Generativa en la educación superior y responde a tres barreras críticas identificadas en este ámbito: el costo de la infraestructura, la precisión pedagógica y la gobernanza de los datos.

Diversos estudios recientes han documentado el despliegue de asistentes virtuales académicos respaldados por modelos comerciales alojados en la nube, como OpenAI (GPT-4) o Anthropic (Claude) (Gamez Lagos, 2025; Toro et al., 2025). Si bien estas implementaciones han demostrado altos niveles de fluidez conversacional, también introducen una dependencia financiera asociada con las tarifas por consumo de tokens y, por consiguiente, costos operativos recurrentes que pueden limitar su escalabilidad en universidades con presupuestos restringidos.

En el diseño propuesto, la interfaz y la base de datos vectorial embebida, ChromaDB, se integran en Streamlit Cloud, mientras que la inferencia generativa se delega a una instancia remota y desacoplada de Ollama. Esta configuración reduce los requerimientos de infraestructura tecnológica institucional y evita la dependencia directa de servicios comerciales de inferencia basados en el consumo de tokens. El enfoque coincide con Gayathri et al. (2025), quienes argumentan que el desacoplamiento entre el middleware lógico y la inferencia de mayor demanda computacional constituye una alternativa para ampliar el acceso a soluciones de IA en instituciones públicas.

Otro problema recurrente reportado en los estudios sobre LLM aplicados a la enseñanza es la susceptibilidad de estos modelos a generar respuestas erróneas, alucinaciones o contenidos descontextualizados respecto del currículo institucional (Johno et al., 2026; Saad et al., 2025). Dzhaliuk et al. (2026); y Giapantzis et al. (2026) abordan el ajuste fino como una de las alternativas para mejorar el comportamiento de estos sistemas. No obstante, Dzhaliuk et al. (2026) señalan las exigencias computacionales asociadas con este proceso, a lo que se suma la necesidad de actualizar periódicamente los contenidos académicos cuando estos cambian entre ciclos formativos.

Frente a esta problemática, en la arquitectura propuesta el LLM opera mediante un esquema de recuperación contextual sustentado en ChromaDB, embebido en la aplicación, que permite proporcionar al modelo información procedente del material bibliográfico previamente incorporado. Esta configuración responde directamente a la preocupación expresada por el 52,38 % de los docentes de la muestra respecto de la necesidad de reducir las alucinaciones y favorece que las respuestas se fundamenten en los contenidos académicos disponibles para la consulta. El planteamiento se encuentra en consonancia con Li et al. (2025), quienes destacan la necesidad de establecer límites semánticos en los entornos de tutoría automatizada.

En materia de ciberseguridad, el análisis de la literatura permitió identificar que algunos modelos de chatbots académicos no incorporan de manera suficiente capas intermedias para la validación y protección de las interacciones (Alafnan et al., 2023; Kumar et al., 2024). En contraste, la arquitectura propuesta incorpora mecanismos específicos de seguridad en diferentes puntos del flujo de información y responde a las necesidades identificadas en el diagnóstico institucional, en el que el 84,76 % de los docentes manifestó preocupación o conocimientos insuficientes relacionados con la privacidad y los riesgos asociados con el uso de estas tecnologías.

La integración de diferentes capas de ciberseguridad permite plantear un entorno de asistencia académica en el que la protección de los datos no depende exclusivamente de las prácticas del usuario, sino también de controles incorporados en la propia arquitectura. La combinación del desacoplamiento de los componentes, los mecanismos de validación de entradas, la detección de posibles intentos de inyección de prompts, el control de información sensible y la recuperación contextual mediante una base vectorial constituye una respuesta tecnológica integrada a las necesidades de seguridad, privacidad y confiabilidad identificadas en el diagnóstico docente.

Asimismo, el empleo predominante de herramientas de código abierto y servicios de acceso gratuito reduce la dependencia de licencias propietarias y favorece la viabilidad de la propuesta en instituciones con recursos tecnológicos limitados.

No obstante, el diseño arquitectónico presenta limitaciones que deben considerarse al interpretar sus alcances. En esta fase no se realizaron pruebas empíricas de latencia y rendimiento bajo cargas masivas de usuarios simultáneos ni se evaluó cuantitativamente el impacto del agente conversacional sobre el rendimiento académico de los estudiantes. Tampoco corresponde interpretar los mecanismos incorporados como una eliminación absoluta de los riesgos de fuga de datos, inyección de prompts o alucinaciones, sino como controles destinados a su mitigación. Estos aspectos deberán ser examinados en fases posteriores mediante pruebas de estrés, evaluaciones específicas de ciberseguridad y estudios empíricos que permitan determinar el desempeño técnico y el impacto educativo de la arquitectura en condiciones reales de implementación.

CONCLUSIONES

Los resultados de la investigación demuestran que es posible desplegar un chatbot inteligente, descentralizado y desacoplado articulado en una capa de presentación y de almacenamiento del material bibliográfico embebido en (ChromaDB) sobre la plataforma serverless de Streamlit Cloud, conectada de forma remota a un motor de inferencia desacoplado (Ollama / Llama-3). Esta configuración elimina por completo los costos de inversión aplicable a instituciones con restricciones presupuestarias.

Las exigencias de privacidad manifestadas por el 84.76% de los encuestados se integra en la capa de ciberseguridad integrada por Pydantic para la validación sintáctica, Rebuff.ai para la detección heurística y vectorial de inyecciones de prompts, y Microsoft Presidio para la información sensible o fuga de datos, cambiando a estos por etiquetas genéricas o valores ficticios, antes de que el texto sea procesado por el modelo de lenguaje o mostrado en la pantalla. El diseño del agente conversacional se consolida como un marco de referencia replicable, de código abierto, riguroso y alineado a la ética de la educación superior.

REFERENCIAS

Alafnan, M. A., Dishari, S., Jovic, M., & Lomidze, K. (2023). ChatGPT as an educational tool: Opportunities, challenges, and recommendations for communication, business writing, and composition courses. ***Journal of Artificial Intelligence and Technology, 3***(2), 60–68. https://doi.org/10.37965/jait.2023.0184

Anghel, C., Craciun, M. V., Cocu, A., Anghel, A. A., & Istrate, A. (2026). SteadyEval: Robust LLM exam graders via adversarial training and distillation. ***Computers, 15***(1). https://doi.org/10.3390/computers15010055

Chroma Core Inc. (2024). ***The open-source embedding database*** (Version 0.5.x) [Software]. https://www.trychroma.com/

Cotton, D. R. E., Cotton, P. A., & Shipway, J. R. (2024). Chatting and cheating: Ensuring academic integrity in the era of ChatGPT. ***Innovations in Education and Teaching International, 61***(2), 228–239. https://doi.org/10.1080/14703297.2023.2190148

De Freitas, J., & Cohen, I. G. (2025). Disclosure, humanizing, and contextual vulnerability of generative AI chatbots. ***NEJM AI, 2***(2). https://doi.org/10.1056/aipc2400464

Dzhaliuk, N., Sabodashko, D., Khoma, V., Khoma, Y., Kolchenko, V., & Podpora, M. (2026). Comparative evaluation of machine learning methods for protecting LLMs from prompt injection attacks. ***International Journal of Information Security, 25***. https://doi.org/10.1007/s10207-026-01264-8

Gamez Lagos, N. M. (2025). ***Asistente virtual académico basado en sistemas LLM para procesar y extraer información normativa de instituciones de educación superior***. https://www.researchgate.net/publication/395710153_Asistente_Virtual_Academico_Basado_en_Sistemas_LLM_para_Procesar_y_Extraer_Informacion_Normativa_de_Instituciones_de_Educacion_Superior

Gayathri, T., Durusoju, S., Mergoju, L., & Reddy, P. V. (2025). MedicAI: AI-powered system for enhanced medical information access. ***9th International Conference on Inventive Systems and Control (ICISC)***. Coimbatore, India.

Giapantzis, K., Bountakas, P., Zarras, A., Farao, A., Bolgouras, V., & Xenakis, C. (2026). Adaptive DeSeTra: Adaptive deformable-span self-attention transformer for LLM security. ***Information Sciences, 754***. https://doi.org/10.1016/j.ins.2026.123670

Gundu, T. (2023). Chatbots: A framework for improving information security behaviours using ChatGPT. En S. Furnell & N. Clarke (Eds.), ***Human aspects of information security and assurance: HAISA 2023*** (Vol. 674, pp. 418–431). Springer.

Johno, H., Amakawa, A., Komaba, A., Tozuka, R., Johno, Y., Sato, J., Yoshimura, K., Nakamoto, K., & Ichikawa, S. (2026). Open-source offline-deployable retrieval-augmented large language model for assisting pancreatic cancer staging. ***Japanese Journal of Radiology***. https://doi.org/10.1007/s11604-026-02049-8

Kumar, S. S., Cummings, M. L., & Stimpson, A. (2024). Strengthening LLM trust boundaries: A survey of prompt injection attacks. *4th International Conference on Human-Machine Systems (ICHMS)*. Toronto, Canada.

Li, Y., Fu, X., Verma, G., Buitelaar, P., & Liu, M. (2025). *Mitigating hallucination in large language models (LLMs): An application-oriented survey on RAG, reasoning, and agentic systems*. arXiv. http://arxiv.org/abs/2510.24476

Luo, H., & Ji, C. (2025). Federated learning-based data collaboration method for enhancing edge cloud AI system security using large language models. *5th International Symposium on Computer Technology and Information Science (ISCTIS 2025)* (pp. 163–166). IEEE. Xi'an, China.

Microsoft. (2023). *Presidio: Data protection and anonymization SDK* [Software]. Microsoft Open Source. https://presidio.dataprivacystack.org/

Ollama. (2023). *Get up and running with Llama 3 and other large language models locally* [Software]. https://ollama.com/

Rebuff AI. (2023). *Rebuff: Prompt injection detection service for LLM applications* [Software]. GitHub. https://github.com/rebuff-ai/rebuff

Saad, M., Mughal, S. F., Mustafa, M., Ahmed, I., & Masood, A. (2025). Assessing and mitigating multi-turn jailbreak vulnerabilities in large language models: A Crescendo attack study. *27th International Multitopic Conference (INMIC 2025)*. Islamabad, Pakistan.

Streamlit. (2026). *Streamlit Community Cloud*. Snowflake Inc. https://share.streamlit.io/

Toro, Y., Toro, I., Calvo, C., Urquizo, D., & Chapalbay, M. (2025). Implementación de un asistente virtual de inteligencia artificial en universidades latinoamericanas. *Invecom*, *6(*3). https://revistainvecom.org/index.php/invecom/article/view/4258/4885

Trinks, M., Gondim, J., & Albuquerque, R. (2023). Multi-agent architecture for passive rootkit detection with data enrichment. In M. V. Garcia & C. Gordón-Gallegos (Eds.), *CSEI: International Conference on Computer Science, Electronics and Industrial Engineering (CSEI 2022)* (Vol. 678, pp. 29–41). Springer.

Conflictos de interés:

Los autores declaran no tener conflictos de interés.

Contribución de los autores:

Jorge Wilson Saa-Saltos, Nancy Magali Rodríguez-Gavilanes: Conceptualización, curación de datos, análisis formal, investigación, metodología, supervisión, validación, visualización, redacción del borrador original y redacción, revisión y edición.

Declaración ética:

La participación fue voluntaria e informada. Se garantizó el anonimato, la confidencialidad y el uso de la información recopilada exclusivamente con fines de investigación.