▶ Recibido: 1 de marzo de 2026
▶ Aceptado: 28 de abril de 2026

Aprender con la tecnología: Chat portátil de IAGen con corpus para el aprendizaje de la Historia en secundaria y público en general

Learning with technology: a portable corpora GenAI chatbot for secondary education and the general public history' learning

Karla Karina Ruiz-Mendoza 0000-0001-8978-8364
Universidad Autónoma de Baja California, México | ruiz.karla32@uabc.edu.mx
Autora de correspondencia

Azucena Yoselin González-García 0000-0002-1217-7247
Universidad Autónoma del Estado de Hidalgo, México | go477665@uaeh.edu.mx

Resumen. Este artículo sistematiza el desarrollo de una Biblioteca sin Internet (BiS), mediante un chat educativo portátil basado en inteligencia artificial generativa (IAGen), grandes modelos de lenguaje (LLM) y Generación Aumentada por Recuperación (RAG), para apoyar la enseñanza de la Historia en secundaria. El estudio corresponde a una fase preempírica de diseño tecnológico: se revisaron código, corpus, parámetros de indexación y registros de ejecución local. El prototipo opera sin internet mediante un LLM local, fragmentos documentales, *embeddings* e índice FAISS, de modo que las respuestas están condicionadas a fuentes previamente seleccionadas. Los resultados muestran factibilidad técnica y latencias de 22 a 43 segundos para responder consultas, esa variación de tiempo depende también del tamaño del *prompt* (instrucciones). Se concluye que una BiS puede funcionar como biblioteca conversacional situada, siempre que el corpus sea transparente, plural y acompañado por actividades que promuevan contextualización, causalidad y verificación histórica.
Versión en lengua de señas mexicana

Palabras clave: involucramiento parental, rendimiento académico, contexto familiar.

Abstract. This article systematizes the development of an Offline Library (BiS, by its Spanish acronym), implemented through a portable educational chatbot powered by generative artificial intelligence (GenAI), large language models (LLMs), and Retrieval-Augmented Generation (RAG) to support secondary school history instruction. The study corresponds to a pre-empirical technological design phase, evaluating code, corpora, indexing parameters, and local execution logs. Operating offline, the prototype relies on a local LLM, document chunks, embeddings, and a FAISS index, ensuring that responses remain grounded in pre-selected sources. The results demonstrate technical feasibility, with response latencies ranging from 22 to 43 seconds depending on prompt size and instruction complexity. The study concludes that a BiS can effectively serve as a situated conversational library, provided the corpus remains transparent, pluralistic, and accompanied by instructional activities that foster historical contextualization, causality, and source verification.
Keywords: artificial intelligence; educational technology; history teaching; information and communication technologies; secondary education.

Introducción

La inteligencia artificial (IA) se ha consolidado como infraestructura educativa y herramienta cotidiana. Sus usos abarcan desde la personalización y la tutoría hasta el apoyo a la evaluación y la elaboración de materiales didácticos (Echeverría Quiñonez y Otero Mendoza, 2025). En particular, la IA generativa (IAGen) puede producir contenido nuevo —texto, imagen, audio o video—, a partir de instrucciones y mediante interfaces conversacionales en lenguaje natural (Fuertes Alpiste, 2024).

Dentro de la IAGen, los grandes modelos de lenguaje (LLM, por sus siglas en inglés) se volvieron centrales porque permiten una interacción relativamente natural donde el usuario pregunta, el sistema responde; el usuario precisa, el sistema ajusta. En educación, la literatura reciente los describe como motores de chatbots y asistentes capaces de explicar temas, contestar dudas, generar cuestionarios, resumir contenidos y apoyar la planeación docente (Li et al., 2025). La rapidez de esta expansión obliga a desplazar la discusión desde la novedad tecnológica hacia las condiciones pedagógicas, éticas y curriculares de uso.

Sin embargo, la pregunta central no es si la IAGen sirve o no sirve, sino bajo qué condiciones pedagógicas puede aportar al aprendizaje. La investigación reciente ha documentado beneficios potenciales, como personalización, creatividad y pensamiento crítico, pero también subraya que el impacto depende de la mediación docente, diseño de actividades y consideraciones éticas explícitas (Echeverría Quiñonez y Otero Mendoza, 2025). En

ISSN 2594-1828 • www.eduscientia.com

Historia, estas tensiones son especialmente visibles: los textos generados pueden parecer correctos y bien escritos, pero obligan a replantear tareas de escritura, evaluación formativa y evaluación sumativa (Kindenberg, 2024).

Hay, además, un límite técnico que en el proceso de aprendizaje se vuelve pedagógico: la alucinación (respuestas plausibles, pero incorrectas) y la estaticidad del conocimiento del modelo. En un área donde la producción de conocimiento epistemológicamente válido importa —como en la Historia—, esas fallas reducen confianza y pueden distorsionar aprendizajes (Li et al., 2025). Por eso, ha tomado fuerza un enfoque arquitectónico que cambia la lógica del sistema: la Generación Aumentada por Recuperación (RAG, por sus siglas en inglés). La idea es que antes de responder, el sistema busque evidencia en un corpus externo y después, genere la respuesta con base en esos fragmentos recuperados (Li et al., 2025). En educación se ha señalado que la RAG ataca precisamente el problema que más compromete a los chatbots basados en LLM: las respuestas fluidas sin sustento, y que su implementación puede ser relativamente directa cuando el objetivo es específico (Swacha y Gracel, 2025).

Con ese trasfondo, este artículo documenta —con enfoque reproducible— el desarrollo de un chat portátil de IAGen especializado por corpus para Historia Universal, dirigido principalmente a secundaria, pero utilizable por público general. El fin no es entrenar el modelo desde cero, sino anclarlo a un conjunto curado de archivos PDF y permitir la sustitución rápida de esos materiales para especializar el tema (por ejemplo, historia de México, ciencias o literatura), sin rehacer la arquitectura. En esa lógica, la IAGen no opera como docente artificial, sino como herramienta que solo cobra sentido dentro de un marco pedagógico definido (Fuertes Alpiste, 2024).

El problema educativo que orienta el trabajo es la necesidad de contar con recursos de consulta histórica que sean verificables, disponibles sin conectividad y capaces de estimular pensamiento histórico, no solo la recuperación automática de datos. La pregunta de investigación fue: ¿Cómo diseñar y documentar un chat educativo portátil, basado en RAG y corpus local, para apoyar la enseñanza de la Historia en secundaria con criterios de trazabilidad y mediación pedagógica? El objetivo fue sistematizar el diseño técnico-pedagógico del prototipo BiS. Como supuesto analítico, se planteó que un corpus curado, un *prompt* con reglas de evidencia y un uso guiado por docentes, pueden convertir el chat en mediador para contextualizar, corroborar y explicar procesos históricos. La población destinataria proyectada corresponde a estudiantes y docentes de secundaria, aunque esta fase no trabajó con muestra ni participantes.

ISSN 2594-1828 • www.eduscientia.com

Eduscientia. Divulgación de la ciencia educativa
Año IX • n.º 18 • agosto 2026

Marco teórico

IAGen y LLM en educación

Se ha descrito a la IAGen como un conjunto de herramientas capaces de apoyar tareas escolares y académicas variadas: estructurar ideas, reformular, guiar búsquedas, resumir, descomponer tareas complejas, acompañar procesos de escritura y producir contenido multimodal (Fuertes Alpiste, 2024). Ese abanico explica su expansión como apoyo didáctico; pero su utilidad queda condicionada por un riesgo básico: que produzca información falsa o no verificable si el sistema no está "amarrado" a fuentes (Li et al., 2025).

En educación, la discusión empírica sugiere que el foco no debería encerrarse en el plagio, sino desplazarse hacia un uso más formativo: apoyo para diseñar tareas, organizar entregas, acompañar escritura y abrir espacios de revisión crítica (Kindenberg, 2024). Si la IAGen entra al aula como ayuda, conviene diseñarla para que no sustituya el trabajo intelectual del estudiante, sino que lo ordene, lo empuje y lo obligue a pensar mejor (Echeverría Quiñonez y Otero Mendoza, 2025).

Aquí, la arquitectura RAG se vuelve una justificación técnica con consecuencias didácticas, al recuperar primero información relevante desde un corpus y generar respuestas después con base en esa evidencia; así, el sistema puede mejorar exactitud y transparencia, reduciendo el margen de respuestas bien escritas, pero infundadas (Li et al., 2025). Y no solo importa el cómo funciona, sino también cómo se evalúa. En revisiones de chatbots RAG en educación aparecen criterios recurrentes como fidelidad a fuentes, cobertura, pertinencia y utilidad, directamente conectados con la validez práctica del recurso en aula (Swacha y Gracel, 2025).

De aprender de la tecnología a aprender con la tecnología

La incorporación de tecnologías digitales en los procesos de enseñanza y aprendizaje ha sido objeto de análisis desde sus primeras fases de implementación. Diversos autores establecieron una distinción conceptual clave entre aprender de la tecnología y aprender con la tecnología, subrayando que esta diferencia no es meramente lingüística, sino epistemológica y pedagógica (Derry y Lajoie, 1993; Salomon et al., 1991; Jonassen, 1996).

El enfoque de aprender de la tecnología se consolidó con el desarrollo de la Instrucción Asistida por Computadora (IAC) y los sistemas de tutoría inteligente (STI). Estas propuestas se apoyaron inicialmente en marcos conductistas —centrados en la relación estímulo–respuesta y el refuerzo—, y posteriormente en perspectivas cognitivistas que atendían a procesos como la memoria de trabajo, la organización de esquemas previos,

ISSN 2594-1828 • www.eduscientia.com

la recuperación de información y la retroalimentación personalizada. En este modelo, la tecnología asume el rol de agente instructivo, ya que presenta contenidos, formula ejercicios y proporciona retroalimentación, reproduciendo una lógica educativa centrada en la transmisión y con una participación limitada del estudiante en la regulación de su propio aprendizaje (Derry y Lajoie, 1993; Salomon et al., 1991).

En contraste, el enfoque de aprender con la tecnología desplaza el eje hacia el estudiante y redefine el papel de las herramientas digitales. En lugar de funcionar como tutor automatizado, la tecnología se concibe como instrumento cognitivo que posibilita realizar tareas intelectuales de mayor complejidad. Desde esta perspectiva, la actividad mental no se delega en el sistema, sino que permanece en el aprendiz, quien utiliza las herramientas para representar, explorar y transformar información de maneras que serían difíciles o inviables sin su mediación (Jonassen, 1996).

Este planteamiento se fundamenta en el constructivismo, entendido como una orientación pedagógica en la que el docente diseña situaciones de aprendizaje con objetivos definidos, donde el estudiante construye activamente el conocimiento. Iiyoshi et al. (2005) describen este proceso a través de cinco operaciones cognitivas interrelacionadas: búsqueda de información, presentación de información, organización del conocimiento, integración del conocimiento y generación de conocimiento. La tecnología, en este marco, no transmite contenidos de forma unidireccional, sino que amplía las posibilidades de representación, articulación y producción intelectual y, por otro lado, reta a los profesores a buscar, diseñar y usar nuevas estrategias para la realización de tareas y exámenes escolares que exijan el desarrollo del pensamiento complejo, crítico y creativo de los alumnos.

Asimismo, esta concepción puede vincularse con el construccionismo propuesto por Papert (1982), en el que el aprendizaje se fortalece cuando los estudiantes crean artefactos significativos que materializan sus ideas. En esta línea, la actividad con tecnología no se limita a consumir información, sino que se orienta a la producción y externalización del pensamiento. De igual forma, desde una perspectiva socioconstructivista, el aprendizaje emerge de la interacción entre sujetos y entorno, la tecnología actúa a modo de mediadora en dinámicas colaborativas donde el conocimiento se construye de manera compartida.

En esa línea, Fuertes Alpiste (2024) propone entender las aplicaciones de IAGen como herramientas para la cognición (HPC/mindtools), siempre que el diseño de la actividad preserve del lado del aprendiz las funciones ejecutivas, entre las que están el decidir, evaluar, argumentar, verificar, sintetizar. Eso importa especialmente en un chat educativo porque el

ISSN 2594-1828 • www.eduscientia.com

Eduscientia. Divulgación de la ciencia educativa
Año IX • n.º 18 • agosto 2026

objetivo no es que el sistema conteste por el alumno, sino que sostenga procesos: formular preguntas, organizar conocimiento y producir reflexión crítica, sin reemplazarlo (2024).

Importancia de Historia y de los recursos didácticos: comprensión histórica y mediación por textos

En la enseñanza de la Historia, la herramienta (sea libro, recurso digital o chatbot) debe sostener la finalidad de reflexionar las acciones ocurridas en el pasado, no solo transmitir datos. En investigación sobre educación histórica, se han conceptualizado dimensiones de comprensión histórica como posturas o *stances* que incluyen identificación, análisis causal, respuesta moral y capacidad de exhibir y organizar la información histórica (Barton y Levstik, 2004, citados en Kindenberg, 2024). Esta conceptualización es útil para orientar un chat; una respuesta correcta no basta si no ayuda al estudiante a explicar causas y consecuencias, ubicar contextos o justificar interpretaciones.

Los recursos didácticos (en particular los textos escolares) no son neutrales, ya que suelen reproducir lo que se considera oficial, junto con normas, valores y categorías. Los estudios sobre libros de texto sostienen que estos funcionan como instrumentos que transmiten conocimiento sistemático y significados culturales, a veces de modo implícito (Luku, 2024). Para un chat especializado por corpus eso implica que seleccionar los PDF no es una tarea solo técnica, sino una decisión curricular y cultural porque el corpus delimita la narrativa histórica, los énfasis regionales y las perspectivas disponibles.

En el campo de la enseñanza de la Historia se han reportado experiencias donde la IA se integra para incrementar motivación, participación y comprensión en contenidos históricos, especialmente cuando se acompañan de estrategias docentes (Ibarra Álvarez et al., 2026). En un nivel más especializado, las propuestas metodológicas para el aprendizaje de la historia (por ejemplo, Historia Antigua) han argumentado que el modelo tradicional centrado en exposición y memorización ha sido superado y, por tanto, conviene transitar hacia metodologías activas apoyadas en tecnologías emergentes, incluida la IAGen (Cabezas Guzmán y Rovira Marcelino, 2025).

Marco conceptual para un chat portátil: corpus sustituible más RAG como control de calidad

Con base en lo anterior, un chat portátil para la enseñanza y el aprendizaje de contenidos históricos se justifica como una implementación donde la especialización reside en el corpus y en el diseño de interacción, no en atribuir autoridad epistémica al modelo.

ISSN 2594-1828 • www.eduscientia.com

En términos técnicos, el objetivo clave es operar con un esquema de generación aumentada por recuperación (RAG): (a) indexación del corpus, (b) recuperación de fragmentos pertinentes ante una consulta, y (c) generación condicionada por la evidencia recuperada (Li et al., 2025). Esta arquitectura responde a dos límites estructurales de los LLM en contextos escolares: la tendencia a producir respuestas plausibles, pero incorrectas, y la dificultad para reflejar actualizaciones del conocimiento con precisión. Al forzar la respuesta apoyándose en un conjunto explícito de textos, RAG opera como un control de calidad de primer orden para tareas de alta sensibilidad factual, como Historia escolar (2025).

La literatura reciente no solo describe el flujo técnico de RAG, sino también sus usos educativos y formas de evaluación. Li et al. (2025) sintetizan aplicaciones para sistemas interactivos —preguntas y respuestas, chatbots y tutoría—, desarrollo de contenidos y despliegues en ecosistemas educativos; en todos los casos, la premisa es elevar precisión y trazabilidad mediante anclaje documental. Asimismo, Swacha y Gracel (2025) muestran que los chatbots educativos con RAG suelen evaluarse por fidelidad a fuentes, pertinencia, cobertura y utilidad pedagógica, además de reportar decisiones de diseño como modelo, corpus y estrategia de recuperación.

Materiales y método

Diseño del estudio y propósito metodológico

El trabajo se planteó como un estudio de desarrollo tecnológico con sistematización metodológica y alcance descriptivo-propositivo. Su propósito fue documentar, de manera replicable y comprensible para una audiencia educativa, la construcción de un prototipo de chat portátil de IAGen para la enseñanza de la Historia en secundaria mediante RAG y corpus local. No se buscó medir impacto en el aprendizaje, sino describir el diseño técnico-pedagógico del sistema, sus materiales, parámetros de operación e implicaciones didácticas. La población destinataria proyectada son estudiantes y docentes de fase 6 de educación básica en México; sin embargo, al tratarse de una fase preempírica, no hubo población participante, muestra o aplicación en aula.

Las fuentes de información fueron: (a) el código y los archivos de interfaz; (b) el corpus documental en PDF; (c) los archivos de persistencia del índice; (d) la bitácora de ejecución; y (e) el *prompt* pedagógico. Las técnicas de recolección consistieron en revisión documental del prototipo, extracción de parámetros técnicos, registro de métricas de latencia y descripción de decisiones de curaduría. El análisis fue descriptivo y se orientó por

ISSN 2594-1828 • www.eduscientia.com

criterios de reproducibilidad, trazabilidad de fuentes, operabilidad sin internet y pertinencia para promover pensamiento histórico.

Unidad de análisis y materiales revisados

La unidad de análisis fue el prototipo funcional BiS y su pipeline de consulta. Para evitar una presentación excesivamente técnica, se revisaron los componentes solo en la medida en que explican su uso educativo: interfaz local, carpeta de textos, segmentación del corpus, recuperación de fragmentos y generación de respuestas con un LLM local. Esta revisión permitió reconstruir cómo una pregunta del usuario se transforma en búsqueda de evidencia y, posteriormente, en una respuesta que debe indicar límites y sustento documental (Li et al., 2025; Swacha y Gracel, 2025).

Arquitectura del chat portátil y flujo de operación (RAG)

El prototipo, nombrado Biblioteca sin Internet (BiS), se organizó como una arquitectura local cliente-servidor. La interfaz web se abre en un navegador del mismo equipo y envía preguntas al *endpoint* / chat mediante solicitudes POST. El *backend* recibe la pregunta, recupera los fragmentos más parecidos del corpus y genera una respuesta con un LLM local; después, devuelve la salida en formato JSON para mostrarla en la pantalla. En términos prácticos, el flujo puede entenderse en tres momentos: primero se preparan e indexan los PDF; luego, se recuperan fragmentos pertinentes ante cada consulta; finalmente, el modelo redacta la respuesta con base en esa evidencia. La Figura 1 resume este procedimiento.

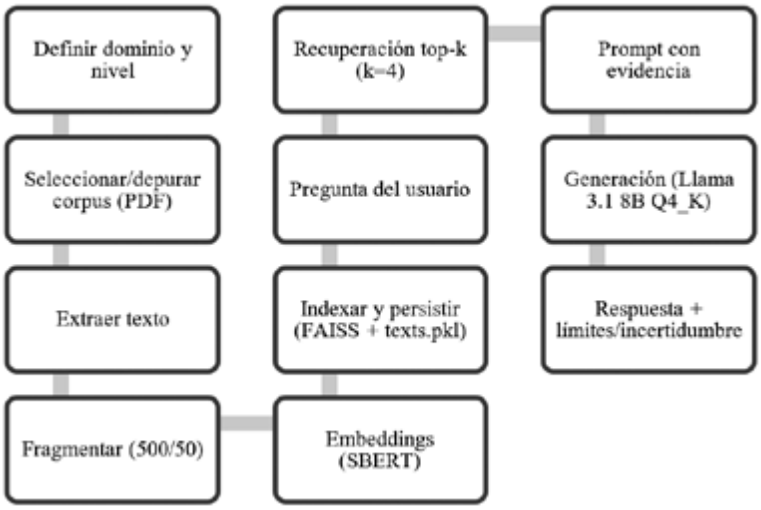
Asimismo, es importante tomar en cuenta lo siguiente:

1. Recuperación: vectorización de la consulta → búsqueda top-k;
2. Generación condicionada: *prompt* + contexto recuperado → respuesta del LLM.

Este flujo es consistente con la formulación canónica de RAG y con revisiones que documentan su adopción en chatbots educativos para reducir alucinaciones y mejorar trazabilidad (Lewis et al., 2020; Li et al., 2025; Swacha y Gracel, 2025).

ISSN 2594-1828 • www.eduscientia.com

Figura 1. Procedimiento general del chat portátil por corpus (RAG)



Fuente: Elaboración propia.

Parámetros técnicos observados y configuración de ejecución

Para que la replicación sea verificable, se registraron los parámetros mínimos que controlan el comportamiento del sistema: tamaño y solapamiento de fragmentos, modelo de embeddings, número de fragmentos recuperados, modelo local y configuración de inferencia. Estos datos no se presentan como tecnicismos aislados, sino como decisiones que afectan la experiencia educativa: precisión de las respuestas, trazabilidad de la evidencia, tiempo de espera y posibilidad de uso en equipos sin internet (Li et al., 2025; Swacha y Gracel, 2025).

La Tabla 1 resume los parámetros observados en el prototipo y la ejecución reportada en una MacBook Pro M1 Pro; también muestra la separación de componentes (corpus, embeddings, recuperación, generación), para facilitar la trazabilidad del pipeline RAG.

ISSN 2594-1828 • www.eduscientia.com

Tabla 1. Parámetros técnicos observados en el prototipo para reproducir el chat portátil

Componente	Parámetro	Valor en ejecución	Interpretación operativa
Corpus	Persistencia	texts.pkl + faiss.index	El sistema reutiliza índice y fragmentos desde disco (sin reabrir los PDF).
Embeddings	Modelo	all-MiniLM-L6-v2	Sentence-BERT para embeddings semánticos (Reimers y Gurevych, 2019).
Embeddings	Dimensión	384	Dimensión típica del modelo MiniLM; impacta memoria/tiempo.
Chunking	chunk_size	500 caracteres	Fragmentación por caracteres (aprox. 90–150 tokens, según idioma).
Chunking	chunk_overlap	50 caracteres	Solapamiento para continuidad local.
Recuperación	k (top-k)	4	Número de fragmentos recuperados por consulta.
Índice	Tipo	FAISS (búsqueda por similitud)	Implementación de índice vectorial para recuperar contexto.
LLM local	Modelo GGUF	Meta-Llama-3.1-8B-Instruct-Q4_K_M.gguf	Modelo cuantizado para inferencia local.
Inferencia	n_ctx	4096	Ventana de contexto usada (aunque el modelo declare mayor contexto de entrenamiento).
Inferencia	n_batch	512	Lote de evaluación reportado por llama.cpp.
Rendimiento	n_threads	8	Paralelismo CPU.
Rendimiento	n_gpu_layers	1	Offload parcial a GPU (Metal) para 1 capa.
Servicio	Endpoint	http://127.0.0.1:8000/chat	Servidor local (uvicorn) que atiende solicitudes POST.

Nota: Los valores se derivan del script rag_pdf.py y de la bitácora de ejecución. La combinación chunking-k-inferencias condiciona la relación precisión/latencia y debe reportarse explícitamente en estudios RAG (Li et al., 2025; Swacha y Gracel, 2025).

Fuente: Elaboración propia.

ISSN 2594-1828 • www.eduscientia.com

Procedimiento replicable

Para facilitar la reproducción se describe el procedimiento como una secuencia de pasos. El orden sigue el pipeline estándar de RAG (Lewis et al., 2020) y prácticas de implementación en entornos educativos (Li et al., 2025; Swacha y Gracel, 2025). La redacción se simplificó para una audiencia educativa no especializada:

Paso 1. Definir tema y público. Se establece el dominio (Historia Universal) y el nivel (secundaria), lo que guía el corpus y las reglas de respuesta (Fuertes Alpiste, 2024; Li et al., 2025).
Paso 2. Reunir textos en carpeta local. Se colocan los PDF del dominio en el directorio del proyecto (Li et al., 2025).
Paso 3. Extraer texto y depurar. El sistema obtiene texto de cada documento y registra fallos sin detener todo el proceso (Li et al., 2025).
Paso 4. Fragmentar el corpus. Los textos se dividen en fragmentos breves y solapados para conservar continuidad local (Li et al., 2025; Swacha y Gracel, 2025).
Paso 5. Calcular embeddings. Cada fragmento se convierte en un vector semántico con SBERT/all-MiniLM-L6-v2 (Reimers y Gurevych, 2019).
Paso 6. Indexar y persistir. Los vectores se guardan en un índice FAISS y los textos/metadatos se almacenan para reutilizarse (Lewis et al., 2020; Li et al., 2025).
Paso 7. Recuperar contexto ante cada consulta. La pregunta del usuario se compara con el índice y se recuperan los fragmentos más pertinentes (Lewis et al., 2020; Swacha y Gracel, 2025).
Paso 8. Generar respuesta condicionada. Los fragmentos recuperados se incorporan al *prompt* y el LLM local redacta la respuesta (Lewis et al., 2020; Li et al., 2025).
Paso 9. Especializar el tema. Se sustituyen los PDF, se reindexa el corpus y se ajustan las instrucciones, manteniendo la misma arquitectura (Swacha y Gracel, 2025; Luku, 2024).

Curaduría del corpus: criterios de selección y sustitución

La curaduría del corpus se concibió como decisión curricular y técnica. Curricular, porque el corpus delimita narrativas y explicaciones disponibles; técnica, porque la calidad de los PDF condiciona extracción y recuperación. Por ende, se seleccionaron textos según estas secciones: (a) fundamentos del oficio histórico, (b) síntesis globales accesibles y (c) didáctica de la Historia con énfasis en el desarrollo del pensamiento histórico, para orientar respuestas hacia evidencia, causalidad, contextualización y corroboración (Wineburg, 2001;

Seixas y Morton, 2012; Lévesque, 2008). Además, en la selección de materiales se tomaron en cuenta las representaciones y sesgos curriculares presentes en textos, dado que el corpus no es neutral (Luku, 2024).

Partiendo de estas consideraciones, se incorporó un subcorpus de materiales descargables en PDF y de acceso abierto (por licencias abiertas tipo *Creative Commons*, repositorios institucionales de distribución gratuita o colecciones patrimoniales). La selección preliminar de estos recursos se apoyó en asistentes de IAGen para identificar repositorios legales y de uso educativo, y fue posteriormente revisada por dos especialistas en Historia para verificar: (a) pertinencia para educación básica secundaria, (b) cobertura mínima por periodos y procesos, y (c) equilibrio entre síntesis, currículo nacional y fuentes para actividades. Este procedimiento mixto —sugerencia asistida por modelos y validación experta— permite sostener un enfoque de corpus situado: el chat no se alimenta de internet abierto, sino de un conjunto deliberado de textos que acota el universo explicativo y hace explícita la mediación curricular.

Prompt pedagógico y control de calidad de respuestas

El comportamiento del chat se controla principalmente mediante: (a) selección y composición del corpus, (b) reglas de recuperación (k, chunking) y (c) un *prompt* pedagógico que delimita lenguaje, alcance y manejo de incertidumbre. Para transparencia, el *prompt* del sistema se documenta íntegramente en el Anexo 1 (y se versiona junto al corpus), de modo que terceros puedan reproducir tanto el estilo de respuesta como la política de "no inventar" cuando el corpus no aporte evidencia suficiente. Esta orientación se alinea con el encuadre de la IAGen como herramienta cognitiva, donde el estudiante conserva funciones ejecutivas y el sistema opera como mediación, no como sustitución (Fuertes Alpiste, 2024). En RAG educativo, estas reglas también contribuyen a fidelidad y trazabilidad de la respuesta (Li et al., 2025; Swacha y Gracel, 2025).

ISSN 2594-1828 • www.eduscientia.com

Tabla 2. Corpus abierto/descargable en PDF para Historia (secundaria) y chat RAG portátil

Núm.	Fuente / recurso	Tipo	Idioma	Cobertura	Uso sugerido en el chat
1	Comisión Nacional de Libros de Texto Gratuitos [CONALITEG], (s.f.),SEP–Libros de texto gratuitos (Secundaria)	Libro de texto	ES	Currículo México	Base curricular y vocabulario escolar
2	Secretaría de Educación Pública, Dirección General de Bachillerato (2024)–*Historia Universal Contemporánea*	Guía/libro	ES	Siglos XVIII–XXI	Puente a media superior; repaso contemporáneo
3	Instituto Nacional Electoral (s. f.) (PDF alojado)–Nueva historia mínima de México ilustrada	Síntesis	ES	México (contexto)	Conectar lo global con caso nacional
4	OpenStax (s.f.$_a$)/ Open Textbook Library. (s. f.$_c$)–*World History, Vol. 1: to 1500*	Libro abierto	EN	Antigüedad–1500	Unidades didácticas “volumen 1”
5	OpenStax (s.f.$_b$)/ Open Textbook Library. (s. f.$_d$)–*World History, Vol. 2: from 1400*	Libro abierto	EN	1400–moderno	Modernidad y contemporaneidad
6	University of North Georgia Press. (s. f.)–*World History: Cultures, States, and Societies to 1500*	Libro abierto	EN	Hasta 1500	Alternativa didáctica para tareas
7	Open Textbook Library (s.f.$_b$)– *World History Since 1500*	Libro abierto	EN	1500–contemporáneo	Texto conciso para guías y fichas
8	Open Textbook Library (s.f.$_a$)– *Modern World History*	Libro abierto	EN	Siglos XVIII–XX	Refuerzo de temas modernos
9	Biblioteca Virtual Miguel de Cervantes (s.f.)–Historia Universal	Biblioteca digital	ES	Variable	Lecturas abiertas y compendios
10	Library of Congress (s.f.)/ World Digital Library	Colección patrimonial	Multi	Fuentes primarias	Proyectos con evidencias (mapas, facsímiles)

Nota: Para reproducibilidad y control, cada PDF incorporado al corpus debe registrarse con: (a) fuente/repositorio, (b) fecha de descarga, (c) licencia/condición de uso indicada, (d) versión/edición si aplica, y (e) huella hash (p. ej., SHA-256), para garantizar que el corpus sea verificable entre corridas.

Fuente: Elaboración propia.

ISSN 2594-1828 • www.eduscientia.com

Criterios de análisis técnico-pedagógico

Para articular la dimensión técnica con la educativa, los hallazgos se interpretaron mediante cuatro criterios: (a) reproducibilidad, entendida como posibilidad de reconstruir el sistema con los mismos parámetros; (b) trazabilidad, entendida como capacidad de vincular respuestas con fuentes del corpus; (c) pertinencia didáctica, entendida como potencial para formular preguntas, contextualizar y corroborar; y (d) viabilidad situada, entendida como funcionamiento local en contextos con conectividad limitada. Estos criterios permitieron traducir métricas técnicas —por ejemplo, número de fragmentos, latencia y longitud de *prompt*— en implicaciones para el diseño de actividades de Historia.

Debido a que no se aplicaron instrumentos con estudiantes, los resultados no se interpretan como evidencia de eficacia educativa. En cambio, se presentan como condiciones de posibilidad para una fase empírica posterior, en la que podrían valorarse comprensión histórica, calidad de las preguntas, uso de evidencia y percepción docente.

Consideraciones éticas de la fase actual

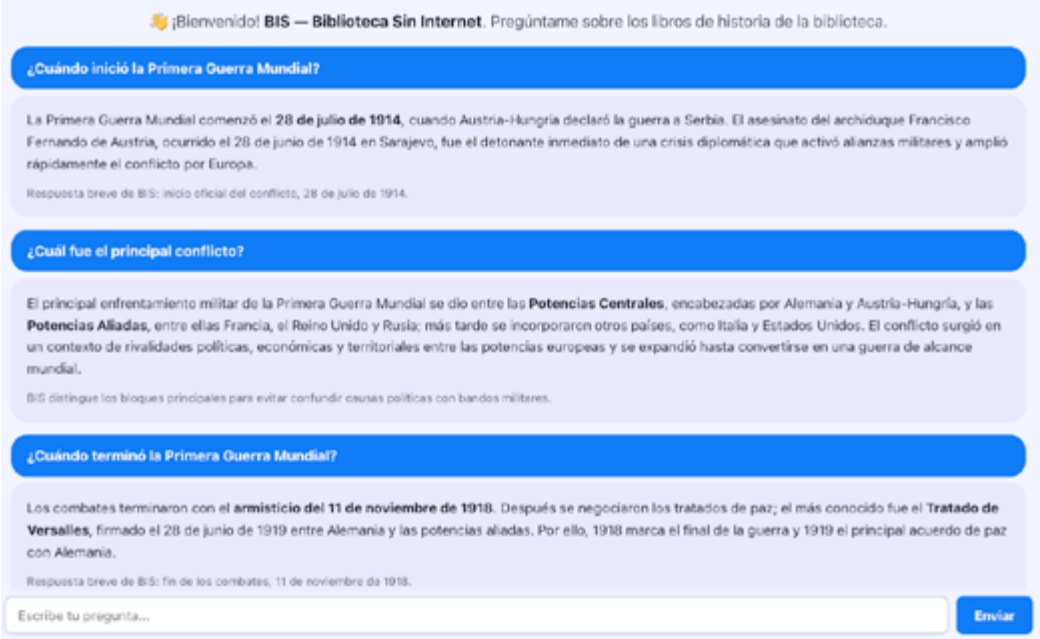
Esta fase no involucró participantes humanos ni análisis de datos personales; se limitó a documentación técnica y curaduría de textos. Sin embargo, se recomienda sostener principios de transparencia (corpus explícito, límites de respuesta) y gestión de riesgo en fases posteriores con población escolar (Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura [UNESCO], 2023; National Institute of Standards and Technology [NIST], 2024).

Resultados

Los resultados se organizan en dos niveles. Primero, se reportan las características técnicas que permiten reproducir el prototipo; segundo, se explican sus implicaciones pedagógicas para el uso en Historia. Por tratarse de una fase preempírica, los datos no representan aprendizaje de estudiantes, sino comportamiento del sistema y decisiones de diseño. En la Figura 2 se presenta la interfaz del chat desarrollado. Aunque se trata de una implementación básica, ha sido suficiente para este primer avance del proyecto al permitir la consulta y visualización de respuestas en un entorno local. En etapas posteriores se prevé mejorar el diseño de la interfaz y añadir funciones orientadas al uso educativo, como el guardado y recuperación de conversaciones, con el fin de apoyar el seguimiento de actividades y la trazabilidad del proceso de consulta.

ISSN 2594-1828 • www.eduscientia.com

Figura 2. Toma de captura de pantalla del chat BiS



Características del corpus indexado

Según lo realizado en el proceso de indexación y persistencia del corpus, la Tabla 3 sintetiza el volumen y la estructura del material que alimenta la recuperación del sistema: se registraron 141 263 fragmentos almacenados en texts.pkl, con una longitud media de 432.1 caracteres y mediana de 461, además de un rango típico P25–P75 de 436–482 caracteres, lo que confirma una fragmentación relativamente estable y cercana al objetivo del chunking (Tabla 3). En conjunto, el corpus acumulado representa ≈61.04 millones de caracteres, lo que sugiere una base documental amplia para consultas diversas; a su vez, la dimensión del embedding (384) y la memoria teórica de embeddings (≈206.93 MiB, float32) permiten dimensionar el costo mínimo de almacenamiento vectorial, sin considerar la sobrecarga propia del índice ni los metadatos (Tabla 3).

Tabla 3. Estadísticas del corpus indexado (derivadas de texts.pkl)

Indicador	Valor
Fragmentos indexados	141,263
Longitud media de fragmento	432.1 caracteres
Mediana de fragmento	461 caracteres
Rango típico (P25–P75)	436–482 caracteres
Total aproximado de texto	61.04 millones de caracteres
Dimensión de embedding	384
Memoria teórica de embeddings (float32)	206.93 MiB

Nota: texts.pkl = archivo de persistencia que almacena fragmentos de texto y metadatos asociados al corpus; embedding = vector semántico que representa un fragmento de texto; MiB = mebibyte (2^20 bytes = 1 048 576 bytes); P25–P75 = percentiles 25 y 75 (rango intercuartílico).

Fuente: Elaboración propia.

Entorno de ejecución y configuración observada

Como hallazgo técnico, en la Tabla 4 se documenta que el prototipo alcanzó inferencia local en Apple Silicon bajo un esquema de aceleración heterogénea (CPU + Metal), pero con *offload* limitado a GPU: solo 1 de 33 capas se delegó a Metal (n_gpu_layers=1), mientras que la mayor parte del grafo de cómputo permaneció en CPU (Tabla 4). Esta configuración constituye un determinante operativo del rendimiento observado, ya que el *throughput* y la latencia quedan condicionados por el componente de cómputo dominante y por el costo de evaluación del contexto.

Asimismo, aunque el modelo declara un contexto de entrenamiento alto (n_ctx_train= 131072), la ejecución se realizó con una ventana efectiva restringida (n_ctx=4096) y un n_batch=512, lo que delimita la cantidad de evidencia recuperada que puede incorporarse al *prompt* sin truncamiento y, por tanto, fija un techo práctico para la extensión del contexto RAG en esta corrida (Tabla 4). También, hubo un KV caché total de 512 MiB (principalmente en CPU) y el despliegue del servicio en http://127.0.0.1:8000, elementos que estabilizan la reproducibilidad del entorno y permiten interpretar los resultados de desempeño como propios de una arquitectura RAG ejecutada en un escenario local, con restricciones explícitas de memoria y contexto (Tabla 4).

ISSN 2594-1828 • www.eduscientia.com

Tabla 4. Configuración y estado de carga del modelo (bitácora)

Indicador	Valor
Equipo	MacBook Pro (Apple M1 Pro)
Aceleración	Metal
Modelo	Meta-Llama-3.1-8B-Instruct
Cuantización / formato	Q4_K, GGUF V3
Tamaño del archivo	4.58 GiB
Capas offload	1/33 capas a GPU (n_gpu_layers=1)
Contexto usado	n_ctx=4096 (modelo declara n_ctx_train=131072)
Batch	n_batch=512
KV cache	512 MiB (≈496 MiB CPU + 16 MiB Metal)
Servidor	uvicorn en http://127.0.0.1:8000

Nota: Metal = API de Apple para cómputo/aceleración usada por llama.cpp en Apple Silicon; GGUF = formato de archivo para modelos cuantizados compatible con llama.cpp; Q4_K = esquema de cuantización de 4 bits (familia K); GiB = gibibyte (2^30 bytes); MiB = mebibyte (2^20 bytes); n_gpu_layers = número de capas del modelo descargadas a GPU; n_ctx = longitud de la ventana de contexto utilizada en inferencia; n_ctx_train = longitud de contexto reportada en metadatos del modelo (entrenamiento); n_batch = tamaño de lote para evaluación; KV cache = memoria de claves/valores usada por el mecanismo de atención para acelerar la generación; uvicorn = servidor ASGI para exponer el endpoint local del chat.

Fuente: Elaboración propia.

Rendimiento por solicitud /chat (latencia y *throughput*)

Esta sección reporta el desempeño del prototipo por petición utilizando los registros automáticos de llama_perf_context_print. Para cada solicitud se documentan: (a) tokens evaluados en el *prompt*, (b) tokens generados, (c) tiempos de evaluación y (d) *throughput* aproximado (tokens/segundo). Adicionalmente, en solicitudes posteriores, se observó el indicador *prefix-match hit* en la bitácora, consistente con reutilización parcial de un prefijo del *prompt* y, por tanto, con reducción del costo de evaluación del *prompt* en interacciones subsecuentes. En términos de tiempo de respuesta, el sistema registró latencias totales por interacción en el orden de decenas de segundos, con un rango aproximado entre ~22 s y ~43 s, bajo la configuración local descrita.

ISSN 2594-1828 • www.eduscientia.com

En la Tabla 5 se presentan las métricas por solicitud atendida por el *endpoint* /chat en ejecución local. Los resultados muestran que la latencia total por petición osciló entre 21.937 s y 42.683 s. En términos del perfil temporal, el componente dominante corresponde a la generación (tiempo de evaluación de tokens generados), mientras que el tiempo de evaluación del *prompt* aumenta cuando el *prompt* es más extenso, lo cual es consistente con la incorporación de mayor contexto recuperado al *prompt* en un esquema RAG (Li et al., 2025; Swacha y Gracel, 2025). Asimismo, el *throughput* de generación se mantuvo en el rango 13.39–21.23 tokens/s, lo que caracteriza el comportamiento del modelo cuantizado en inferencia local bajo las condiciones de ejecución reportadas (Tabla 5).

Para sintetizar el comportamiento observado, la Tabla 6 resume estadísticas descriptivas de seis solicitudes que se observaron en este ejercicio. En conjunto, el desempeño promedio se ubicó en 33.46 s por interacción, con una mediana de 34.57 s. El *throughput* medio de evaluación del *prompt* fue 68.29 ± 11.49 tokens/s, mientras que el *throughput* medio de generación fue 16.17 ± 1.87 tokens/s (Tabla 6). Estos resultados son coherentes con el patrón esperado en RAG: la latencia depende tanto del volumen de contexto incorporado al *prompt* como del costo de generación del modelo, por lo que variaciones en longitud de *prompt* y longitud de salida explican parte importante de la variabilidad entre solicitudes.

ISSN 2594-1828 • www.eduscientia.com

Tabla 5. Métricas de desempeño por solicitud /chat (MacBook Pro M1 Pro, inferencia local)

req	*prompt*_tokens	*prompt*_s	*prompt*_toks/s	gen_tokens	gen_s	gen_toks/s	total_s	total_tokens
1	168	3.609	46.55	511	38.155	13.39	42.000	679
2	333	4.497	74.06	369	20.363	18.12	25.003	702
3	316	4.058	77.87	511	27.781	18.39	32.096	827
4	589	8.424	69.92	213	13.442	15.85	21.937	802
5	365	4.828	75.59	511	31.964	15.99	37.042	876
6	589	8.954	65.78	511	33.485	15.26	42.683	1100

Nota: req = número de solicitud; *prompt*_tokens = tokens evaluados en el *prompt*; *prompt*_s = tiempo de evaluación del *prompt* (s); *prompt*_toks/s = *throughput* de evaluación del *prompt* (tokens/s); gen_tokens = tokens generados; gen_s = tiempo de generación (s); gen_toks/s = *throughput* de generación (tokens/s); total_s = tiempo total por solicitud (s); total_tokens = tokens totales (*prompt* + generación). s = segundos.

Fuente: Elaboración propia.

Tabla 6. Resumen estadístico de desempeño (derivado de las seis solicitudes)

Indicador	Valor
Rango de latencia total	21.94–42.68 s
Mediana de latencia total	34.57 s
Promedio de latencia total	33.46 s
Throughput prompt (media ± DE)	68.29 ± 11.49 tokens/s
Throughput generación (media ± DE)	16.17 ± 1.87 tokens/s
Carga inicial del modelo (reportada)	3.61 s

Nota: DE = desviación estándar; s = segundos; tokens/s = tokens por segundo. Los valores se derivan de los reportes automáticos de llama_perf_context_print para las seis solicitudes consecutivas.

Fuente: Elaboración propia.

ISSN 2594-1828 • www.eduscientia.com

Implicaciones pedagógicas de los resultados técnicos

Los resultados técnicos adquieren sentido educativo cuando se traducen en decisiones de mediación. La portabilidad y el anclaje documental favorecen continuidad de consulta y verificación, mientras que la latencia y el límite de contexto obligan a diseñar preguntas breves, tareas por etapas y actividades de revisión de evidencias. La Tabla 7 sintetiza esta lectura pedagógica.

Tabla 7. Lectura pedagógica de los principales resultados del prototipo

Resultado técnico observado	Lectura educativa	Decisión didáctica recomendada
Corpus local e indexado	Delimita las fuentes disponibles y permite discutir qué voces, periodos y perspectivas entran al chat.	Registrar versiones del corpus, revisar sesgos y proponer actividades de comparación entre fuentes.
Respuesta condicionada por fragmentos RAG	La consulta puede orientarse hacia evidencia y no solo hacia respuesta inmediata.	Pedir al estudiante que identifique la evidencia, contraste fragmentos y explique causas/consecuencias.
Latencias de 22 a 43 segundos	La interacción no debe plantearse como diálogo instantáneo continuo, sino como consulta reflexiva.	Usar preguntas breves, secuencias por etapas y tiempos de lectura mientras se genera la respuesta.
Prompt con límites e incertidumbre	Favorece una cultura de verificación al declarar cuándo el corpus no alcanza para responder.	Incluir preguntas de verificación, reformulación de dudas y criterios para distinguir hecho e interpretación.

Nota: La tabla no reporta impacto en aprendizaje; traduce los resultados de funcionamiento del prototipo en criterios de diseño didáctico para una fase empírica posterior.
Fuente: Elaboración propia.

ISSN 2594-1828 • www.eduscientia.com

Discusión y conclusiones

En relación con la pregunta de investigación, los resultados muestran que el diseño de un chat educativo portátil basado en RAG es viable cuando se documentan cuatro decisiones: corpus, recuperación, *prompt* y entorno de ejecución. Esto responde al objetivo del artículo porque no solo se informa que el prototipo funciona, sino cómo lo hace, con qué límites y bajo qué condiciones pedagógicas debería emplearse.

El aporte se ubica en un punto previo a los estudios de impacto. A diferencia de investigaciones que analizan narrativas históricas generadas por estudiantes o por ChatGPT (Kindenberg, 2024), estrategias didácticas con IA en ciencias sociales (Ibarra Álvarez et al., 2026) o propuestas metodológicas con IAGen y modelos 3D (Cabezas Guzmán y Rovira Marcelino, 2025), este trabajo documenta la infraestructura local y el control del corpus como condición para un uso escolar verificable. En diálogo con las revisiones sobre RAG educativo (Li et al., 2025; Swacha y Gracel, 2025), el prototipo aporta un caso situado para Historia y muestra que la exactitud no depende solo del modelo, sino de la curaduría, el *prompt*, la transparencia y la actividad didáctica.

El desarrollo de chats portátiles basados en RAG y ejecutados localmente (sin internet) permite replantear la integración de la IAGen en contextos donde la conectividad es intermitente o inexistente. En esos escenarios, la portabilidad no es un atributo secundario, sino una condición de posibilidad: el sistema funciona como una infraestructura mínima que mantiene disponible una biblioteca conversacional, aun cuando el entorno no sostenga servicios en la nube. Esta decisión técnica (modelo local más el corpus local) también implica una decisión pedagógica: el aprendizaje puede organizarse alrededor de tareas y recursos anclados en un contexto específico, en lugar de depender de una web abierta y desigual en acceso, calidad y verificabilidad (Li et al., 2025; Swacha y Gracel, 2025; UNESCO, 2023).

En términos técnicos, los resultados muestran que el prototipo es operable como sistema local con corpus persistido e inferencia en Apple Silicon, y que su comportamiento está condicionado por decisiones típicas de RAG: segmentación del corpus (*chunking*), volumen de contexto incorporado al *prompt* y costo de generación del modelo. El corpus quedó indexado en una cantidad alta de fragmentos con longitudes consistentes con la estrategia de segmentación, lo que favorece estabilidad en la recuperación y reduce la probabilidad de inyectar fragmentos excesivos al contexto. Aun así, el desempeño por interacción indica que la latencia se concentra principalmente en la generación y que aumenta cuando crece el *prompt*, lo que refuerza la necesidad de reglas de respuesta breves, estructura fija y manejo

ISSN 2594-1828 • www.eduscientia.com

explícito de incertidumbre cuando el corpus no aporta evidencia suficiente. Como área de oportunidad, se requiere fortalecer la trazabilidad respuesta–evidencia; además de declarar límites del corpus, conviene que el sistema muestre de forma sistemática los fragmentos (o referencias) que sustentan cada respuesta y, cuando sea posible, su procedencia (p. ej., documento y sección), de modo que el usuario pueda verificar y aprender a corroborar, no solo a consultar. En conjunto, estos hallazgos apoyan la factibilidad del enfoque portátil y señalan un margen claro para optimizar la experiencia de uso mediante ajustes de *prompt*, control de longitud de salida, curaduría del corpus y mejoras de transparencia en la evidencia recuperada, sin modificar la arquitectura de base (Li et al., 2025; Swacha y Gracel, 2025).

Desde la perspectiva del aprendizaje situado, estas aplicaciones pueden entenderse como herramientas situadas cuando el conocimiento que median se define por prácticas, problemas y lenguajes propios de una comunidad (Lave y Wenger, 1991). En la tradición clásica, aprender no equivale a recibir contenidos, sino a participar en prácticas sociales y construir significado en contexto; la noción de *legitimate peripheral participation* subraya que la competencia se desarrolla gradualmente mediante participación guiada en comunidades de práctica (Lave y Wenger, 1991; Wenger, 1998). En esa línea, un chat portátil por corpus puede operar como un mediador que apoya la participación (formular preguntas, comparar versiones, justificar interpretaciones) siempre que las actividades se diseñen para que la agencia cognitiva permanezca del lado del estudiante y del grupo (Brown et al., 1989; Fuertes Alpiste, 2024). Dicho de otro modo: el valor educativo no está en tener respuestas, sino en usar el dispositivo para sostener prácticas de indagación histórica (corroborar, contextualizar, explicar causalidad), en consonancia con enfoques del desarrollo de pensamiento histórico (Wineburg, 2001; Seixas y Morton, 2012).

Desde la perspectiva de las herramientas para la cognición (HPC), el ChatBIS puede entenderse como una implementación concreta del enfoque de aprender con la tecnología y no de aprender de la tecnología. A diferencia de los modelos donde la tecnología asume el rol de tutor que transmite contenidos, aquí el sistema opera como artefacto mediador que organiza la actividad intelectual del estudiante sin sustituirla (Derry y Lajoie, 1993; Jonassen, 1996). La arquitectura RAG, al exigir recuperación de evidencia y delimitación del corpus, estructura procesos como búsqueda, organización e integración de información, coherentes con las operaciones cognitivas descritas por Iiyoshi et al. (2005). En este sentido, el valor pedagógico del prototipo no reside en automatizar respuestas, sino en sostener funciones ejecutivas del aprendiz —formular preguntas, contrastar fuentes, elaborar explicaciones causales— en consonancia con la propuesta de entender la IAGen como *mindtool*

ISSN 2594-1828 • www.eduscientia.com

o HPC (Fuertes Alpiste, 2024). Así, el ChatBIS se inscribe en una tradición constructivista y socioconstructivista donde la tecnología amplifica la capacidad de representación y argumentación histórica, manteniendo la agencia cognitiva del estudiante como núcleo del proceso formativo (Papert, 1982; Lave y Wenger, 1991; Wenger, 1998).

Ahora bien, el rasgo más delicado —y a la vez más interesante— de estas arquitecturas portátiles es que habilitan un control explícito del conocimiento al que acceden los estudiantes, ya que el corpus define el mundo consultable. Esa delimitación puede funcionar como mecanismo de calidad, pertinencia curricular y seguridad epistemológica, (por ejemplo, al reducir alucinaciones y respuestas sin fuente) (Lewis et al., 2020; Li et al., 2025). Sin embargo, la misma característica puede convertirse en un dispositivo de cierre epistemológico, en el cual quien selecciona el corpus decide qué narrativas, voces y categorías entran al aula. En Historia, donde el conocimiento es inseparable de interpretación y perspectiva, el corpus no es neutral; como ocurre con los libros de texto, puede transmitir significados culturales e ideológicos de forma implícita (Luku, 2024; Apple, 1993). Por eso, un chat situado no debería confundirse con un chat restringido: lo situado exige pertinencia contextual, pluralidad y mecanismos para visibilizar sesgos y ausencias (Wineburg, 2001; UNESCO, 2023).

En términos de diseño educativo, lo anterior conduce a una implicación central: un chat portátil por corpus no es solo una herramienta tecnológica, sino una forma de gobernanza curricular en miniatura. La calidad del aprendizaje dependerá de reglas de construcción del corpus (criterios, versiones, trazabilidad), de reglas de interacción (declaración de límites, incertidumbre, solicitud de precisión) y de prácticas docentes que conviertan la consulta en actividad reflexiva, no en consumo de respuesta (Fuertes Alpiste, 2024; Li et al., 2025). En coherencia con marcos de gestión de riesgo, esta gobernanza debe incluir transparencia (qué fuentes contiene, cuándo se actualizó), auditoría (qué cambió entre versiones) y evaluación contextual (si el recurso mejora comprensión y no solo velocidad) (NIST, 2024; UNESCO, 2023). Además, se identifican estas implicaciones prácticas para comunidades con conectividad limitada:

- Equidad por infraestructura mínima: Lo portátil reduce la dependencia de conectividad y permite continuidad didáctica en escuelas con acceso limitado.
- Control de calidad por anclaje documental: RAG puede reducir respuestas infundadas al exigir evidencia desde el corpus (Lewis et al., 2020; Li et al., 2025).

ISSN 2594-1828 • www.eduscientia.com

- Currículo situado (no genérico): El corpus puede configurarse para un territorio, una región o una problemática concreta, habilitando microcurrículos de alta pertinencia (Lave y Wenger, 1991; Wenger, 1998).
- Riesgo de monocultura narrativa: Si el corpus es único, el chat puede estabilizar una sola versión del pasado; por ello se recomienda pluralidad de fuentes y explicitar qué no está incluido (Apple, 1993; Luku, 2024).
- Diseño como participación: El corpus debe construirse con participación docente y, cuando sea posible, comunitaria, para que lo situado no sea imposición, sino apropiación local (Wenger, 1998).

Ante lo planteado, los chats portátiles por corpus materializan una tensión productiva porque pueden ser tecnologías de acceso en territorios sin internet y, simultáneamente, tecnologías de regulación del conocimiento disponible. Su legitimidad pedagógica no proviene de reemplazar al docente, sino de operar como biblioteca conversacional que haga practicable el aprendizaje situado: preguntas mejores, evidencia rastreable, interpretaciones justificadas y límites explícitos (Brown et al., 1989; Lave y Wenger, 1991; Li et al., 2025). La condición para que ese control sea educativo —y no meramente restrictivo— es que el corpus sea gobernado con criterios de pluralidad, transparencia y revisión, de modo que el dispositivo no cierre el mundo, sino que lo vuelva discutible.

Como limitaciones, este trabajo se sitúa en una fase preempírica y de caracterización técnico-pedagógica. Primero, los resultados no incluyen todavía evaluación sistemática con estudiantes o docentes; por ello no es posible inferir el impacto en el aprendizaje de la Historia ni en el desarrollo del pensamiento histórico, sino únicamente la factibilidad y el comportamiento del prototipo bajo condiciones controladas. Segundo, el desempeño observado depende de una configuración específica de hardware, parámetros de inferencia y composición del corpus, lo que exige cautela al generalizar métricas de latencia o fluidez a otros equipos y conjuntos documentales. Estas limitaciones no disminuyen el aporte, porque la contribución principal es documentar un procedimiento reproducible y un diseño portátil que permite operar sin internet y con control explícito del corpus. Como trabajo futuro, se propone una fase empírica con docentes y estudiantes de secundaria, instrumentos de observación y entrevistas, así como rúbricas para valorar comprensión histórica, uso de evidencia, calidad de preguntas y percepción de utilidad. En paralelo, se recomienda continuar optimizando la interfaz, la trazabilidad respuesta–evidencia, el guardado de

ISSN 2594-1828 • www.eduscientia.com

conversaciones, el control de longitud de salida y la curaduría plural del corpus, manteniendo la arquitectura RAG como base (Li et al., 2025; Swacha y Gracel, 2025). SC

Referencias

Apple, M. W. (1993). *Official Knowledge: Democratic Education in a Conservative Age*. Routledge.

Barton, K. C., y Levstik, L. S. (2004). *Teaching History for the Common Good*. Lawrence Erlbaum Associates.

Biblioteca Virtual Miguel de Cervantes. (s. f.). Historia Universal [Colección/búsqueda de recursos]. https://www.cervantesvirtual.com

Brown, J. S., Collins, A., y Duguid, P. (1989). Situated Cognition and the Culture of Learning. *Educational Researcher*, *18*(1), 32–42. https://doi.org/10.3102/0013189X018001032

Cabezas Guzmán, G., y Rovira Marcelino, A. (2025). Tecnologías emergentes aplicadas a la enseñanza de la Historia Antigua: Una propuesta metodológica basada en IA generativa y modelos 3D. *AI & Antiquity: Journal of Teaching and Technology in Ancient Studies*, *1*(1), 93–103. https://doi.org/10.64946/aiantiquity.v1i1.005

Comisión Nacional de Libros de Texto Gratuitos. (s. f.). Libros de texto gratuitos de nivel educación secundaria. Ciclo escolar 2024-2025[Catálogo]. https://libros.conaliteg.gob.mx/secundaria.html

Derry, S. J., y Lajoie, S. P. (1993). A middle camp for (un)intelligent instructional computing: An introduction. En S. P. Lajoie y S. J. Derry (Eds.), *Computers as cognitive tools* (pp. 1–11). https://books.google.co.ls/books?id=xGeasuCsN2MC&printsec=frontcover#v=onepage&q&f=false

Echeverría Quiñonez, B. R., y Otero Mendoza, L. K. (2025). Inteligencia Artificial Generativa como herramienta pedagógica: una revisión sistemática sobre su impacto en los procesos de enseñanza-aprendizaje. *SAGA Revista Científica Multidisciplinar*, *2*(3), 537–550.

Fuertes Alpiste, M. (2024). Enmarcando las aplicaciones de IA generativa como herramientas para la cognición en educación. *Pixel-Bit. Revista de Medios y Educación*, 71, 42–57. https://doi.org/10.12795/pixelbit.107697

Ibarra Álvarez, J. V., Benítez Maldonado, F. R., Fernández Cobas, L. C., y Vergel Parejo, E. E. (2026). Estrategia didáctica basada en inteligencia artificial para el aprendizaje de ciencias sociales en quinto grado. *Revista REG*, *5*(1), 711–740. https://doi.org/10.70577/reg.v5i1.482

Iiyoshi, T., Hannafin, M. J., y Wang, F. (2005). Cognitive tools and student-centered learning: Rethinking tools, functions, and applications. *Educational Media International*, *42*(4), 281–296. https://doi.org/10.1080/09523980500161346

Instituto Nacional Electoral. (s. f.). *Nueva historia mínima de México ilustrada*. https://portalanterior.ine.mx/archivos2/portal/servicio-profesional-electoral/concurso-publico/2016-2017/primera-convocatoria/docs/Otros/36-historia-minima-de-mexico.pdf

Jonassen, D. H. (1996). *Computers in the classroom: Mindtools for critical thinking*. Prentice Hall.

Kindenberg, B. (2024). ChatGPT-Generated and Student-Written Historical Narratives: A Comparative Analysis. *Education Sciences, 14*(5), 530. https://doi.org/10.3390/educsci14050530

Lave, J., y Wenger, E. (1991). *Situated learning: Legitimate peripheral participation*. Cambridge University Press. https://doi.org/10.1017/CBO9780511815355

Lévesque, S. (2008). *Thinking Historically: Educating Students for the Twenty-First Century*. University of Toronto Press.

Lewis, P., Pérez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W. T., Rocktäschel, T., Riedel, S., y Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems*, 33, 9459–9474. https://doi.org/10.48550/arXiv.2005.11401

Li, Z., Wang, Z., Wang, W., Hung, K., Xie, H., y Wang, F. L. (2025). Retrieval-augmented generation for educational application: A systematic survey. *Computers and Education: Artificial Intelligence*, 8, 100417. https://doi.org/10.1016/j.caeai.2025.100417

Library of Congress. (s. f.). World Digital Library [Colección digital]. https://www.loc.gov

Luku, E. (2024). Free from gender bias? An Analysis of Gender Representations in Civics and Mathematics Textbooks for Primary Schools in Albania. *Journal of Educational Media, Memory, and Society*, *16*(2), 113–142. https://doi.org/10.3167/jemms.2024.160205

National Institute of Standards and Technology. (2024). *Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile (NIST AI 600-1)*. U.S. Department of Commerce. https://doi.org/10.6028/NIST.AI.600-1

Open Textbook Library. (s. f.-a). *Modern World History* [Open textbook, PDF]. https://open.umn.edu/opentextbooks

Open Textbook Library. (s. f.-b). *World History Since 1500: An open and free textbook* [Open textbook, PDF]. https://open.umn.edu/opentextbooks

Open Textbook Library. (s. f.-c). *World History*, volume 1: to 1500 [Registro del libro, PDF]. https://open.umn.edu/opentextbooks

Open Textbook Library. (s. f.-d). *World History*, volume 2: From 1400 [Registro del libro, PDF]. https://open.umn.edu/opentextbooks

OpenStax. (s. f.a). *World History*, volume 1: to 1500 [Open textbook, PDF]. Rice University. https://openstax.org/details/books/world-history-volume-1

OpenStax. (s. f.b). *World History*, volume 2: from 1400 [Open textbook, PDF]. Rice University. https://openstax.org/details/books/world-history-volume-2

Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura. (2023). Guidance for generative AI in education and research. https://unesdoc.unesco.org/ark:/48223/pf0000386693

Papert, S. (1982). *Mindstorms: Children, computers, and powerful ideas*. Basic Books.

Reimers, N., y Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. *In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing* (EMNLP-IJCNLP) (pp. 3982–3992). Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1410

Salomon, G., Perkins, D. N., y Globerson, T. (1991). Partners in Cognition: Extending Human Intelligence with Intelligent Technologies. *Educational Researcher*, *20*(3), 2–9. https://doi.org/10.3102/0013189X020003002

Secretaría de Educación Pública, Dirección General de Bachillerato. (2024). *Historia universal contemporánea* [Libro de texto gratuito, PDF]. https://dgb.sep.gob.mx/storage/recursos/2024/09/xviD7hv8rg-Historia-Universal-Contemporanea.pdf

Seixas, P., y Morton, T. (2012). *The Big Six Historical Thinking Concepts*. Nelson Education.

Swacha, J., y Gracel, M. (2025). Retrieval-Augmented Generation (RAG) Chatbots for Education: A Survey of Applications. *Applied Sciences, 15*(8), 4234. https://doi.org/10.3390/app15084234

University of North Georgia Press. (s. f.). *World History: Cultures, States, and Societies to 1500* [Open textbook, PDF]. https://ung.edu/university-press

Wenger, E. (1998). *Communities of Practice: Learning, Meaning, and Identity*. Cambridge University Press. https://doi.org/10.1017/CBO9780511803932

Wineburg, S. (2001). *Historical thinking and other Unnatural acts: Charting the Future of Teaching the Past*. Temple University Press.

World Digital Library. (s. f.). World Digital Library [Colección digital]. https://www.wdl.org