EVALUACIÓN FORMATIVA AUTOMATIZADA CON IA EN EDUCACIÓN SUPERIOR

AUTOMATED FORMATIVE ASSESSMENT WITH AI IN HIGHER EDUCATION

Paúl Geovanny Amén Mora[1*]

[1] Docente de la Carrera Educación, Facultad de Ciencias Sociales, Humanísticas y de la Educación, Universidad Estatal del Sur de Manabí, Ecuador. ORCID: https://orcid.org/0009-0007-1962-4015. Correo: paul.amen@unesum.edu.ec

Rodrigo Alexander Rincón Zambrano[2]

[2] Docente de la Carrera Educación, Facultad de Ciencias Sociales, Humanísticas y de la Educación, Universidad Estatal del Sur de Manabí, Ecuador. ORCID: https://orcid.org/0000-0002-2713-5111. Correo: jahiry.molineros@unesum.edu.ec

Jahiry del Rosario Molineros Ronquillo[3]

[3] Docente de la Carrera Educación, Facultad de Ciencias Sociales, Humanísticas y de la Educación, Universidad Estatal del Sur de Manabí, Ecuador. ORCID: https://orcid.org/0009-0009-7013-5801. Correo: jahiry.molineros@unesum.edu.ec

José Miguel Bacusoy Sanchez[4]

[4] Docente de la Carrera Educación, Facultad de Ciencias Sociales, Humanísticas y de la Educación, Universidad Estatal del Sur de Manabí, Ecuador. ORCID: https://orcid.org/0009-0007-4611-8097. Correo: jose.bacusoy@unesum.edu.ec

*** Autor para correspondencia: paul.amen@unesum.edu.ec**

Resumen

Las preguntas de respuesta abierta en evaluación formativa requieren retroalimentación oportuna, específica y criterial para ser pedagógicamente efectivas, pero su producción manual es costosa en tiempo y presenta variabilidad entre evaluadores. Este estudio analiza la implementación de un sistema de evaluación formativa asistido por inteligencia artificial desplegado en la Universidad Estatal del Sur de Manabí (UNESUM), que

automatiza la calificación de preguntas abiertas mediante rúbricas analíticas explícitas y genera retroalimentación personalizada a escala de curso. La arquitectura integra Google Forms para recolección, Google Sheets como repositorio con rúbricas estructuradas, Google Apps Script para orquestación con procesamiento por lotes y validación de salidas JSON, y Gmail para comunicación automática de feedback HTML. Se evaluó el desempeño operativo en seis paralelos (n=113 estudiantes) midiendo latencia de evaluación, robustez del sistema y calidad de la retroalimentación. Los resultados muestran tiempos de procesamiento de 1:42 a 2:01 minutos por paralelo, con comunicación de retroalimentación en aproximadamente 1 segundo por estudiante, representando una reducción del 96.8% al 98.4% del tiempo docente respecto a métodos manuales. El sistema genera feedback criterial (escala 1-4 por criterio) con justificaciones específicas, recomendaciones accionables y transparencia mediante inclusión de respuestas originales. La solución demostró ser escalable, replicable y pedagógicamente alineada con principios de evaluación formativa, redefiniendo el rol docente hacia funciones de diseño y supervisión mientras mantiene el control profesional sobre las decisiones evaluativas.

Palabras clave: evaluación; retroalimentación; automatización; rúbricas; inteligencia; aprendizaje

Abstract

Open-ended questions in formative assessment require timely, specific, and criteria-based feedback to be pedagogically effective, but their manual production is time-consuming and presents inter-rater variability. This study analyzes the implementation of an artificial intelligence-assisted formative assessment system deployed at the Universidad Estatal del Sur de Manabí (UNESUM), which automates the grading of open-ended questions using explicit analytic rubrics and generates personalized feedback at a course scale. The architecture integrates Google Forms for collection, Google Sheets as a repository with structured rubrics, Google Apps Script for orchestration with batch processing and JSON output validation, and Gmail for the automatic communication of HTML feedback. Operational performance was evaluated across six parallel courses (n=113 students) by measuring evaluation latency, system robustness, and feedback quality. The results show processing times ranging from 1:42 to 2:01 minutes per parallel course, with feedback communication at approximately 1 second per student, representing a 96.8% to 98.4% reduction in teacher time compared to manual methods. The system generates criteria-based feedback (a 1-4 scale per criterion) with specific justifications, actionable recommendations, and transparency through the inclusion of original answers. The solution proved to be scalable, replicable, and pedagogically aligned with formative assessment principles, redefining the teacher's role towards design and supervision functions while maintaining professional control over evaluative decisions.

Keywords: *assessment; feedback; automation; rubrics; artificial intelligence; learning*

Fecha de recibido: 09/05/2025
Fecha de aceptado: 08/08/2025
Fecha de publicado: 13/08/2025

Introducción

Las preguntas de respuesta abierta permiten observar razonamiento de orden superior: análisis, síntesis, transferencia y argumentación. También hacen visibles concepciones erróneas que los ítems cerrados suelen ocultar. Su valor formativo depende de la retroalimentación: esta es eficaz cuando reduce la distancia entre el desempeño actual y el esperado al responder con claridad a las preguntas ¿hacia dónde ir?, ¿cómo va? y ¿qué sigue?, y cuando llega en el momento oportuno para orientar la siguiente acción de aprendizaje (Hattie & Timperley, 2007).

El principal desafío operativo es el costo de producir esta retroalimentación: leer cada respuesta, contrastar con rúbricas criteriales, justificar los puntajes y redactar recomendaciones accionables exige varios minutos por estudiante. En grupos medianos o grandes, el volumen de trabajo y los retrasos resultantes terminan afectando el aprendizaje. La evidencia reciente en educación superior indica que la oportunidad, la claridad y la consistencia de la retroalimentación siguen siendo aspectos problemáticos desde la percepción estudiantil, precisamente los atributos que más determinan su utilidad. Esto refuerza la necesidad de rediseñar procesos para sostener la calidad sin aumentar la carga docente (Williams, 2024).

El problema no es únicamente de volumen, sino también de fiabilidad. Investigaciones en calificación de respuestas construidas muestran variabilidad entre evaluadores relevante; prácticas como el holistic scoring con un solo calificador pueden no alcanzar coeficientes de generalizabilidad aceptables, lo que plantea la necesidad de estandarizar procedimientos y apoyar la labor docente con herramientas que reduzcan la variabilidad y aceleren la devolución (Huang & Whipple, 2023). En paralelo, en contextos sumativos de alto impacto, se ha cuestionado la superioridad intrínseca del formato abierto frente a la opción múltiple; la discusión, no obstante, resulta menos pertinente para la evaluación formativa, donde la riqueza de la evidencia cualitativa exige devoluciones rápidas, específicas y transparentes (Hift, 2014).

Las soluciones institucionales más extendidas (Google Workspace y los LMS) resuelven con solvencia la recolección y organización de evidencias, pero dejan sin cubrir el ciclo completo que exige la evaluación formativa de respuestas abiertas: calificación criterial consistente, salida estructurada y comunicación de retroalimentación de alta calidad a escala.

Google Forms permite convertir formularios en cuestionarios, definir claves de respuesta y puntajes, y añadir comentarios; sin embargo, el modelo de calificación está orientado a ítems cerrados o coincidencias exactas y no implementa rúbricas de desempeño para texto abierto (p. ej., escala 1–4 por criterio con descriptores), por lo que la revisión de respuestas abiertas sigue siendo manual.

Google Classroom sí ofrece rúbricas, pero ligadas a tareas dentro de Classroom; no existe un acoplamiento nativo que aplique automáticamente una rúbrica de Classroom a respuestas abiertas recogidas en Forms/Sheets, lo que obliga a procesos separados o desarrollos ad hoc.

Ni Forms ni Sheets incluyen mecanismos nativos para (a) traducir rúbricas a instrucciones ejecutables, (b) invocar modelos de IA y controlar su salida, y (c) generar artefactos normalizados (p. ej., tablas por criterio con explicaciones) listos para auditoría. La literatura reciente mapea beneficios y desafíos del uso de IA en educación superior, pero predominan experiencias en plataformas propietarias o prototipos puntuales; hay

pocas descripciones operativas y replicables de pipelines integrados Forms → Sheets → Apps Script → correo (Bond et al., 2024).

Cualquier solución práctica en Apps Script está sujeta a cuotas y límites (p. ej., tiempo máximo de ejecución, límites diarios por servicio y usuario), lo que hace necesario el procesamiento por lotes y la persistencia incremental para evitar fallos en corridas largas.

Asimismo, el envío masivo de retroalimentación por correo debe respetar los límites de envío de Gmail/Google Workspace (p. ej., destinatarios/mensajes por día), que condicionan la cadencia y requieren estrategias de rate-limiting.

Confiabilidad de la salida de IA: necesidad de validación estructural

Los modelos generativos pueden producir salidas plausibles pero no fieles o con estructura inconsistente, lo que impide su uso directo para calificación criterial sin capas de validación (p. ej., exigir JSON y verificar esquema). El fenómeno de las alucinaciones y sus mitigaciones está ampliamente documentado, por lo que se requieren controles de formato y diseño de prompts adecuados (Ji et al., 2023).

Cerrar esta brecha exige un enfoque socio-técnico: diseño pedagógico (rúbricas explícitas y criterios operativos) más ingeniería ligera (Apps Script) para resolver orquestación, loteo, validación de salidas y comunicación bajo las cuotas reales del ecosistema. El caso práctico que se presenta aborda precisamente estas dimensiones.

El objetivo general de este estudio es analizar un caso práctico operativo de un sistema de evaluación formativa asistido por IA, desplegado en la UNESUM, que apoyado en Google Forms, Google Sheets y Google Apps Script automatiza la calificación de preguntas abiertas mediante rúbricas explícitas y genera retroalimentación personalizada y trazable a escala de curso; específicamente, se abordan:

(i) El diseño e implementación end-to-end de una arquitectura con recolección en Forms, centralización y rúbricas en Sheets, un motor de evaluación en Apps Script con procesamiento por lotes, salida JSON validada y capa agnóstica de proveedor de IA (OpenAI/Gemini), además de un módulo de comunicación por correo en HTML;

(ii) La evaluación del desempeño operativo se realizó mediante la medición de la latencia de evaluación (desde el disparo hasta la escritura completa en Sheets), del tiempo de comunicación (envío de correos personalizados), de la robustez (tasa de reintentos por formato/JSON y finalización sin errores) y del uso de cuotas (cumplimiento de los límites de Apps Script y Gmail);

(iii) La caracterización de la calidad de la retroalimentación a partir del análisis de los artefactos generados: hojas “Resultado P1/P2…”, hoja resumen por estudiante y correos HTML; en términos de claridad, especificidad por criterio (escala 1–4 con justificación), transparencia (inclusión de la respuesta original) y orientaciones accionables;

(iv) La replicabilidad y gobernanza, con lineamientos técnicos y plantillas (estructuras de hojas, configuración y fragmentos de código clave) y prácticas de seguridad y control (uso de PropertiesService, bitácora de ejecución, versionado de rúbricas y prompts) para facilitar la adopción institucional; y

(v) La alineación pedagógica, mostrando cómo la automatización sostiene principios de evaluación formativa: oportunidad, especificidad y consistencia del feedback y redefine el rol docente hacia el diseño y la auditoría del proceso evaluativo sin delegar el juicio profesional.

Este trabajo se enmarca dentro del proyecto de Vinculación "Alfabetización e Integración de TIC en prácticas pedagógicas de unidades educativas y comunidades del Sur de Manabí" y contribuye al proyecto de investigación "Fortalecimiento del nivel didáctico-pedagógico de los docentes en las instituciones educativas del cantón Jipijapa", aportando una solución práctica y replicable para la automatización de procesos evaluativos formativos.

Materiales y métodos

La evaluación formativa es un ciclo continuo: recoger y leer evidencias para acercar el desempeño a las metas. El feedback guía el aprendizaje si contesta tres preguntas clave: ¿hacia dónde vamos? (objetivos y criterios), ¿cómo vamos? (diagnóstico frente a estándares) y ¿qué sigue? (pasos de mejora). Esta arquitectura conceptual se ha vinculado de manera consistente con mejoras en el aprendizaje, siempre que el mensaje aporte información útil y esté alineado con las metas de la tarea (Hattie & Timperley, 2007; Black & Wiliam, 1998).

Además de la estructura, el momento de entrega (*timing*) condiciona la efectividad formativa: devoluciones rápidas, específicas y accionables favorecen la corrección de errores y la transferencia de estrategias; los retrasos prolongados disminuyen su utilidad pedagógica. Síntesis influyentes subrayan que el feedback debe ser informativo más que meramente evaluativo, centrado en la tarea y el proceso, y comunicado con suficiente inmediatez para orientar el siguiente intento (Shute, 2008).

La orientación por criterios fortalece la transparencia y la autorregulación del estudiantado. Cuando la devolución se refiere a rúbricas explícitas (criterios, descriptores y niveles), el estudiante puede identificar con precisión brechas específicas por ejemplo, coherencia argumentativa o uso de evidencias y planificar acciones para cerrarlas. Este enfoque se articula con modelos que vinculan evaluación formativa y aprendizaje autorregulado, y con la noción de feedback literacy, entendida como las capacidades para interpretar y usar la retroalimentación de manera productiva (Nicol & Macfarlane-Dick, 2006; Carless & Boud, 2018).

La evidencia de síntesis cuantitativa respalda estas pautas: metaanálisis recientes reportan efectos positivos moderados, modulados por la calidad informativa del feedback, su nivel de foco (tarea, proceso, autorregulación) y su oportunidad dentro del ciclo instruccional (Wisniewski et al., 2020). En términos operativos, ello exige procesos capaces de ofrecer devoluciones criteriales, claras y oportunas a escala de curso, minimizando la variabilidad y los retrasos que suelen afectar la práctica docente.

Rúbricas analíticas vs. holísticas; fiabilidad entre evaluadores

Las rúbricas holísticas producen un juicio global del desempeño a partir de una valoración integrada, mientras que las analíticas desagregan la tarea en criterios y niveles con descriptores específicos. Esta distinción tiene consecuencias pedagógicas y métricas: las rúbricas holísticas son eficientes para decisiones globales rápidas, pero ofrecen menor diagnosticidad; en cambio, las analíticas favorecen la retroalimentación formativa al hacer visibles las dimensiones del desempeño y orientar acciones de mejora vinculadas a cada criterio (Moskal & Leydens, 2000; Jonsson & Svingby, 2007).

La fiabilidad entre evaluadores es un punto crítico en la calificación de respuestas construidas. La variabilidad entre calificadores aumenta cuando los criterios son ambiguos o los descriptores no son observables, y disminuye cuando se emplean rúbricas analíticas con descriptores claros, se realiza calibración con ejemplos ancla (anchor papers) y se establece un protocolo para resolver discrepancias (Jonsson & Svingby, 2007). La evidencia comparativa reciente en educación indica, además, que los procedimientos analíticos tienden a mejorar el acuerdo entre evaluadores frente a enfoques holísticos, especialmente cuando el entrenamiento incluye discusiones de criterios y ejemplos representativos (Jönsson, 2021).

Desde la perspectiva psicométrica, la fiabilidad entre evaluadores se estima con coeficientes de acuerdo (p. ej., kappa de Cohen) o de consistencia (p. ej., correlaciones intraclase, ICC). En enfoques más integrales, la Teoría de la Generalizabilidad (TG) descompone la varianza en facetas: tarea, criterio y calificador, y permite simular cómo cambian los coeficientes de generalización al variar el número de tareas o de calificadores, lo que aporta una base para decisiones de diseño (más criterios, doble puntaje, mayor entrenamiento) según el umbral de fiabilidad deseado (Brennan, 2003; Tasdelen Teker, 2016). En términos operativos, la literatura converge en tres implicaciones: el diseño de la rúbrica importa (criterios claros y descriptores observables), el proceso importa (calibración y doble corrección con resolución de discrepancias) y la gestión importa (planificación informada por TG). Cuando estas condiciones no se cumplen, incluso rúbricas bienintencionadas pueden producir puntajes inestables o incoherentes entre correctores, con efectos sobre la equidad y la interpretabilidad de los resultados (Moskal & Leydens, 2000; Jonsson & Svingby, 2007).

Uso de LLMs en evaluación y generación de feedback (riesgos y sesgos)

La investigación reciente sobre modelos de lenguaje de gran tamaño (LLMs) en educación superior señala que pueden acelerar y personalizar la retroalimentación en tareas de respuesta construida; sin embargo, advierte límites de calidad, transparencia y control que exigen marcos socio-técnicos prudentes (diseño de criterios, validación de salidas y supervisión docente). Una revisión sistemática en educación superior concluye que los LLMs son útiles como apoyo para generar comentarios específicos y rápidos, siempre que la institución defina parámetros de uso, mantenga intervención humana y documente los procesos para fines de trazabilidad y mejora (Lee & Moore, 2024).

Un riesgo clave es la alucinación: el sistema puede generar respuestas verosímiles, pero erróneas o sin respaldo. La literatura reciente sugiere mitigarse con formatos estructurados (p. ej., JSON), fuentes acotadas, referencias explícitas al texto del estudiante y chequeos automáticos de consistencia antes de emplear los resultados en evaluación (Ji et al., 2023).

Otros riesgos se relacionan con sesgos y gobernanza: los LLMs pueden reproducir patrones indeseados del corpus de entrenamiento y operar como “loros estocásticos”, lo que plantea desafíos de justicia, privacidad y rendición de cuentas. La literatura recomienda evitar delegaciones automáticas de juicio, registrar metadatos (modelo, versión de prompt, parámetros), establecer auditorías por subgrupos y conservar la revisión docente previa a la comunicación de resultados de alto impacto (Bender et al., 2021).

Implicación de diseño. En aplicaciones formativas, el uso responsable de LLMs para generar feedback requiere: (a) alinear la salida a una rúbrica criterial y validarla estructuralmente; (b) referenciar el texto del estudiante en cada justificación; (c) reducir la aleatoriedad en inferencia y documentar parámetros; (d) auditar sesgos y errores con muestreos periódicos; y (e) mantener control humano en la edición y envío del feedback.

Automatización con Google Workspace/Apps Script en contextos educativos

La automatización "ligera" en Google Workspace mediante Apps Script permite articular flujos Forms → Sheets → Gmail/Classroom sin servidores dedicados, lo que la convierte en una vía pragmática para escalar procesos evaluativos y de retroalimentación en educación superior. Estudios recientes señalan, además, que la adopción docente de Workspace se sustenta en percepciones de utilidad y facilidad de uso, y que su integración en docencia y evaluación se beneficia de marcos de gobernanza y capacitación institucional, reforzando la transferibilidad de soluciones basadas en Apps Script. (Ayanwale et al., 2024).

Por su parte, se han descrito aplicaciones académicas que combinan Google Drive/Sheets con Apps Script para automatizar procesos críticos con trazabilidad y control operativo. Un caso en una biblioteca universitaria muestra un sistema de préstamo digital, creado con Drive y Apps Script, que integra registro de usuarios, control de accesos y reportes. El ejemplo ilustra que Apps Script puede respaldar servicios educativos con exigencias de gobernanza y auditoría (Xu et al., 2021). Aunque el dominio es bibliotecario y no de evaluación de aprendizajes, el patrón técnico hojas como base operativa, scripts para orquestación y artefactos normalizados es análogo al requerido en automatizaciones evaluativas.

En la misma línea, un artículo universitario muestra cómo combinar AppSheet y Apps Script para crear aplicaciones de gestión académica dentro de Google Workspace, integrando módulos y reglas de negocio. La idea central es que pueden levantarse sistemas funcionales y escalables con bajo costo, algo especialmente útil para automatizaciones docentes que necesitan persistencia de datos, procesamiento por lotes y mensajería sin salir del entorno institucional (Pham, 2024).

En suma, la evidencia justifica automatizar la evaluación sin salir de Workspace. Sheets actúa como base de datos viva, Apps Script orquesta los procesos y Gmail/Classroom entregan los resultados bajo reglas de gobernanza conocidas. Con estos patrones, es posible montar *pipelines* reproducibles, auditables y que escalen al tamaño de un curso.

La literatura coincide en que el feedback formativo eficaz debe ser oportuno, específico y alineado con metas y criterios, pero también reconoce que su producción es costosa en tareas de respuesta construida (Hattie & Timperley, 2007). A ello se suma un problema de fiabilidad: incluso con buenas intenciones, las decisiones de calificación varían entre correctores cuando los criterios son ambiguos o poco observables; por eso las rúbricas analíticas y la calibración sistemática se recomiendan frente a enfoques holísticos para sostener la consistencia (Jonsson & Svingby, 2007). En paralelo, los avances recientes en modelos de lenguaje ofrecen atajos para sintetizar y personalizar devoluciones, pero arrastran riesgos de alucinación, opacidad y sesgo que impiden su uso directo sin controles (Bender et al., 2021).

El vacío se sitúa en la intersección de estas tres líneas: abundan principios y recomendaciones (feedback oportuno y criterial; rúbricas analíticas; supervisión humana del juicio), pero faltan descripciones operativas, replicables y auditables de pipelines end-to-end que, dentro de entornos institucionales ampliamente disponibles (Google Workspace), integren: (a) recolección de respuestas abiertas, (b) calificación por criterios asistida por IA con validación estructural de salidas, y (c) comunicación automática de retroalimentación de calidad a escala, todo ello bajo cuotas reales de ejecución y con trazabilidad (metadatos de modelo, versión de rúbrica y prompt). La mayor parte de trabajos reporta beneficios o dilemas en piezas aisladas (p. ej., uso

de rúbricas; potencial de la IA; experiencias de adopción tecnológica), pero ofrece poca guía técnica para el docente–desarrollador que debe convertir esos principios en un flujo reproducible y sostenible en el aula.

La propuesta documentada en este artículo aborda ese vacío con una solución socio-técnica que ancla la automatización en los fundamentos pedagógicos y los hace operables en un stack de alta disponibilidad:

En lugar de delegar el juicio, formaliza la rúbrica en criterios y niveles (analítico) y la convierte en instrucciones ejecutables para la IA; la salida se restringe a JSON validado contra un esquema, lo que reduce la ambigüedad y favorece la auditoría (Jonsson & Svingby, 2007).

Para responder a la exigencia de oportunidad, orquesta un pipeline Forms → Sheets → Apps Script → correo HTML capaz de generar y enviar devoluciones minutos después del envío, preservando la estructura feed up / feed back / feed forward propia del buen feedback (Hattie & Timperley, 2007).

Frente a los riesgos de los LLMs, incorpora controles de calidad: prompts explícitos con referencias al texto del estudiante, baja aleatoriedad en inferencia, validación de formato y revisión docente previa al envío; además, registra metadatos (modelo, parámetros, versión de rúbrica/prompt) para facilitar la rendición de cuentas y el análisis posterior (Bender et al., 2021).

Finalmente, resuelve la viabilidad operativa a nivel institucional mediante procesamiento por lotes y persistencia incremental que respetan las cuotas del entorno, asegurando que el flujo cierre a escala de curso sin infraestructura adicional.

Con ello, el trabajo traslada recomendaciones dispersas en la literatura a un procedimiento implementable y replicable, ofreciendo evidencias de rendimiento y artefactos verificables (hojas de resultados y correos HTML) que pueden inspirar, adaptar y auditar en otras asignaturas y contextos.

La literatura coincide en que el feedback formativo eficaz debe ser oportuno, específico y alineado con metas y criterios, pero también reconoce que su producción es costosa en tareas de respuesta construida (Hattie & Timperley, 2007). A ello se suma un problema de fiabilidad: incluso con buenas intenciones, las decisiones de calificación varían entre correctores cuando los criterios son ambiguos o poco observables; por eso las rúbricas analíticas y la calibración sistemática se recomiendan frente a enfoques holísticos para sostener la consistencia (Jonsson & Svingby, 2007). En paralelo, los avances recientes en modelos de lenguaje ofrecen atajos para sintetizar y personalizar devoluciones, pero arrastran riesgos de alucinación, opacidad y sesgo que impiden su uso directo sin controles (Bender et al., 2021).

El vacío se sitúa en la intersección de estas tres líneas: abundan principios y recomendaciones (feedback oportuno y criterial; rúbricas analíticas; supervisión humana del juicio), pero faltan descripciones operativas, replicables y auditables de pipelines end-to-end que, dentro de entornos institucionales ampliamente disponibles (Google Workspace), integren: (a) recolección de respuestas abiertas, (b) calificación por criterios asistida por IA con validación estructural de salidas, y (c) comunicación automática de retroalimentación de calidad a escala, todo ello bajo cuotas reales de ejecución y con trazabilidad (metadatos de modelo, versión de rúbrica y prompt). La mayor parte de trabajos reporta beneficios o dilemas en piezas aisladas (p. ej., uso de rúbricas; potencial de la IA; experiencias de adopción tecnológica), pero ofrece poca guía técnica para el docente–desarrollador que debe convertir esos principios en un flujo reproducible y sostenible en el aula.

La propuesta documentada en este artículo aborda ese vacío con una solución socio-técnica que ancla la automatización en los fundamentos pedagógicos y los hace operables en un stack de alta disponibilidad:

En lugar de delegar el juicio, formaliza la rúbrica en criterios y niveles (analítico) y la convierte en instrucciones ejecutables para la IA; la salida se restringe a JSON validado contra un esquema, lo que reduce la ambigüedad y favorece la auditoría (Jonsson & Svingby, 2007).

Para responder a la exigencia de oportunidad, orquesta un pipeline Forms → Sheets → Apps Script → correo HTML capaz de generar y enviar devoluciones minutos después del envío, preservando la estructura feed up / feed back / feed forward propia del buen feedback (Hattie & Timperley, 2007).

Frente a los riesgos de los LLMs, incorpora controles de calidad: prompts explícitos con referencias al texto del estudiante, baja aleatoriedad en inferencia, validación de formato y revisión docente previa al envío; además, registra metadatos (modelo, parámetros, versión de rúbrica/prompt) para facilitar la rendición de cuentas y el análisis posterior (Bender et al., 2021).

Finalmente, resuelve la viabilidad operativa a nivel institucional mediante procesamiento por lotes y persistencia incremental que respetan las cuotas del entorno, asegurando que el flujo cierre a escala de curso sin infraestructura adicional.

Con ello, el trabajo traslada recomendaciones dispersas en la literatura a un procedimiento implementable y replicable, ofreciendo evidencias de rendimiento y artefactos verificables (hojas de resultados y correos HTML) que pueden inspirar, adaptar y auditar en otras asignaturas y contextos.

Contexto y participantes

El estudio se llevó a cabo en la Universidad Estatal del Sur de Manabí (UNESUM), en la asignatura Informática Aplicada a la Educación, con seis paralelos de tercer nivel (3A–3F) que, en condiciones regulares, cuentan con ≈35 estudiantes por paralelo. La intervención incluyó a quienes pudieron conectarse a la hoja de cálculo en el momento de la actividad (muestreo por conveniencia): 3A = 20, 3B = 16, 3C = 23, 3D = 19, 3E = 16 y 3F = 19 (n = 113). No se realizaron ensayos controlados; todas las observaciones provienen del funcionamiento regular del sistema durante sesiones de clase.

Materiales

Entorno tecnológico: Google Forms (recolección), Google Sheets (repositorio y tablero de control), Google Apps Script (orquestación), Gmail de Google Workspace (comunicación) y un conector configurable hacia una API de IA generativa (OpenAI o Google) por medio de un adaptador común.

Estructura lógica de datos en Sheets: pestaña de respuestas (timestamp, ID, P1, P2…), pestaña de preguntas (ID, enunciado, peso), una pestaña de rúbrica por pregunta (criterios, descriptores 1–4 y pesos) y una pestaña de configuración (tamaño de lote, modelo de IA, umbrales, plantilla de correo). Las salidas se registran en las hojas “Resultado P1”, “Resultado P2”, etc., y en una síntesis por estudiante.

Diseño técnico y arquitectura

El flujo Forms → Sheets → Apps Script → IA → Sheets → Gmail se implementó en módulos:

a) Orquestación. La función `evaluarRespuestas()` detecta nuevos envíos, carga preguntas y rúbricas, empaqueta solicitudes por estudiante y por pregunta, y ejecuta procesamiento por lotes (p. ej., 5 estudiantes por ciclo) para respetar cuotas y tiempos máximos de Apps Script, con persistencia del progreso para tolerar interrupciones.
b) Prompting y validación. Se construye dinámicamente un prompt con rol de evaluador, escala 1–4 por criterio y requisito de salida en JSON con campos de puntaje total, porcentaje, tabla de criterios con justificaciones, fortalezas, oportunidades y recomendación final. Un validador interno comprueba el esquema; si la estructura falla, se reintenta con un prompt de corrección mínima (sin recalificar).
c) Capa agnóstica de proveedor. Un servicio LLMService unifica llamadas a OpenAI o Google según la pestaña de configuración (endpoint, clave, temperatura y timeouts), sin afectar el resto del código.
d) Escritura y formato. La función `escribirResultados_()` genera las hojas por pregunta y la síntesis por estudiante, con formato condicional tipo "semáforo", filtros y celdas protegidas.
e) Comunicación. La función `enviarCorreos_()` compone un correo HTML con encabezado (docencia/asignatura), puntuación total y porcentaje destacados, transcripción de las respuestas y tabla por criterios con justificaciones, además de un mensaje final de ánimo; se registra la bitácora de envío.

Procedimiento

- Preparación: definición de preguntas y rúbricas analíticas (criterios, descriptores y pesos) y carga de parámetros en "Config".
- Recolección: aplicación del formulario en clase o de forma asincrónica y verificación del volcado a "Respuestas…".
- Evaluación automática: ejecución de `evaluarRespuestas()`; el script procesa por lotes, invoca la IA, valida el JSON y persiste resultados.
- Revisión docente (opcional): inspección de casos límite (respuestas muy breves o fuera de tópico) y ajustes puntuales.
- Comunicación: ejecución de `enviarCorreos_()` y registro en bitácora.
- Cierre: consolidación de métricas (tiempos, reintentos y finalización sin errores) para análisis descriptivo.

Variables e indicadores

Rendimiento operativo: tiempo de evaluación (`T_eval`), medido desde el disparo de `evaluarRespuestas()` hasta la escritura final en Sheets; tiempo de envío (`T_mail`) para completar los correos; tasa de reintentos por formato JSON; y finalización sin interrupciones por límites de Apps Script/Gmail.

Calidad del feedback: en muestra estratificada, especificidad por criterio (escala 1–4), coherencia entre la justificación y la evidencia en el texto del estudiante, transparencia (inclusión de la transcripción) y accionabilidad (recomendación vinculada a la rúbrica).

Eficiencia docente: estimación del tiempo manual (min/respuesta + min/correo) frente al tiempo automatizado observado.

Consideraciones éticas y de privacidad

El procesamiento se efectuó en cuentas institucionales con accesos restringidos al equipo docente, sin recopilar datos sensibles más allá de la identificación académica; las claves y tokens se gestionaron mediante PropertiesService (nunca en celdas); se informó previamente al estudiantado que la retroalimentación es asistida por IA y revisada por el docente, con disponibilidad de solicitud de revisión ante discrepancias; y el uso fue estrictamente formativo, sin efectos sancionatorios.

Resultados y discusión

Los tiempos de ejecución se obtuvieron a partir de las marcas de tiempo registradas en la bitácora de Apps Script y en los registros de ejecución durante el uso real del sistema en los seis paralelos. Se definió `T_eval` como el intervalo entre el disparo de `evaluarRespuestas()` y la escritura completa de resultados en Sheets, medido por paralelo y, cuando correspondía, por lote; y `T_mail` como el intervalo total de `enviarCorreos_()` por paralelo. Los resultados se describieron mediante estadística descriptiva (medianas, rangos y proporciones); la calidad del feedback se caracterizó con una lista de cotejo que consideró claridad, especificidad por criterio, coherencia con la evidencia textual y accionabilidad. La eficiencia docente se estimó comparando el tiempo observado en la operación regular con una línea base manual conservadora (minutos por respuesta + minutos por envío), sin experimentación controlada. Para cada paralelo se registraron `T_eval` y `T_mail`; la síntesis de los valores observados se presenta en la Tabla 1.

Tabla 1. Tiempos de ejecución por paralelo.

Paralelo	n estudiantes	T_eval	T_mail
3A	20	1 min 46 s	1 s / mail
3B	16	1 min 42 s	1 s / mail
3C	23	2 min 01 s	1 s / mail
3D	19	1 min 49 s	1 s / mail
3E	16	1 min 44 s	1 s / mail
3F	19	1 min 45 s	1 s / mail

En seis paralelos (n = 113), el sistema completó la etapa de evaluación (`T_eval`) entre 1:42 y 2:01 por paralelo (media ≈ 1:48, DE ≈ 7,0 s). Sumado en bloque, el procesamiento de los 113 estudiantes tomó 10 min 47 s (647 s), lo que equivale a una tasa de 10,5 estudiantes por minuto y un tiempo medio de ≈5,7 s por estudiante para calcular y volcar los resultados en las hojas de cálculo. La comunicación automática de la retroalimentación personalizada registró ≈1 s por correo, de modo que el envío total a la cohorte demandó 1 min 53 s adicionales. En conjunto, el ciclo evaluación + comunicación se completó en ≈12 min 40 s de tiempo de reloj.

Para dimensionar el impacto operativo, se comparó este desempeño con una línea base manual conservadora. Si un docente dedicara 3–6 minutos a calificar una respuesta abierta con rúbrica y 30–60 segundos a preparar/enviar el correo por estudiante, la corrección y comunicación de 113 informes requeriría ≈6,6 a 13,2 horas. Frente a ese rango, el proceso automatizado de ≈12,7 minutos supone una reducción del tiempo de trabajo de ≈96,8% (escenario conservador bajo) a ≈98,4% (escenario alto), manteniendo además la estructura criterial del feedback (puntaje por criterio, justificación y recomendación final).

Los resultados muestran que el sistema cumplió el objetivo de oportunidad y escala de la retroalimentación: la etapa de evaluación (`T_eval`) se completó por paralelo entre 1:42 y 2:01 (media ≈ 1:48) y la comunicación registró ≈1 s por correo, de modo que el ciclo evaluación + comunicación para 113 estudiantes se cerró en ≈12:40 de tiempo de reloj. Traducido a eficiencia, frente a una línea base manual conservadora (3–6 min/ respuesta abierta + 30–60 s por correo), la automatización supuso una reducción de ≈96,8% a ≈98,4% del tiempo docente.

Respecto de la calidad del artefacto de feedback, el sistema entregó devoluciones criteriales (escala 1–4 por criterio) con justificaciones y recomendaciones accionables, además de la transcripción de la respuesta del estudiante. Esta estructura hace visible la correspondencia entre evidencia y juicio y favorece el uso pedagógico de la devolución. La capa de validación estructural (salida obligatoria en JSON y verificación de esquema) y el procesamiento por lotes contribuyeron a la estabilidad operativa; si bien no se midió fiabilidad entre evaluadores porque no se comparó con una segunda corrección humana paralela, la estandarización del formato reduce fuentes habituales de variación (p. ej., omisiones o cambios de criterio al redactar comentarios).

Los hallazgos son coherentes con la evidencia que vincula la oportunidad del feedback con su efectividad formativa: devolver comentarios minutos después de la actividad maximiza su utilidad para regular el aprendizaje entre intentos. Frente a descripciones previas centradas en piezas aisladas (p. ej., rúbricas o generación de comentarios con IA), este caso aporta un pipeline operativo completo dentro de Google Workspace, con trazabilidad (bitácora, metadatos de modelo y versión de prompt), control docente y artefactos verificables (hojas de resultados y correos HTML).

Pedagógicamente, el sistema desplaza el foco del docente desde la tarea repetitiva de redactar comentarios hacia funciones de diseño y moderación: elaborar rúbricas analíticas claras, ajustar prompts para alinear criterios y monitorear casos límite. La inmediatez y la estructura criterial de la devolución favorecen prácticas de autorregulación (el estudiante entiende qué mejorar y cómo), y la transparencia (transcripción de la respuesta) permite auditoría y autoexplicación.

El diseño es escalable por tres razones: (1) se apoya en infraestructura existente (Forms/Sheets/Apps Script/Gmail); (2) implementa loteo y control de cuotas; y (3) separa configuración (modelo, tamaño de lote, plantillas) del código. Para crecer a cohortes mayores, bastan ajustes de TAMANO_LOTE, activación de triggers temporizados y, si es necesario, escalonar el envío de correos para respetar límites de Workspace.

En adopción institucional, conviene definir roles: un docente responsable (o equipo) que mantenga rúbricas y prompts; un técnico-pedagógico que supervise parámetros y bitácoras; y el docente de aula que interpreta y, si procede, ajusta devoluciones en casos atípicos. El entrenamiento debe cubrir:

a) Diseño de rúbricas analíticas con descriptores observables;
b) principios básicos de prompting criterial;
c) lectura de bitácoras y resolución de incidencias; y
d) pautas de privacidad y comunicación responsable.

El mantenimiento incluye rotación y resguardo de claves (PropertiesService), revisión periódica de cuotas, actualización de plantillas de correo y verificación de cambios en APIs/modelos.

Conclusiones

La implementación de un sistema de evaluación formativa asistido por inteligencia artificial demostró ser una solución viable para automatizar la retroalimentación en preguntas abiertas a escala institucional. Los resultados evidencian una significativa reducción del tiempo docente dedicado a tareas evaluativas repetitivas, manteniendo la calidad pedagógica del feedback y preservando la estructura criterial necesaria para el aprendizaje formativo.

El enfoque socio-técnico adoptado logró equilibrar la eficiencia operativa con la calidad pedagógica, generando retroalimentación estructurada, específica por criterio y accionable que favorece los procesos de autorregulación del estudiantado. La validación estructural de las salidas de IA y la estandarización de formatos contribuyeron a reducir la variabilidad típica de los procesos de evaluación manual, mejorando la consistencia y transparencia del feedback.

La arquitectura desarrollada dentro del ecosistema Google Workspace ofrece un marco replicable y escalable que puede ser adoptado por otras instituciones educativas sin requerir infraestructura adicional. Este modelo redefine el rol docente hacia funciones de mayor valor pedagógico, centrándose en el diseño de criterios de evaluación y la supervisión del proceso automatizado, mientras mantiene el control y la responsabilidad profesional sobre las decisiones evaluativas.

Referencias

Ayanwale, M. A., Molefi, R. R., & Liapeng, S. (2024). Unlocking educational frontiers: Exploring higher educators' adoption of Google Workspace technology tools for teaching and assessment in Lesotho dynamic landscape. Heliyon, 10(9), e30049. https://doi.org/10.1016/j.heliyon.2024.e30049

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (pp. 610–623). https://doi.org/10.1145/3442188.3445922

Black, P., & Wiliam, D. (1998). Inside the black box: Raising standards through classroom assessment. Phi Delta Kappan, 80(2), 139–148

Bond, M., Khosravi, H., De Laat, M., Bergdahl, N., Negrea, V., Oxley, E., Pham, P., Chong, S. W., & Siemens, G. (2024). A meta systematic review of artificial intelligence in higher education: A call for increased ethics, collaboration, and rigour. International Journal of Educational Technology in Higher Education, 21, 4. https://doi.org/10.1186/s41239-023-00436-z

Brennan, R. L. (2003). Coefficients and indices in generalizability theory (CASMA Research Report No. 1). University of Iowa. https://education.uiowa.edu/sites/education.uiowa.edu/files/2021-11/casma-research-report-1.pdf

Carless, D., & Boud, D. (2018). The development of student feedback literacy: Enabling uptake of feedback. Assessment & Evaluation in Higher Education, 43(8), 1315–1325. https://doi.org/10.1080/02602938.2018.1463354

Hattie, J., & Timperley, H. (2007). The power of feedback. Review of Educational Research, 77(1), 81–112. https://doi.org/10.3102/003465430298487

Hift, R. J. (2014). Should essays and other "open-ended"-type questions retain a place in written summative assessment in clinical medicine? BMC Medical Education, 14, 249. https://doi.org/10.1186/s12909-014-0249-2

Huang, J., & Whipple, P. B. (2023). Rater variability and reliability of constructed response questions in New York state high-stakes tests of English language arts and mathematics: Implications for educational assessment policy. Humanities and Social Sciences Communications, 10, 860. https://doi.org/10.1057/s41599-023-02385-4

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y., Chen, D., Dai, W., Chan, H. S., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. ACM Computing Surveys, 55(12), Article 248. https://doi.org/10.1145/3571730

Jönsson, A. (2021). Analytic or holistic? A study about how to increase the agreement between teachers' assessments. Assessment in Education: Principles, Policy & Practice, 28(4), 384–401. https://doi.org/10.1080/0969594X.2021.1884041

Jonsson, A., & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. Educational Research Review, 2(2), 130–144. https://doi.org/10.1016/j.edurev.2007.05.002

Lee, S. S., & Moore, R. L. (2024). Harnessing generative AI (GenAI) for automated feedback in higher education: A systematic review. Online Learning, 28(3), 82–104. https://doi.org/10.24059/olj.v28i3.4593

Moskal, B. M., & Leydens, J. A. (2000). Scoring rubric development: Validity and reliability. Practical Assessment, Research, and Evaluation, 7(1), 1–7. https://doi.org/10.7275/q7rm-gg74

Nicol, D. J., & Macfarlane-Dick, D. (2006). Formative assessment and self-regulated learning: A model and seven principles of good feedback practice. Studies in Higher Education, 31(2), 199–218. https://doi.org/10.1080/03075070600572090

Pham, N. S. (2024). Using AppSheet and Apps Script to develop management and implementation applications for university education programs. VNU Journal of Science: Education Research, 40(1). https://doi.org/10.25073/2588-1159/vnuer.47362

Shute, V. J. (2008). Focus on formative feedback. Review of Educational Research, 78(1), 153–189. https://doi.org/10.3102/0034654307313795

Tasdelen Teker, G. (2016). Using generalizability theory to investigate the reliability of performance assessments. International Journal of Assessment Tools in Education, 3(2), 166–178. https://files.eric.ed.gov/fulltext/ED585251.pdf

Williams, A. (2024). Delivering effective student feedback in higher education: An evaluation of the challenges and best practice. International Journal of Research in Education and Science, 10(2), 473–501. https://doi.org/10.46328/ijres.3404

Wisniewski, B., Zierer, K., & Hattie, J. (2020). The power of feedback revisited: A meta-analysis of educational feedback research. Frontiers in Psychology, 10, 3087. https://doi.org/10.3389/fpsyg.2019.03087

Xu, Q., Lin, L., & Wu, X. (2021). Implementing controlled digital lending with Google Drive and Apps Script: A case study at the NYU Shanghai Library. International Journal of Librarianship, 6(1), 37–54. https://doi.org/10.23974/ijol.2021.vol6.1.193