INNOVACADEMIA
Revista Científica en Educación de la Universidad Autónoma de Nuevo León

DOI: https://doi.org/10.29105/innoacad.v1i2.36
Recibido - Received: 27 de febrero de 2025
Aceptado - Accepted: 8 de mayo de 2025

Ensayo

Modelos de lenguaje en educación: Inteligencia Artificial Generativa para optimizar el análisis del desempeño docente

Language Models in Education: Generative Artificial Intelligence for Optimizing Teacher Performance Analysis

Roberto E. Ramos-Rivera[1], Pedro C. Santana-Mancilla*, Jesús García-Mancilla[2], Laura S. Gaytán-Lugo[3]

[1]Facultad de Telemática, Universidad de Colima, México. Ingeniería Civil por la Universidad de Colima.
ORCID: 0009-0007-9588-3115.
Contacto: roberto_ramos@ucol.mx

*Autor para correspondencia. Facultad de Telemática, Universidad de Colima, México. Doctorado en Tecnologías de la Información y Comunicaciones por la Universidad de Vigo. ORCID: 0000-0002-4184-0116.
Contacto: psantana@ucol.mx

[2] AI & Digital Solutions, Argomai, Estados Unidos de América. Maestría en Ciencias en Computación por el Instituto Tecnológico Autónomo de México. ORCID: 0000-0002-2104-8033.
Contacto: jesusmancilla@argomai.com

[3] Facultad de Ingeniería Mecánica y Eléctrica, Universidad de Colima, México. Doctorado en Tecnologías de Información por la Universidad de Guadalajara. ORCID: 0000-0002-7007-7500.
Contacto: laura@ucol.mx

RESUMEN

Este artículo explora el uso de la Inteligencia Artificial Generativa, específicamente los Grandes Modelos de Lenguaje (LLM, por sus siglas en inglés), para analizar respuestas abiertas en evaluaciones del desempeño docente, justificando su elección frente a métodos tradicionales y evaluando sus fortalezas, limitaciones y viabilidad práctica. Aunque los LLM ofrecen capacidades avanzadas para interpretar y clasificar datos textuales, su tendencia a generar 'alucinaciones' plantea desafíos en contextos donde la precisión es crucial. Para mitigar estos riesgos, se presentan tres enfoques: los LLM de dominio específico, entrenados con datos educativos para mejorar su relevancia; los Pequeños Modelos de Lenguaje (SLM por sus siglas en inglés), modelos más ligeros que optimizan la eficiencia y reducen la posibilidad de errores; y el uso de modelos en la nube con entrenamiento *few-shot*, que permiten ajustes rápidos mediante ejemplos representativos, aunque con implicaciones en privacidad y protección de datos. Finalmente, se proponen métricas clave como indicadores para medir el impacto y la calidad de estas soluciones en contextos educativos y se describen los beneficios de estas herramientas para las instituciones educativas, incluyendo la mejora en la toma de decisiones, la accesibilidad tecnológica y adaptación a entornos con recursos limitados.

Palabras clave:

inteligencia artificial generativa, grandes modelos de lenguaje (LLM), evaluación del desempeño docente.

ABSTRACT

This work explores using Generative Artificial Intelligence, specifically Large Language Models (LLM), to analyze open-ended responses in teacher performance assessments, justifying their choice over traditional methods and assessing their strengths, limitations, and practical feasibility. Although LLM offers advanced capabilities for interpreting and classifying textual data, its tendency to generate 'hallucinations' presents challenges in contexts where precision is crucial. Three approaches are presented to mitigate these risks: domain-specific LLMs, fine-tuned with educational data to enhance their relevance; Small Language Models (SLM), lighter models designed to optimize efficiency and reduce errors; and cloud-based models using few-shot learning, which allow rapid adaptation with representative examples but pose privacy concerns when processing sensitive educational data. Finally, key metrics are proposed as indicators to measure the impact and quality of these solutions in educational contexts, and the benefits of these tools for academic institutions are discussed, including improved decision-making, technological accessibility, and ecological sustainability.

Keywords:

generative artificial intelligence, large language models (LLM), teacher performance assessment.

Cómo referenciar:
Ramos-Rivera, R. E., Santana-Mancilla, P. C., García-Mancilla, J. y Gaytán-Lugo, L. S. (2025). Modelos de lenguaje en educación: Inteligencia Artificial Generativa para optimizar el análisis del desempeño docente. *INNOVACADEMIA, 1*(2), 70-81.
https://doi.org/10.29105/innoacad.v1i2.36

Introducción

En los últimos años, gracias a los avances tecnológicos en el aprendizaje profundo, específicamente en el desarrollo de Grandes Modelos de Lenguaje (LLM, por sus siglas en inglés, Large Language Models), la implementación de sistemas conversacionales ha ganado una importancia significativa en diversas áreas como la salud, los negocios y la educación. En particular, en la gestión educativa, la evaluación periódica del desempeño docente es fundamental para garantizar la mejora continua de la calidad en las instituciones. Aunque estos instrumentos incluyen tanto preguntas cerradas como abiertas, el análisis cualitativo de estas últimas implica grandes desafíos debido a la complejidad, variedad y volumen de los datos generados, requiriendo métodos más efectivos que los tradicionales. El objetivo principal de este artículo es explorar el potencial de la Inteligencia Artificial Generativa (IA Gen), particularmente mediante los LLM, para optimizar el análisis de respuestas abiertas en evaluaciones del desempeño docente, examinando sus beneficios, limitaciones y estrategias específicas para superar desafíos como la precisión y las alucinaciones generadas por estos modelos.

Los LLM son modelos de IA Gen capaces de procesar y generar texto con un alto grado de coherencia y contexto. Están entrenados con enormes cantidades de datos textuales obtenidos de múltiples fuentes, lo que les permite adquirir un conocimiento implícito del lenguaje y la relación entre palabras, frases e ideas. Mediante la utilización de arquitecturas basadas en atención, conocidas como transformadores (*transformers*), estos modelos pueden captar dependencias a largo plazo y comprender el sentido global de los textos (Lin et al., 2023). Esta habilidad les facilita desempeñar tareas complejas, como responder preguntas abiertas, resumir información, interpretar comentarios o mantener conversaciones fluidas. Además, los LLM no se limitan a un único dominio, pueden adaptarse y especializarse, dando lugar a modelos específicos que responden con mayor precisión a las necesidades particulares de un sector o disciplina (Jansen et al., 2023).

La evaluación periódica del personal docente es una herramienta fundamental para garantizar la evolución y mejora continua de la calidad de una institución educativa (Wang et al., 2023). A través de estas evaluaciones, el estudiantado proporciona una valiosa perspectiva sobre diversos aspectos como la calidad de la enseñanza, el contenido educativo y las áreas de mejora, permitiendo al equipo que gestiona el desarrollo del personal docente, obtener información significativa sobre las percepciones y experiencias de los estudiantes respecto a sus docentes.

Estos instrumentos de evaluación del desempeño docente suelen estructurarse en torno a preguntas cerradas y escalas de valoración, pero también incluyen preguntas abiertas que permiten a los estudiantes manifestar sus impresiones con mayor libertad y detalle. Por medio de estas preguntas abiertas, se puede identificar aquellas prácticas pedagógicas que consideran más valiosas, como el uso de ejemplos concretos o el acompañamiento personalizado y, al mismo tiempo, señalar los aspectos que requieren atención y mejora. Por ejemplo, la necesidad de mayor claridad en las explicaciones o una comunicación más efectiva con el grupo. Este tipo de retroalimentación cualitativa aporta una perspectiva más completa y matizada del desempeño docente.

Sin embargo, analizar las respuestas abiertas de estas encuestas presenta desafíos significativos debido a la variedad y al volumen de datos, así como a la necesidad de interpretaciones precisas. Los métodos habituales, como los chatbots tradicionales o los modelos basados en aprendizaje automático supervisado, aunque presentan una menor complejidad técnica y menores costos iniciales, tienen limitaciones significativas en cuanto a escalabilidad, flexibilidad semántica y capacidad de generalización frente a respuestas abiertas (Remadi et al., 2024). Estas limitaciones generan sesgos, lo que dificulta la obtención de conclusiones fiables.

La evolución de los LLM ha permitido el desarrollo de interfaces conversacionales avanzadas alojadas en la nube como GPT-4 de OpenAI, Gemini de Google, o Copilot de Microsoft, los cuales han mostrado avances notables en la capacidad para procesar y analizar lenguaje natural. Paralelamente, el surgimiento de modelos descargables

como Llama y DeepSeek ha abierto nuevas posibilidades para el desarrollo de LLM, ya que permiten entrenar y desplegar modelos localmente sin depender de servicios en la nube ni comprometer la privacidad de los datos. Esta diversidad de enfoques ofrece soluciones flexibles según las necesidades y recursos de cada institución, combinando la escalabilidad de los modelos en la nube con el control y adaptabilidad de los modelos locales. En el análisis de encuestas, estas capacidades resultan particularmente valiosas para interpretar respuestas abiertas que contienen información rica y matizada, difícil de captar mediante métodos tradicionales.

Desarrollo

Evaluación del desempeño docente

La evaluación del desempeño docente es un pilar fundamental para garantizar la calidad educativa, ya que permite identificar fortalezas, áreas de mejora y oportunidades para el desarrollo profesional del profesorado (Stronge, 2018). Sin embargo, diversos estudios señalan que las evaluaciones que se aplican a los docentes presentan múltiples limitaciones. Entre los problemas más recurrentes se encuentran instrumentos con falta de validez y fiabilidad, el uso indebido de estadísticas, sesgos de género y popularidad, y la reducción de la enseñanza a percepciones subjetivas de satisfacción (Boring et al., 2016; Hornstein, 2017).

Particularmente, la evaluación que realizan los estudiantes sobre el desempeño docente tiende a ser interpretada como indicadores objetivos, ignorando que muchos de sus ítems están más relacionados con la experiencia del estudiante que con criterios pedagógicos sustantivos. Esto resulta especialmente problemático cuando las instituciones basan decisiones de promoción o permanencia únicamente en estos instrumentos (Hornstein, 2017; Stark & Freishtat, 2014).

Por otra parte, un problema creciente respecto a la evaluación docente es la retroalimentación realizada utilizando comentarios abiertos. Si bien este tipo de herramientas pueden ofrecer información valiosa sobre la experiencia del alumno en clase, su alcance para evaluar el desempeño docente es limitado. Como señaló Hornstein (2017), estos comentarios deben interpretarse con cautela, ya que los estudiantes pueden describir aspectos vivenciales del curso, pero no están en posición de juzgar con precisión elementos pedagógicos complejos o el diseño didáctico de la enseñanza. Además, Heffernan (2022) expuso que estos comentarios son una fuente de ansiedad y estrés para los docentes, dado que pueden recibir comentarios abusivos y prejuiciosos, alejados de aspectos pedagógicos. Por su parte, Kreitzer y Sweet-Cushman (2022) agregaron que, los comentarios cualitativos presentan serios problemas de confiabilidad y equidad, pues existe evidencia consistente de sesgo de género, así como de novedad y negatividad.

Dado lo anterior, resulta importante repensar los mecanismos de evaluación del desempeño docente, privilegiando enfoques más holísticos, formativos y centrados en la mejora continua. Las instituciones educativas deben complementar las opiniones estudiantiles con múltiples fuentes de evidencia, como la observación entre pares, la autoevaluación reflexiva y el análisis del diseño curricular. Asimismo, es fundamental establecer procesos claros y responsables para la revisión de los comentarios cualitativos, de modo que estos se analicen con criterios éticos y pedagógicos, protegiendo al profesorado de juicios injustos o prejuiciosos. Lo anterior, con el fin de avanzar hacia una evaluación más justa, equitativa y útil, que no solo rinda cuentas, sino que también fomente el desarrollo profesional y la calidad de la enseñanza.

Modelos de lenguaje: transformando el análisis educativo

En este contexto tan complejo, múltiples estudios han explorado el uso de LLM para el análisis de datos textuales provenientes de encuestas y evaluaciones. Si bien se han observado resultados prometedores, se han identificado también limitaciones. Para que estos modelos funcionen de manera óptima, se requiere contar con datos de entrenamiento lo suficientemente extensos y representativos. Además, existe el reto de ajustar el modelo para que comprenda contextos educativos

muy específicos, ya que en sí la habilidad lingüística no garantiza una interpretación precisa de las dinámicas propias del ámbito docente.

A pesar de esto, la literatura muestra avances significativos. Por ejemplo, Gao et al. (2024) propusieron sistemas de evaluación automática capaces de manejar respuestas abiertas a gran escala. Al emplear LLM, se logró una agilización del proceso, reduciendo el tiempo que antes requería la evaluación manual. Esto sugiere un impacto positivo en la eficiencia institucional, por lo que el personal puede centrarse en interpretar las conclusiones y en implementar mejoras, en lugar de dedicar recursos considerables a tareas repetitivas.

Otro ejemplo es el trabajo de Remadi et al. (2024), que aborda la integración de múltiples fuentes de datos para simplificar la extracción de información. Esto puede ser especialmente beneficioso, debido a que los comentarios de los estudiantes suelen presentarse de manera dispersa en encuestas, foros internos, correos electrónicos e incluso publicaciones en plataformas académicas. Un LLM entrenado y afinado para este propósito puede unificar esas fuentes y producir un análisis más coherente, permitiendo decisiones mejor fundamentadas sobre la práctica docente.

Por su parte, Abdou y Eude (2024) demostraron el potencial de modelos pre-entrenados combinados con enfoques de similitud semántica para evaluar automáticamente exámenes. Aunque el enfoque se centra en la calificación de pruebas escritas, los hallazgos son relevantes para el contexto de la evaluación docente. El alto nivel de precisión logrado, con un 96%, indica que, con el entrenamiento y la configuración adecuados, los LLM pueden emular y, en algunos casos, complementar la labor de un evaluador humano. Esta capacidad, trasladada al análisis de respuestas abiertas en encuestas, puede reducir la carga de trabajo y ofrecer una primera clasificación o extracción de patrones clave.

El trabajo de Pinto et al. (2023) destacó la aplicación práctica de ChatGPT, para corregir preguntas abiertas en formación técnica. Lo relevante de su estudio es la comprobación de una concordancia significativa entre las apreciaciones del modelo y las de los expertos humanos. Esto sugiere que los LLM no solo pueden identificar temas o sentimientos en las respuestas, sino también aproximarse al criterio especializado que, en última instancia, determina la calidad de la retroalimentación del alumno en la evaluación docente.

Adicionalmente, Lin y Koedinger (2024) presentaron un sistema que utiliza modelos generativos pre-entrenados para proporcionar retroalimentación explicativa sobre las respuestas de los alumnos. No se limita a señalar errores o fortalezas, sino que explica por qué algo es valioso o problemático. Este avance es útil en la evaluación docente, ya que no solo se obtienen temas a mejorar, sino también sugerencias constructivas que pueden guiar la mejora de la labor educativa. La capacidad de proponer reformulaciones puede extrapolarse al análisis de las respuestas de los estudiantes sobre el profesorado: el LLM identifica áreas de mejora y puede sugerir intervenciones específicas, como el rediseño de una estrategia didáctica o una mayor claridad en la retroalimentación que se ofrece al alumno.

Desafíos de los LLM

A pesar de los avances significativos de los LLM en la comprensión y generación de texto, su uso conlleva riesgos que deben ser cuidadosamente gestionados. Uno de los problemas más señalados es la tendencia de estos modelos a generar 'alucinaciones'. Este fenómeno implica que el modelo produzca información falsa o inexacta, ya sea por malinterpretar el contexto, por carecer de datos suficientes o simplemente por las complejas dinámicas internas de su arquitectura. En entornos donde la fiabilidad es crucial, como las instituciones educativas, este tipo de comportamientos puede resultar sumamente problemático. Una mala interpretación de las respuestas del estudiantado, la sugerencia de prácticas pedagógicas inadecuadas o la identificación errónea de áreas de mejora pueden conducir a una toma de decisiones equivocada, con consecuencias negativas, generando un perjuicio real para la comunidad educativa.

La aparición de alucinaciones se asocia a varios factores. Por un lado, los LLM convencionales suelen estar entrenados con cantidades masivas de datos provenientes de diferentes dominios, temáticas y estilos. Si bien esta

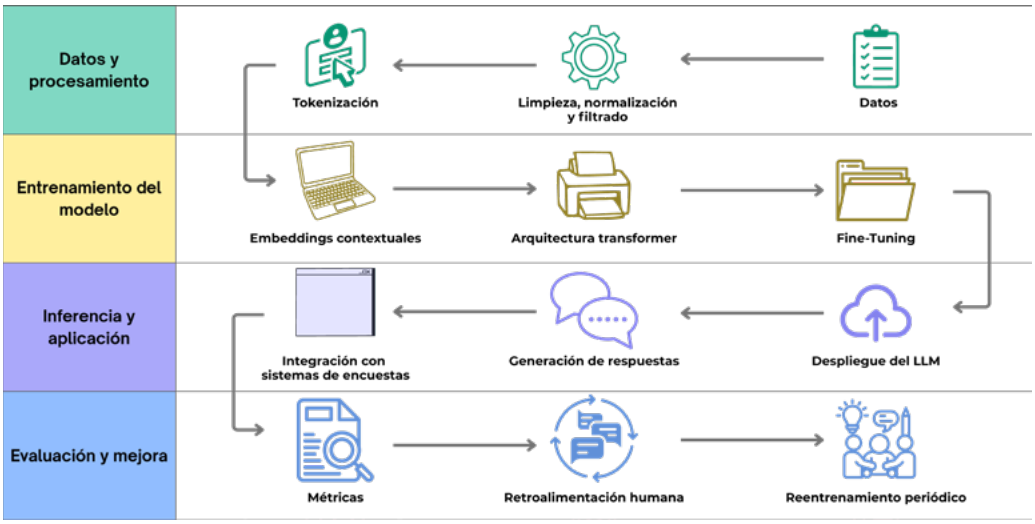
diversidad es útil para lograr versatilidad, también genera un riesgo: el modelo puede combinar fragmentos de información no relacionados y generar respuestas que parecen coherentes, pero carecen de exactitud. Además, las arquitecturas empleadas permiten que el modelo, ante la falta de datos específicos, rellene los huecos con inferencias no verificadas. Un LLM no comprende el mundo como lo haría un ser humano; simplemente, calcula la probabilidad de que una palabra siga a otra a partir de lo visto en su entrenamiento, sin una base ontológica que le asegure la veracidad de lo que afirma.

Modelos de lenguaje especializados

Conscientes de este reto, se han propuesto tres estrategias que buscan minimizar las alucinaciones y ofrecer herramientas más seguras y precisas: los LLM de dominio específico, los Pequeños Modelos de Lenguaje (SLM, por sus siglas en inglés, Small Language Models) y los modelos en la nube con entrenamiento *few-shot*. Estas soluciones parten de la premisa de que no siempre es conveniente o necesario contar con un modelo de gran tamaño, entrenado con datos de todo tipo y con un alcance tan amplio que pierda de vista la precisión en contextos particulares. La estrategia en cada caso consiste en enfocar el entrenamiento y diseño del modelo en un entorno de datos más controlado, reduciendo así el margen de error y mejorando la fiabilidad de las respuestas.

Los LLM de dominio específico se construyen tomando como base un gran modelo genérico, pero luego se le somete a un proceso de afinamiento (*Fine-Tuning*) utilizando datos propios del ámbito educativo, en este caso, las respuestas de las evaluaciones docentes, en lugar de exponer al modelo a todo tipo de textos. Este proceso de especialización refina el entendimiento del modelo, ya que aprende las convenciones, el lenguaje, las necesidades y las expectativas del dominio en cuestión. Con ello, disminuye la probabilidad de que el modelo se desvíe hacia inferencias alejadas del contexto o hacia la generación de información inexacta. La Figura 1 muestra la arquitectura de un LLM de dominio específico.

Figura 1
Arquitectura de un LLM de dominio específico



Nota: Elaboración propia.

La arquitectura consiste en un proceso en varias capas, cada una con una función clara y complementaria. En la primera capa, se reúnen y preparan los datos del dominio, en este caso, la evaluación del desempeño docente, lo que implica recolectar el corpus especializado, eliminar información irrelevante y convertir el contenido en un formato que pueda ser procesado fácilmente por el modelo (técnicas de tokenización). Una vez que los datos están listos, la segunda capa se centra en la representación y el entrenamiento del modelo. Para ello, se representa el significado de las palabras en función del contexto en el que aparecen, lo que permite identificar relaciones y patrones en el lenguaje. Se utilizan redes neuronales avanzadas con mecanismos de atención, que ayudan al modelo a enfocarse en las partes más relevantes del texto y a comprenderlo de manera más detallada. Después de este afinamiento (*Fine-Tuning*) del conjunto de datos del dominio, el modelo pasa a la tercera capa, en la cual se despliega y entra en la fase de aplicación, siendo capaz de responder preguntas abiertas, integrarse con sistemas de análisis de encuestas y proporcionar información útil a los responsables de la toma de decisiones. Finalmente, la cuarta capa se dedica a la evaluación y mejora continua. En este nivel, se aplican métricas para medir la calidad de las respuestas y se incorpora la retroalimentación humana. Esta retroalimentación, junto con la actualización periódica de datos y ajustes del modelo, garantiza que el LLM se mantenga relevante, preciso y alineado con las necesidades cambiantes del entorno educativo.

Esta arquitectura, no solo reduce la probabilidad de alucinaciones, sino que también garantiza una mayor pertinencia en las respuestas. Por ejemplo, si un LLM de dominio específico es utilizado para analizar respuestas abiertas de encuestas docentes, ya no se limitará a reconocer patrones estadísticos del lenguaje general, sino que comprenderá mejor qué tipo de comentarios son comunes en evaluaciones de profesores, cuáles son las temáticas más habituales y cómo variar el matiz según el nivel educativo o la disciplina. De este modo, las respuestas generadas resultarán más relevantes, contextualizadas y útiles para la toma de decisiones educativas.

Otra vía propuesta para controlar las alucinaciones y mejorar la fiabilidad son los SLM. En lugar de perseguir la magnitud y la complejidad de los LLM tradicionales, que contienen miles de millones de parámetros, la idea es trabajar con modelos más compactos y manejables. Estos modelos más pequeños requieren menos recursos computacionales y son más sencillos de entrenar y desplegar. Pero su principal ventaja en el contexto educativo es que, al ser más reducidos, se facilita el control sobre sus parámetros y se puede intervenir de forma más precisa en su comportamiento.

Los SLM permiten incorporar restricciones explícitas que limiten la capacidad del modelo de generar alucinaciones. Esto se logra mediante el uso de técnicas de regularización, *Fine-Tuning* con conjuntos de datos cuidadosamente curados o integración de mecanismos de validación cruzada. La relativa simplicidad de estos modelos les da a los responsables técnicos y académicos un mayor margen para entender su funcionamiento interno y realizar modificaciones oportunas. De este modo, las instituciones pueden adecuar el modelo a sus necesidades, ajustando el tipo de información que se espera y filtrando los resultados antes de presentarlos a los responsables de la toma de decisiones. Esta menor complejidad también facilita la transparencia y la interpretabilidad, aspectos fundamentales para generar confianza en las herramientas tecnológicas empleadas en el ámbito educativo.

Entre las principales características de los SLM (Figura 2), se destaca la accesibilidad, dado que no requieren las costosas infraestructuras de cómputo que suelen exigir los LLM de mayor escala. Esto favorece su implementación en instituciones educativas con recursos limitados, democratizando el uso de herramientas avanzadas de análisis de texto. Al ser más ligeros, resultan también más eficientes en términos de tiempo de entrenamiento y respuesta, permitiendo una integración ágil con los sistemas existentes. Además, su naturaleza reducida contribuye a disminuir la probabilidad de alucinaciones, puesto que, con menos parámetros que ajustar, es más sencillo controlar su comportamiento y garantizar una mayor precisión. Finalmente, su tamaño se traduce en un menor consumo energético, lo que los

hace más ecológicos al reducir la huella ambiental del procesamiento y la adaptación continua del modelo.

Figura 2
Características de los pequeños modelos de lenguaje (SLM)

Nota: Elaboración propia.

Una tercera estrategia es el uso de LLM en la nube, como ChatGPT, Gemini o Claude, que pueden beneficiarse del enfoque de entrenamiento *few-shot* para mejorar su precisión en tareas específicas sin necesidad de un ajuste completo. A diferencia de los LLM de dominio específico, que requieren un proceso de *Fine-Tuning* intensivo, los modelos en la nube pueden adaptarse a tareas concretas mediante ejemplos representativos que les ayuden a comprender el contexto y reducir el margen de error en sus respuestas. En este enfoque, se proporciona al modelo un conjunto limitado de ejemplos de preguntas y respuestas esperadas, lo que le permite captar patrones y ajustar sus predicciones de manera más precisa.

El entrenamiento *few-shot* es particularmente útil en el análisis de evaluaciones docentes porque permite a los modelos pre-entrenados interpretar respuestas abiertas con mayor precisión, sin necesidad de acceso directo a grandes volúmenes de datos. Este enfoque reduce el riesgo de sesgos y alucinaciones, y facilita la implementación de modelos personalizados sin los costos computacionales y técnicos del *Fine-Tuning* tradicional.

Sin embargo, estos enfoques tienen implicaciones particulares que pueden afectar su adopción efectiva. Los LLM de dominio específico requieren una infraestructura costosa para su entrenamiento inicial, así como personal técnico especializado en procesamiento de datos y en la optimización del modelo. Esto puede ser un obstáculo significativo para instituciones educativas pequeñas o con recursos limitados (Jansen et al., 2023). La principal desventaja de los SLM radica en su capacidad limitada para manejar tareas lingüísticas complejas o contextos muy diversos, dado que su reducido tamaño puede restringir significativamente la precisión en análisis que requieran mayor profundidad o matices semánticos complejos (Gao et al., 2024). Respecto a los modelos basados en la nube con entrenamiento *few-shot*, su uso puede generar costos recurrentes (ligados a la cantidad de consultas realizadas), así como implicaciones significativas en términos de privacidad y protección de datos personales, dada la externalización del procesamiento (Zhang y Tian, 2024).

Estas consideraciones permiten identificar que las tres estrategias ofrecen diferentes niveles de especialización y adaptabilidad para el uso de modelos de lenguaje en contextos educativos heterogéneos. Mientras que los LLM de dominio específico garantizan un alto nivel de personalización y precisión a costa de mayores

requerimientos computacionales, los SLM proporcionan una alternativa eficiente y accesible para tareas más controladas. Por su parte, los modelos en la nube con entrenamiento *few-shot* permiten flexibilidad y rápida implementación en instituciones que buscan mejorar la interpretación de datos cualitativos sin necesidad de entrenar modelos desde cero. Cada enfoque responde a necesidades distintas, pero en todos los casos, el objetivo es el mismo: mejorar la calidad del análisis de los datos y reducir los riesgos asociados a las alucinaciones en modelos de IA Gen.

Casos de éxito

El uso de LLM en la evaluación del desempeño docente ha comenzado a generar resultados prometedores, especialmente en el ya mencionado análisis automatizado de respuestas abiertas en dichas evaluaciones. Un estudio reciente llevado a cabo en la Escuela Eshelman de Farmacia de la University of North Carolina exploró la viabilidad de emplear ChatGPT para procesar los comentarios que los estudiantes realizan en las encuestas de evaluación docente (Fuller et al., 2024). Tradicionalmente, la revisión manual de estos comentarios implica una inversión considerable de tiempo por parte del profesorado. Se utilizó la versión GPT-3.5 turbo de ChatGPT para identificar patrones recurrentes en los comentarios estudiantiles y compararlos con los hallazgos obtenidos por los instructores humanos.

Los resultados mostraron una alta concordancia entre los temas identificados por ChatGPT y los seleccionados manualmente, alcanzando niveles de coincidencia de hasta el 82% en aspectos generales del curso y entre el 53% y el 81% en la clasificación de áreas de mejora. Además, el modelo redujo el tiempo necesario para analizar los comentarios, completando la tarea en un rango de 10 a 12 minutos, en contraste con los casi 30 minutos requeridos por los docentes. Esto sugiere que los LLM pueden desempeñar un papel complementario en la evaluación docente al agilizar la síntesis de información sin comprometer la calidad del análisis. Sin embargo, el estudio también señaló algunas limitaciones, como la posible omisión de comentarios minoritarios pero relevantes, así como la necesidad de garantizar la privacidad de los datos al utilizar modelos alojados en la nube. En general, la investigación concluye que, aunque ChatGPT no reemplaza el juicio experto del profesorado, puede ser una herramienta valiosa para optimizar el proceso de análisis de evaluaciones docentes y permitir a los docentes enfocar sus esfuerzos en la mejora de su práctica educativa.

Otro estudio, realizado en una escuela secundaria en China, exploró el uso de LLM para clasificar respuestas abiertas en encuestas de evaluación docente (Zhang y Tian, 2024). Para ello, se utilizó ChatGPT en dos modalidades: sin ejemplos previos de clasificación (*zero-shot*) y con algunos ejemplos de referencia (aprendizaje *few-shot*). Su desempeño se comparó con modelos de aprendizaje profundo previamente entrenados en datos específicos, incluyendo un LLM, basado en BERT, de dominio específico.

Los hallazgos mostraron que ChatGPT logró una clasificación aceptable de los comentarios abiertos, con un desempeño especialmente notable en los casos donde se le proporcionaron ejemplos previos. Esto sugiere que, incluso sin un entrenamiento especializado, los LLM en la nube pueden adaptarse a tareas educativas y ofrecer análisis útiles con ajustes mínimos. No obstante, el modelo de dominio específico superó en precisión a ChatGPT, lo que sugiere que, si se dispone de grandes volúmenes de datos etiquetados, los modelos ajustados a un contexto particular (LLM de dominio específico) pueden ser más efectivos.

Impacto en las instituciones educativas

Tanto los LLM de dominio específico como los SLM y los LLM en la nube con entrenamiento *few-shot* ofrecen beneficios concretos para las instituciones educativas que deseen incorporar herramientas de análisis automatizado. La reducción de alucinaciones no es solo una cuestión de mejorar la exactitud de la información, sino también de reforzar la credibilidad del sistema. Al disponer de modelos que operan sobre datos pertinentes y delimitados, el personal responsable de la toma de decisiones puede adoptar estas soluciones con mayor confianza.

Además, la implementación de estos modelos con un enfoque más acotado en términos de dominio o tamaño permite optimizar el uso de recursos. Las instituciones educativas, a menudo con presupuestos ajustados, pueden encontrar en los SLM una alternativa más eficiente y económica, ya que no requieren la misma infraestructura computacional ni los costos asociados al mantenimiento de un LLM tradicional. Del mismo modo, los LLM de dominio específico pueden ser entrenados a partir de un modelo base, lo que reduce la necesidad de comenzar desde cero y abarata el proceso de entrenamiento. Esta eficiencia económica, sumada a la mejora en la calidad informativa, constituye un incentivo claro para la adopción de estas estrategias. Por su parte, los modelos en la nube con entrenamiento *few-shot* permiten flexibilidad y rápida implementación en instituciones que buscan mejorar la interpretación de datos cualitativos sin necesidad de entrenar modelos desde cero ni mantener una infraestructura tecnológica.

En última instancia, la adopción de alguno de estos modelos ofrece a las instituciones educativas la posibilidad de beneficiarse plenamente de la IA Gen aplicada al análisis de datos cualitativos sin sacrificar la precisión ni la pertinencia. Al reducir el riesgo de alucinaciones, estas herramientas mejoran la calidad de la retroalimentación obtenida de las evaluaciones de desempeño docentes, facilitando la identificación de fortalezas y la detección temprana de áreas de mejora. Asimismo, permiten un aprovechamiento más ágil y eficaz de los recursos, tanto en términos computacionales como de personal. El resultado es una toma de decisiones más informada, estratégica y confiable, que redunda en la mejora de la experiencia de enseñanza-aprendizaje.

Justificación conceptual

Este ensayo se enmarca en el paradigma del análisis automatizado de datos cualitativos mediante IA Gen, particularmente con LLM. Desde un enfoque conceptual, Vaswani et al. (2017) señalaron que los LLM se fundamentan en la teoría de redes neuronales transformadoras comúnmente conocidos como *transformers*, la cual ha revolucionado el procesamiento del lenguaje natural por su capacidad de captar dependencias contextuales a largo plazo. Esta capacidad le permite abordar el desafío específico del análisis de respuestas abiertas en evaluaciones del desempeño docente, un contexto caracterizado por su complejidad lingüística y semántica, difícil de manejar con métodos tradicionales como el análisis temático manual o técnicas estadísticas básicas.

La elección de los enfoques discutidos en este ensayo: LLM de dominio específico, Pequeños Modelos de Lenguaje (SLM), y modelos basados en entrenamiento *few-shot*, se justifica por su potencial para mitigar las limitaciones inherentes a los LLM generales, principalmente la generación de alucinaciones y la falta de precisión contextual. Cada uno de estos enfoques tiene fortalezas distintivas: los LLM de dominio específico se destacan por su alta precisión en contextos educativos específicos debido a su especialización mediante datos del dominio educativo (Jansen et al., 2023); los SLM ofrecen eficiencia operativa, accesibilidad tecnológica y mayor facilidad de ajuste y control (Gao et al., 2024), y finalmente, los modelos con entrenamiento *few-shot* en la nube aportan flexibilidad y rápida implementación, siendo útiles especialmente cuando la disponibilidad de datos es limitada (Zhang y Tian, 2024).

Se consideraron también enfoques alternativos, como métodos tradicionales de análisis temático manual, análisis basado en reglas lingüísticas y análisis estadístico no supervisado; sin embargo, se descartaron debido a limitaciones clave como la escalabilidad limitada, el sesgo del evaluador humano y las dificultades para capturar matices semánticos complejos (Remadi et al., 2024; Wang et al., 2023). En conjunto, las fortalezas y limitaciones analizadas proporcionan un fundamento científico claro que sustenta la elección conceptual adoptada para este ensayo.

Conclusiones

La evaluación del desempeño docente es un componente esencial para el fortalecimiento de los sistemas educativos, ya que permite asegurar la calidad de la enseñanza, reconocer buenas prácticas pedagógicas y detectar áreas

de mejora. Cuando se realiza con criterios claros, éticos y formativos, esta evaluación se convierte en una herramienta poderosa para el desarrollo profesional docente y la toma de decisiones institucionales. En este sentido, la incorporación de tecnologías avanzadas para el análisis de evaluaciones, como los LLM, abre nuevas posibilidades para mejorar la gestión y el aprovechamiento de los datos derivados de estos procesos. La integración de LLM en el análisis de la evaluación docente representa un avance significativo, pero también plantea desafíos éticos y técnicos que deben ser considerados con responsabilidad.

En primer lugar, el impacto de estos modelos puede ser profundo y multifacético. Facilitan una gestión más eficiente y precisa de la retroalimentación docente, mejorando la calidad y rapidez del análisis de grandes volúmenes de datos cualitativos. Esto permite focalizar esfuerzos y recursos humanos en la interpretación estratégica de los resultados y en la implementación efectiva de mejoras docentes específicas.

Para evaluar el valor añadido de un sistema de IA Gen en este contexto, se proponen diversas métricas clave: la eficiencia operativa, medida como el tiempo medio de procesamiento por encuesta; el costo por mil respuestas procesadas, útil para comparar la viabilidad económica frente al análisis manual; la satisfacción de los usuarios institucionales, que puede evaluarse mediante encuestas, y el porcentaje de docentes que implementan acciones correctivas basadas en los informes automatizados, lo cual refleja directamente el impacto pedagógico de este enfoque.

Sin embargo, es importante que el personal docente y de gestión educativa comprenda las limitaciones de estas herramientas, evitando una confianza ciega en los resultados del modelo. La supervisión humana sigue siendo esencial, tanto para validar los hallazgos como para interpretar matices que el modelo pueda pasar por alto.

Además, la adopción de LLM en este contexto conlleva consideraciones éticas relevantes. El uso de datos de estudiantes y docentes debe cumplir estrictamente con las normativas de privacidad y protección de datos, especialmente cuando se utilizan modelos en la nube, donde la información es procesada y almacenada en servidores externos a las instituciones educativas.

Finalmente, para que los LLM mantengan su relevancia y efectividad, es fundamental su actualización continua. El lenguaje evoluciona, las metodologías docentes cambian y las expectativas estudiantiles se transforman con el tiempo. Entrenar los modelos periódicamente con datos recientes y representativos garantizará la calidad y pertinencia del análisis.

Agradecimientos

Agradecemos el apoyo de la Secretaría de Ciencia, Humanidades, Tecnología e Innovación (SECIHTI) a través de la beca de Roberto E. Ramos-Rivera (CVU: 826812).

Referencias

Abdou, I., & Eude, T. (2024). Open-ended questions automated evaluation: Proposal of a new generation. *Proceedings of the 2023 International Joint Conference on Robotics and Artificial Intelligence*, 143–147. https://doi.org/10.1145/3632971.3632980

Álvarez, B. A, Acosta-Díaz, R., & Morales-Vanegas, E. A. (2024). Privacy-Aware Artificial Intelligence: A Review of Design Principles and Applications. *Avances en Interacción Humano-Computadora, 9*(1), 209–213. https://doi.org/10.47756/aihc.y9i1.169

Boring, A., Ottoboni, K., & Stark, P. B. (2016). Student evaluations of teaching (mostly) do not measure teaching effectiveness. *ScienceOpen Research*, 1-11. https://doi.org/10.14293/S2199-1006.1.SOR-EDU.AETBZC.v1

Fuller, K. A., Morbitzer, K. A., Zeeman, J. M., Persky, A. M., Savage, A. C., & McLaughlin, J. E. (2024). Exploring the use of ChatGPT to analyze student course evaluation comments. *BMC Medical Education, 24*(423), 1-8. https://doi.org/10.1186/s12909-024-05316-2

Gao, R., Merzdorf, H. E., Anwar, S., Hipwell, M. C., & Srinivasa, A. R. (2024). Automatic assessment of text-based responses in post-secondary education: A systematic review. *Computers and Education: Artificial Intelligence, 6*, 1-15. https://doi.org/10.1016/j.caeai.2024.100206

Heffernan, T. (2022). Sexism, racism, prejudice, and bias: A literature review and synthesis of research surrounding student evaluations of courses and teaching. *Assessment & Evaluation in Higher Education, 47*(1), 144–154. https://doi.org/10.1080/02602938.2021.1888075

Hornstein, H. A. (2017). Student evaluations of teaching are an inadequate assessment tool for evaluating faculty performance. *Cogent Education, 4*(1), 1304016. https://doi.org/10.1080/2331186X.2017.1304016

Jansen, B. J., Jung, S., & Salminen, J. (2023). Employing large language models in survey research. *Natural Language Processing Journal, 4*, 1-7. https://doi.org/10.1016/j.nlp.2023.100020

Kreitzer, R. J., & Sweet-Cushman, J. (2022). Evaluating Student Evaluations of Teaching: A Review of Measurement and Equity Bias in SETs and Recommendations for Ethical Reform. *Journal of Academic Ethics, 20*(1), 73–84. https://doi.org/10.1007/s10805-021-09400-w

Lin, C.-C., Huang, A. Y. Q., & Yang, S. J. H. (2023). A Review of AI-Driven Conversational Chatbots Implementation Methodologies and Challenges (1999–2022). *Sustainability, 15*(5), 1-13. https://doi.org/10.3390/su15054012

Lin, J., & Koedinger, K. R. (2024). HAROR: A System for Highlighting and Rephrasing Open-Ended Responses. *Proceedings of the Eleventh ACM Conference on Learning @ Scale*, 553–555. https://doi.org/10.1145/3657604.3664721

Pinto, G., Cardoso-Pereira, I., Monteiro, D., Lucena, D., Souza, A., & Gama, K. (2023). Large Language Models for Education: Grading Open-Ended Questions Using ChatGPT. *Proceedings of the XXXVII Brazilian Symposium on Software Engineering*, 293–302. https://doi.org/10.1145/3613372.3614197

Remadi, A., El Hage, K., Hobeika, Y., & Bugiotti, F. (2024). To prompt or not to prompt: Navigating the use of Large Language Models for integrating and modeling heterogeneous data. *Data & Knowledge Engineering, 152*, 1-17. https://doi.org/10.1016/j.datak.2024.102313

Stark, P. B., & Freishtat, R. (2014). An evaluation of course evaluations. *ScienceOpen Research*, 1-7. https://doi.org/10.14293/S2199-1006.1.SOR-EDU.AOFRQA.v1

Stronge, J. H. (2018). *Qualities of effective teachers* (3rd edition). ASCD.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). *Attention Is All You Need*, 1-15. https://doi.org/10.48550/ARXIV.1706.03762

Wang, Z., Denny, P., Leinonen, J., & Luxton-Reilly, A. (2023). Leveraging Large Language Models for Analysis of Student Course Feedback. *Proceedings of the 16th Annual ACM India Compute Conference*, 76–79. https://doi.org/10.1145/3627217.3627221

Zhang, B., & Tian, X. (2024). Capturing fine-grained teacher performance from student evaluation of teaching via ChatGPT. *Education and Lifelong Development Research, 1*(4), 156–169. https://doi.org/10.46690/elder.2024.04.01