UNIVERSIDAD DE CONCEPCIÓN
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS

Optimización de la selección de Grandes Modelos de Lenguaje para la generación de Feedback Personalizado en la asignatura de Álgebra I, de la Universidad de Concepción.

Ivette Henríquez Carter

Tesis presentada a la Facultad de Ciencias Físicas y Matemáticas de la Universidad de Concepción para optar al título profesional de Ingeniera Civil Matemática

Julio 2025
Concepción, Chile

Profesor Guía: Alejandra Maldonado Trapp

Profesores Co-guía: Claudio Bustos
Carlos Navarrete
Mónica Selva

A Gaspar

AGRADECIMIENTOS

Quiero dedicar y agradecer este logro, en primer lugar, a mi familia. Gracias por ser mi pilar fundamental, perdón por las ausencias y el desaparecer, pero bueno, es lo que tocó por tratar de sacar esta carrera. Gracias, mamá, por alentarme a seguir adelante cuando el camino se volvía difícil. Gracias, Gasparcito, por ser lo mejor del mundo, mi regalo más preciado y por quien quiero lograr lo que sea posible para que pueda seguir creciendo feliz.

Seguimos con mi pololo, Brayan. Gracias por acompañarme y estar presente en los momentos más tensos del año. Gracias por ser mi compañero y apoyarme de manera incondicional, te amo.

A mi mejor amiga, Scarllet, infinitas gracias por todo el cariño en cada una de las etapas de mi vida universitaria. Comenzamos juntas en 2018, seguimos juntas hoy en 2025, y espero que sean muchos años más. A mi mejor amigo, Jere, con quien estuvimos a la par procesando este camino, vaya que fue duro, pero lo logramos. Éxito en todo y espero poder acompañarte en tus futuras metas.

También gracias a mis amigas del colegio: Franci, Katty, Estefany, Camila, Gabi y Cancino. A mis amigues de la U: Cata, Fran, Isi, Vicho, Mirko, Pato, José Benjamín, Martín y Mateo. Gracias, Dani, por tu compañía y por cada una de las salidas. Gracias, Vale, por cada uno de tus consejos entregados a lo largo de lo que fue difusión.

Expreso mi sincera gratitud a mis profesores y a los profesionales que guiaron mi formación. Gracias por compartir sus conocimientos y enseñanzas de vida. Gracias, profe Ale, por ser una mujer maravillosa con la que pude trabajar y aprender infinitamente. Gracias, profe Moni, por haber sido la mejor jefa de carrera, le deseo el mejor de los éxitos y espero nos podamos vernos pronto. Nataly, gracias por todo, espero poder seguir aprendiendo mucho más y que sigamos trabajando juntas de alguna u otra manera.

Finalmente, agradezco el apoyo institucional y financiero que hizo posible la realización de esta tesis. En particular, al proyecto Fondecyt 11241189 y al proyecto UCO21102 Interdisciplina.

Resumen

En esta tesis se aborda el desafío de seleccionar Grandes Modelos de Lenguaje (LLMs) para la educación, buscando un balance óptimo entre calidad pedagógica y sostenibilidad económica. Se presenta un modelo matemático de optimización para guiar la configuración de LLMs en la generación de Feedback Personalizado para cursos de matemática universitaria. Se realizó un experimento aleatorio con un diseño factorial estratificado en bloques, evaluando cuatro factores clave:

- Modelo para el Análisis Técnico.
- Temperatura para el Análisis Técnico.
- Modelo para el Feedback Personalizado.
- Temperatura para el Feedback Personalizado.

A partir de 9 evaluaciones de respuestas reales de la asignatura de Álgebra I del primer semestre del año 2024, considerando 81 configuraciones derivadas del diseño factorial estratificado en bloques, se comparó el rendimiento de tres familias de modelos líderes: **Gpt-4o-2024-08-06**, **LLaMA-3.3-70b-versatile** y **Gemini-2.0-flash-001**.
Los resultados del análisis de varianza y un índice de rendimiento compuesto demuestran que Gemini-2.0-flash-001 ofrece el mejor equilibrio entre riqueza pedagógica y eficiencia económica. Este marco cuantitativo proporciona una guía objetiva, rigurosa y replicable para la adopción informada y sostenible de LLMs en la enseñanza de las matemáticas.

Índice general

Índice de tablas

Índice de figuras

Capítulo 1

Introducción

La entrega de retroalimentación de alta calidad y oportuna es esencial para el aprendizaje efectivo en asignaturas complejas como la matemática universitaria (Söderström and Palm, 2024). Tradicionalmente, esta tarea recae en la o el docente, lo que limita la frecuencia y personalización de los comentarios debido a la carga de trabajo. Los Grandes Modelos de Lenguaje (LLMs por sus siglas en inglés) han mostrado potencial para automatizar y enriquecer este proceso, generando explicaciones, correcciones y sugerencias adaptadas a cada estudiante (Wan and Chen, 2025).

Sin embargo, el utilizar LLMs en estos procesos tiene dos retos importantes que enfrentar. La primera es la diversidad de respuestas que se deben analizar, por ejemplo, respuestas muy largas que presentan mayormente errores o una nula respuesta a la pregunta. La segunda es el costo asociado a tomar y generar el Feedback Personalizado. Por estas razones surge la necesidad de un enfoque cuantitativo que equilibre la calidad educativa, que se define en términos de precisión en el análisis técnico y riqueza del feedback, con los costos asociados a estas salidas.

El objetivo de esta investigación es diseñar un modelo matemático que permita identificar la configuración óptima de LLMs para generar Feedback Personalizado en cursos de matemática. Para ello, se compararon tres modelos que provienen de distintas familias, evaluados con datos reales a una pregunta del curso de Álgebra I que se imparte para las carreras de Ingeniería.

1.1. Contexto y Relevancia

La capacidad de proporcionar retroalimentación en tiempo real en el contexto educativo actual es esencial para mejorar el rendimiento estudiantil y la comprensión de conceptos complejos. Los LLMs han demostrado un potencial significativo en diversas aplicaciones educativas (Fernández et al., 2024). Investigaciones indican que estos modelos pueden ser utilizados para generar explicaciones y autoevaluaciones, facilitando el aprendizaje de temas difíciles (Aldazharova et al., 2024; Boscardin et al., 2024). Sin embargo, es crucial reconocer que, aunque muestran una efectividad educativa notable en ciertas áreas, también presentan limitaciones, especialmente en la comprensión conceptual de problemas complejos.

La selección de un modelo adecuado va más allá de su eficacia académica; factores como el costo y la eficiencia operativa son igualmente importantes para su implementación sostenible en el ámbito universitario (Boscardin et al., 2024). Se ha destacado que la capacitación adicional de LLMs puede mejorar su precisión y utilidad en contextos educativos (Kaneda et al., 2023; Nori et al., 2023), pero también se requieren consideraciones sobre el entrenamiento y la adaptación a normativas específicas para maximizar su efectividad (Miao et al., 2024). La integración de herramientas en el aula no solo podría optimizar el proceso de enseñanza-aprendizaje, sino también servir como un valioso apoyo para las y los educadores, aliviando parte de su carga de trabajo en la evaluación y retroalimentación (Rath, 2025; Yamtinah et al., 2024). Por lo tanto, se sugiere que los educadores evalúen cuidadosamente ambos aspectos para seleccionar la herramienta que mejor se adapte a sus necesidades educativas y a las de sus estudiantes.

1.2. Problema

A pesar de los avances significativos en la tecnología de los LLMs, la selección adecuada de un modelo que logre un equilibrio eficiente entre coste y rendimiento sigue siendo un desafío crítico, en particular en contextos educativos donde los recursos son frecuentemente limitados. La literatura existente indica que, si bien los LLMs pueden ofrecer retroalimentación educativa de alta calidad, su

implementación en instituciones con presupuestos restringidos no siempre es viable (Quttainah et al., 2024; Yan et al., 2024). Un análisis de los costos asociados con el uso de LLMs revela que solo las grandes empresas tecnológicas pueden permitirse desarrollar y mantener estos modelos a gran escala, lo que genera preocupaciones sobre la equidad en el acceso a la tecnología educativa (Tamayo et al., 2024).

Existen marcos que destacan la importancia de considerar la accesibilidad y la transparencia al seleccionar un LLM para usos educativos, lo que ayuda a las instituciones a tomar decisiones más informadas (Lee et al., 2024). Las investigaciones sugieren que la incorporación de LLMs debe ir acompañada de formación y orientación para maximizar su eficacia mientras se mitigan los riesgos asociados, como la dependencia excesiva en estas herramientas (Pavlova et al., 2024).

Por lo tanto, este trabajo propone un marco de referencia que no solo evalúa la efectividad de los LLMs en términos de calidad del feedback educativo, sino que también considera los aspectos económicos y técnicos que determinarán su viabilidad a largo plazo en el ámbito universitario. La necesidad de herramientas de retroalimentación automáticas y personalizadas sigue siendo apremiante, y la elección adecuada de un LLM puede facilitar una experiencia educativa más diversa, siempre que se implementen las consideraciones mencionadas.

1.3. Objetivos

- **Objetivo General:** Desarrollar un modelo matemático que optimice la selección y configuración de los LLMs para generar Feedback Personalizado, considerando la eficiencia económica y técnica en la asignatura de Álgebra I para ingeniería.
- **Objetivos Específicos:**
  - Definir métricas de calidad a las salidas de los LLMs para el análisis técnico y el Feedback Personalizado.
  - Diseñar el proceso general para la generación de Feedback Personalizado.
  - Evaluar el rendimiento y costo de LLMs en términos de las métricas de calidad establecidas, implementando algunas etapas del proceso general

de Feedback Personalizado.

Capítulo 2

Marco Teórico

2.1. Fundamentos de la Inteligencia Artificial

La Inteligencia Artificial (IA) es un campo de la informática que tiene como objetivo crear máquinas capaces de comportarse de manera inteligente (**?**). Dentro de la IA, el *Machine Learning* (ML) es un subconjunto que se enfoca en desarrollar algoritmos que permitan a las computadoras aprender y hacer predicciones o tomar decisiones basadas en datos (Bengio et al., 2013). El *Deep Learning* (DL) es un tipo específico de ML que involucra redes neuronales con múltiples capas para aprender representaciones de datos con múltiples niveles de abstracción (LeCun et al., 2015). El procesamiento del lenguaje natural (NLP) es una rama de la IA que se enfoca en permitir que las computadoras comprendan, interpreten y generen lenguaje humano (Guo et al., 2017).

El DL, en particular las redes neuronales profundas, ha mostrado un éxito significativo en varias aplicaciones, incluyendo el reconocimiento de patrones y concursos de aprendizaje automático (Schmidhuber, 2015). Permite la extracción de información pertinente a través de la recombinación jerárquica de características para aprender patrones representados en los datos (Mater and Coote, 2019). En el contexto de NLP, las técnicas de aprendizaje profundo aprovechan múltiples capas de nodos de procesamiento no lineal para aprender automáticamente representaciones del lenguaje para realizar tareas complejas (Guo et al., 2017). Además, el éxito de los algoritmos de ML depende en gran medida de la representación de los datos, ya que diferentes representaciones pueden enredar o

revelar diferentes factores explicativos dentro de los datos (Bengio et al., 2013).

2.2. Inteligencia Artificial Generativa

La Inteligencia Artificial Generativa (IAG) es un subcampo específico que utiliza algoritmos para crear nuevos contenidos, como texto, imágenes, música, y más, a partir de patrones aprendidos de datos existentes.
La IAG se centra en la creación y generación de contenido original, utilizando modelos como *Generative Pre-trained Transformers* para crear texto coherente, así como Redes Generativas Adversariales (GANs) para producir imágenes. Esta suele depender de grandes conjuntos de datos y modelos complejos que permiten la creación en lugar de solo clasificación o predicción. Esto implica un nivel más alto de complejidad en el procesamiento y entrenamiento, especialmente para generar contenido de alta calidad que sea coherente y relevante (Valencia Mendoza et al., 2024).

2.3. Definiciones claves para LLMs

Para entender mejor los mecanismos y configuraciones en los LLMs, es importante revisar los siguientes términos y conceptos que son esenciales en el funcionamiento de estos modelos:

2.3.1. Tokens

Los tokens son las unidades básicas de texto que estos modelos procesan durante su funcionamiento. La tokenización, es decir, el proceso de dividir el texto en estas unidades más pequeñas, permite a los LLMs llevar a cabo tareas de procesamiento de lenguaje natural de manera efectiva (Gutierrez-Vasques and Mijangos De La Cruz, 2024). Un token puede corresponder a una palabra completa, un fragmento de palabra o incluso un carácter, lo que facilita que los modelos manejen distintos idiomas y estructuras lingüísticas. Por ejemplo, en modelos basados en WordPiece la palabra “matemática” podría convertirse en los subtokens “mate”, “má” y “tica”, mientras que otros emplean *Byte-Pair Encoding* (BPE), que es un algoritmo de compresión de datos que también se utiliza en el NLP para la tokenización de subpalabras, o para optimizar su vocabulario y su

capacidad de generalización *SentencePiece*. Además, la cuantificación de tokens afecta directamente al rendimiento de los LLMs, cada modelo impone un límite máximo de tokens por entrada y salida, y si este se supera, la respuesta suele cortarse o volverse incoherente.

2.3.2. Temperatura

La temperatura es un parámetro que controla la aleatoriedad de la generación de texto. La temperatura se utiliza en el algoritmo de muestreo que determina cómo se seleccionan las palabras o frases como respuestas generadas por el modelo. Generalmente, la temperatura puede ajustarse para influir en el equilibrio entre la diversidad y la coherencia de las respuestas generadas.

Una temperatura más baja, por ejemplo, cercana a 0, genera respuestas más predecibles y coherentes, lo que significa que el modelo optará por las palabras más probables y seguirá patrones conocidos en los datos de entrenamiento. Esto es útil en situaciones donde se requiere precisión y estabilidad, como la generación de textos técnicos y formales (Windisch et al., 2024). En contraste, al aumentar la temperatura, por ejemplo, a 1,0 o más, el modelo se vuelve más creativo y variado, generando respuestas que pueden ser menos predecibles y más diversas.

Investigaciones han demostrado cómo el ajuste de la temperatura puede impactar significativamente el rendimiento de los LLMs en diversas tareas. Por ejemplo, un estudio indicó que al aumentar la temperatura, el número de diagnósticos únicos en aplicaciones médicas aumentó de 18, con temperatura 0, a 105, con temperatura 1,0, lo que representa un incremento del 583 % en la diversidad de diagnósticos (Jarrett et al., 2025).

2.3.3. Prompts

Un prompt se define como el texto o la instrucción que se proporciona al modelo para guiar su generación de respuestas. Los prompts son esenciales porque establecen el contexto en el que el modelo debe operar y determinan cómo se interpretarán sus respuestas. La formulación del prompt puede influir significativamente en la calidad y relevancia del output generado, siendo fundamental para obtener resultados coherentes y útiles (Fernandes, 2023).

Los prompts pueden variar en complejidad, desde preguntas simples hasta

descripciones detalladas de tareas específicas. Por ejemplo, un prompt sencillo podría ser "¿Qué es un número primo?", mientras que uno más complejo podría ser "Demuestra, paso a paso, el teorema fundamental del Álgebra y discute dos de sus apliacaciones". Esta variabilidad en la formulación permite a los educadores y desarrolladores personalizar las interacciones con el modelo para ajustarse a sus necesidades específicas, sean estas educativas o creativas (Beckford, 2023).

La calidad del prompt también se relaciona directamente con el rendimiento del modelo. Investigaciones sugieren que prompts bien diseñados pueden mejorar la coherencia y relevancia de las respuestas generadas (Gutierrez-Vasques and Mijangos De La Cruz, 2024).

2.3.4. Roles: Sistema, Usuario y Asistente

En los modelos de chat, un "rol" define la función o identidad asignada a cada participante y condiciona su comportamiento en la conversación. Los roles principales son:

- **Mensaje del Sistema:** Es una instrucción interna que configura el comportamiento del modelo (asistente). No proviene de un humano, sino de la propia plataforma que invoca el LLM. Suele especificar el tono, la personalidad o las reglas de interacción.
- **Usuario / Mensaje Humano:** Corresponde a la entrada real del usuario; estas pueden ser las preguntas, solicitudes o comentarios que impulsan la conversación. El asistente analiza estos mensajes para determinar qué información o ayuda proporcionar.
- **Mensaje del Asistente:** Es la respuesta generada por el modelo, fundamentada en las instrucciones del sistema y en las solicitudes del usuario. Puede incluir explicaciones, sugerencias, ejemplos o cualquier contenido que resuelva la consulta planteada y mantenga la coherencia con el contexto.

Esta estructura de roles garantiza una interacción fluida y coherente, permitiendo que el LLM ajuste dinámicamente sus respuestas según las directrices del sistema y las necesidades del usuario.

2.4. Técnicas de Mejora de Rendimiento

- La **Ingeniería de Prompt** es una técnica crucial para guiar y optimizar las respuestas de sistemas basados en Inteligencia Artificial. Implica la elaboración meticulosa de instrucciones bien diseñadas para obtener respuestas precisas y concisas al interactuar con LLMs (Gartlehner et al., 2023).

  Algunas de las técnicas de mejora de rendimiento son:

  a) ***Chain of Thought* (Cadena de pensamiento o CoT por sus siglas en inglés):** implica presentar tareas complejas de razonamiento en múltiples pasos a los LLMs a través de ejemplos de respuestas paso a paso. Esta técnica ha mostrado rendimiento avanzado en tareas desafiantes como la aritmética y razonamiento simbólico, demostrando su capacidad para generar procesos de razonamiento complicado en los LLMs (Kojima et al., 2022). La técnica permite a los modelos manejar tareas sofisticadas que no se ajustan a las leyes de escalamiento estándar, ampliando las capacidades para resolver problemas.

  b) ***Self Consistency*:** Propuesta por Wang et al. (2023), se diseñó para reemplazar la decodificación utilizada en el CoT. Este método, en vez de confiar en una única ruta de razonamiento para llegar a la respuesta, toma múltiples muestras posibles de caminos de razonamiento, cada uno diferente y diverso. El proceso consiste en generar varias rutas de razonamiento para una misma pregunta o problema. Luego, se selecciona la respuesta más consistente entre estas múltiples rutas que se generaron. Esta técnica mejora la fiabilidad y coherencia de las respuestas de los LLM, asegurando que las salidas generadas se alineen con el proceso de razonamiento previsto y mantengan la consistencia lógica (Taveekitworachai et al., 2023).

  c) ***Few-Shot*:** Implica condicionar a los LLM para resolver diversas tareas usando prompts creadas a partir de un número limitado de pares de ejemplos en cuanto a entrada-salida. Esta técnica permite a los LLM generalizar eficazmente a través de tareas con datos de entrenamiento mínimos, mostrando un rendimiento robusto en entornos con pocos recursos y escenarios de cero ejemplos (zero-shot)(Ma et al., 2023).

- El *Fine-tuning* es un proceso donde se toma un LLM pre-entrenado y se ajustan sus parámetros para especializarlo en tareas o dominios específicos (Durán, 2023). Esto se logra continuando el entrenamiento con un conjunto de datos más pequeño y específico para la tarea. Es útil porque permite adaptar un modelo general a necesidades particulares, mejorando su rendimiento en tareas concretas como la traducción de textos técnicos o la asistencia al cliente, haciéndolo más rápido y eficiente para aplicaciones en tiempo real.

- *Retrieval-Augmented Generation* (o RAG por sus siglas en inglés) combina un sistema de recuperación de información con un modelo de generación de secuencias. Primero, el sistema de recuperación busca en una gran base de datos para encontrar información relevante, que luego alimenta al modelo generativo para producir respuestas informadas y contextualizadas. RAG es especialmente útil para tareas que requieren una comprensión profunda y detallada, como la generación de respuestas para preguntas complejas o la elaboración de resúmenes de investigación, donde la precisión es crucial (Lewis et al., 2021).

La principal diferencia entre *fine-tuning* y RAG radica en la forma en que cada técnica aprovecha los datos y modifica el modelo, la primera reentrena un modelo de aprendizaje profundo en un conjunto de datos específico del dominio para optimizar su rendimiento en tareas concretas, ajustando directamente sus pesos y sesgos. La segunda, en cambio, aumenta un modelo de NLP conectándolo en tiempo real a la base de datos de la organización, recuperando documentos relevantes y enriqueciendo dinámicamente el contexto de generación (Belcic and Stryker, 2024).

La elección de la técnica de ingeniería de prompts es crucial para el éxito de tareas que requieren razonamiento complejo. Para el propósito de este estudio, analizar y calificar soluciones matemáticas, se seleccionó la técnica CoT. Investigaciones pioneras como la de Wei et al. (2022) demostraron que al instruir a los LLMs a "pensar paso a paso" y a externalizar su proceso de razonamiento, su rendimiento en tareas aritméticas, de sentido común y de razonamiento simbólico mejora drásticamente.

Esta metodología es particularmente efectiva porque descompone un problema complejo en una secuencia de pasos intermedios y manejables. En el contexto de la evaluación educativa, esto permite guiar al modelo para que primero identifique los conceptos clave, luego compare la respuesta del estudiante con la pauta, y finalmente justifique la asignación de puntaje, emulando el proceso cognitivo de un evaluador humano. Dada su probada eficacia para mejorar la fidelidad y la lógica en problemas de múltiples pasos, el CoT fue considerado el enfoque más adecuado para este trabajo.

2.5. Retroalimentación en la Educación

La retroalimentación, en el contexto educativo, se refiere al proceso a través del cual se proporciona información constructiva a las y los estudiantes o usuarios sobre su desempeño, tareas o interacciones con el sistema. Esta retroalimentación es esencial para el aprendizaje, ya que ayuda a los individuos a identificar áreas de mejora, entender sus errores y corregir sus enfoques en el futuro. En el ámbito educativo ha sido objeto de un amplio estudio y se ha demostrado que tiene efectos positivos significativos en el aprendizaje de los estudiantes. Según Mateos Morfín and Flores Aguirre (2022), la retroalimentación puede ser un factor determinante en el éxito educativo, ya que proporciona a las y los estudiantes información específica sobre su desempeño, lo que les permite ajustar sus estrategias de aprendizaje. Además, Andrés Jiménez and González Zúñiga (2016) destacan que la retroalimentación fomenta la autorregulación al permitir que las y los estudiantes reflexionen sobre su proceso de aprendizaje y tomen decisiones informadas para mejorar.

Los modelos educativos han enfatizado en la importancia de utilizar criterios e indicadores formativos en la retroalimentación. Riojas Rivera et al. (2023) argumentan que la retroalimentación debe ser específica y orientada a la mejora continua. Asimismo, Burga Vargas et al. (2023) enfatizan la necesidad de motivar la agencia de las y los estudiantes, promoviendo su participación activa en el proceso de retroalimentación. Esto se alinea con las propuestas de Hernández Gutiérrez et al. (2017), quienes sugieren que la retroalimentación debe ser un proceso colaborativo que involucre tanto a docentes como a estudiantes.

En conclusión, la retroalimentación es un componente esencial del proceso de

enseñanza-aprendizaje que no solo mejora el rendimiento académico, sino que también promueve la autorregulación y el aprendizaje autónomo. Las prácticas de retroalimentación deben ser cuidadosamente diseñadas y ejecutadas, considerando la interacción entre estudiantes y docentes, así como la utilización de criterios claros y específicos. La implementación de estrategias de retroalimentación efectivas puede transformar la experiencia de aprendizaje, haciendo que las y los estudiantes se conviertan en participantes activos en su propio proceso educativo.

2.5.1. Feedback Personalizado.

El Feedback Personalizado se refiere a la información proporcionada a un estudiante que está adaptada a su producción o comportamiento individual. Su objetivo principal, al igual que la retroalimentación en general, es reducir las discrepancias entre el rendimiento actual del estudiante y un objetivo o estándar deseado. Para ser verdaderamente valiosa, la retroalimentación debe informar sobre una tarea o proceso de aprendizaje específico y estar adaptada al resultado o acciones del estudiante para ayudar a cerrar esa brecha (Akavova et al., 2023).

2.6. Estado del Arte

2.6.1. Aplicaciones y Desempeño de LLMs en la Calificación Educativa

Los LLMs han llamado la atención de manera significativa en varios campos de investigación debido a sus notables capacidades. En el ámbito de la Ingeniería de Software, los LLMs han mostrado avances en generar texto similar al humano, ayudando en tareas relacionadas con el código (Al-Kaswan et al., 2024). Además, en el sector biomédico, se han aprovechado para procesar registros médicos electrónicos no estructurados, ofreciendo oportunidades para mejorar la atención médica (Albarqawi, 2022).

En el campo de la educación, su aplicación más prometedora es la asistencia en la evaluación. Estudios recientes han pasado de afirmaciones generales a pruebas empíricas rigurosas sobre la viabilidad de los LLMs como herramientas de calificación. Por ejemplo, Mok et al. (2024) realizaron una prueba empírica comparando la calificación de problemas de física de pregrado por parte de humanos

frente a varios LLMs (Gemini 1.5 Pro, GPT-4, GPT-4o y Claude 3.5 Sonnet). Sus resultados indican que, aunque la calificación de la IA es propensa a errores matemáticos y alucinaciones, su calidad mejora sustancialmente cuando se le proporciona un esquema de calificación (rúbrica), acercándose al rendimiento humano. El estudio también encontró una correlación entre la capacidad de un LLM para resolver un problema y su habilidad para calificarlo correctamente.

De manera complementaria, Chen and Wan (2024) investigaron el uso de GPT-4o para asignar puntaje parcial a las explicaciones escritas de los estudiantes en problemas de física. Demostraron que, mediante una cuidadosa ingeniería de prompts y el uso de la auto-consistencia (ejecutar la calificación múltiples veces y tomar el resultado más frecuente), es posible alcanzar una precisión a nivel humano. Este enfoque permite evaluar no solo el resultado numérico, sino el proceso de razonamiento del estudiante, un aspecto fundamental en la educación STEM.

Los hallazgos de estos estudios son consistentes con la investigación realizada en otros dominios cuantitativos. En un estudio comparativo en módulos de educación superior como Tecnologías de la Información y Gestión, Ragolane et al. (2024), encontraron que GPT-4 superó notablemente a su predecesor. Mientras que GPT-3.5 mostró una diferencia promedio del 24 % en las calificaciones en comparación con los evaluadores humanos, GPT-4 redujo esta brecha a solo un 4 %. Este estudio también subraya un factor crítico: el rendimiento de la IA está intrínsecamente ligado a la calidad de la rúbrica o la memoria. En módulos con preguntas objetivas y rúbricas bien estructuradas, como Estadística, la IA demostró ser altamente eficaz, proponiendo un modelo híbrido donde la IA gestiona la escala y la objetividad, mientras que los humanos aportan el juicio en tareas subjetivas.

Estos trabajos ilustran que, si bien los LLMs no son perfectos, su rol en la automatización de la calificación es cada vez más viable, especialmente en dominios técnicos donde la evaluación del proceso es tan importante como la respuesta final y donde se pueden diseñar rúbricas claras.

2.6.2. Desafíos en el Flujo de Trabajo y la Evaluación de la Calificación con IA

Más allá del rendimiento intrínseco de los modelos, la implementación práctica de la calificación asistida por IA presenta desafíos significativos en el flujo de trabajo. Aquí, Kortemeyer et al. (2024) exploraron estos desafíos utilizando un examen de termodinámica escrito a mano. El mayor obstáculo identificado fue la conversión de las respuestas manuscritas a un formato legible por máquina mediante Reconocimiento Óptico de Caracteres (OCR). Este proceso es propenso a errores, especialmente con expresiones matemáticas complejas y diagramas dibujados a mano, cuya interpretación por parte de la IA resultó ser menos fiable que la de las derivaciones matemáticas.

Para mitigar la incertidumbre inherente al proceso, se han propuesto estrategias innovadoras. Lo realizado por Chen and Wan (2024) fue que desarrollaron un índice de confianza de calificación basado en la entropía de Shannon. Este índice mide la variabilidad en los resultados de calificación a través de múltiples ejecuciones de auto-consistencia. Las respuestas con alta entropía (es decir, alta variabilidad) pueden ser marcadas automáticamente para una revisión humana, optimizando así la supervisión y combinando la eficiencia de la IA con la precisión del juicio experto.

Estos hallazgos subrayan que la calificación con IA no es simplemente una tarea de entrada y salida, sino un proceso complejo que requiere un diseño cuidadoso del flujo de trabajo, desde la digitalización de las respuestas hasta la gestión de la incertidumbre del modelo.

2.6.3. Modelos

2.6.3.1. Familias de Modelos

Los LLMs se agrupan en distintas familias, desarrolladas por empresas o comunidades, cada una con características y objetivos específicos:

- **Meta:** La familia LLaMA, presentada por Meta en 2023, se caracteriza por su eficiencia computacional y accesibilidad en entornos de recursos limitados. Sus versiones más pequeñas, como el modelo de 8000 Millones (M) de parámetros, permiten su ejecución en hardware más pequeños,

mientras que arquitecturas mixtas de expertos, como LLaMA Scout, tienen 17000M activos sobre 109000M totales, funcionan en una sola GPU H100 con cuantización FP4, reduciendo drásticamente los requisitos de cómputo (Dojo Staff; Malhotra).

- **OpenAI:** La serie GPT (Generative Pre-trained Transformer) destaca por su versatilidad en generación de texto, seguimiento de instrucciones complejas y compromiso ético. A través de técnicas de alineamiento (InstructGPT) mejora la fidelidad y seguridad de las respuestas, reduciendo la generación de contenido tóxico, es decir, expresiones que incluyen insultos, lenguaje ofensivo, discurso de odio o mensajes intimidatorios, y minimizando sesgos indeseados.
- **Google DeepMind:** La familia GEMINI integra capacidades nativas multimodales (texto, imágenes, audio, video) con ventanas de contexto de hasta 2 M tokens. Sus variantes Pro y Ultra implementan modos de razonamiento profundo ("Deep Think") para simulaciones interactivas y análisis avanzado de código y datos, escalando desde dispositivos edge hasta infraestructuras en la nube (Google DeepMind, 2025; Team et al., 2024).
- **DeepSeek:** Desarrollada bajo la dirección de Liang Wenfeng, DeepSeek R1 es un modelo open-source de alto rendimiento entrenado en 57 días usando 2048 GPUs H800 y miles de GPUs de menor potencia, con un costo total de entrenamiento de US$ 5,6 M, muy por debajo de los cientos de millones asociados a otros líderes del mercado, orientado a finanzas, ciencia y políticas públicas con filosofía de IA confiable y colaborativa (huongnguyen253, 2025; Robison, 2025).

Modelo	Familia	Costo input (USD /1M tokens)	Costo output (USD /1M tokens)
Llama 3.1 8B	Meta	0.09	0.09
Llama 3.1 70B	Meta	0.52	0.75
Llama-3.3-70B-Versatile	Meta	0.59	0.79
GPT-4o-2024-08-06	OpenAI	2.50	10.0
GPT-4o-mini-2024-07-18	OpenAI	0.15	0.60
GPT-3.5-turbo	OpenAI	0.50	1.50
Gemini-2.0-flash-001	Google	0.10	0.40
Gemini 1.5 Flash	Google	0.075	0.30
Gemini 1.5 Flash-8B	Google	0.075	0.30

Figura 2.6.1: Modelos y sus costos. Fuente: Elaboración propia.

2.6.3.2. Medición de Modelos (*Benchmarks y Datasets*)

Los *benchmarks* en IA son pruebas estandarizadas que se utilizan para evaluar y comparar el rendimiento de diversos modelos de lenguaje (Dua et al., 2019). Estas pruebas, como MMLU o GSM-8K, son fundamentales para medir capacidades generales. Sin embargo, la investigación aplicada en educación, como la de Mok et al. (2024), a menudo requiere la creación de metodologías de evaluación personalizadas. En lugar de depender únicamente de benchmarks estándar, estos estudios establecen su "verdad absoluta" (*ground truth*) comparando los resultados de la IA con las calificaciones de expertos humanos (profesores o asistentes de docencia) bajo rúbricas específicas, lo que proporciona una medida de rendimiento mucho más contextualizada y relevante para la tarea educativa en cuestión. Entre los *benchmarks* más conocidos en el campo de los LLMs, se definen los siguientes:

a) **MMLU (*Massive Multitask Language Understanding*):** Está diseñado para medir el conocimiento adquirido durante el preentrenamiento mediante evaluaciones zero-shot y few-shot en 57 áreas de STEM, humanidades,

ciencias sociales y más. Varía en dificultad desde nivel elemental hasta profesional, evaluando tanto conocimiento general como resolución de problemas (Hendrycks et al., 2021).

b) **GSM-8K(*Grade School Math 8K*):** Conjunto de datos de 8,5 Kilo (K) problemas matemáticos de escuela primaria, creados por escritores humanos. Aunque conceptualmente simples, requieren múltiples pasos de razonamiento, lo que desafía a los modelos de última generación (Cobbe et al., 2021).

c) **HellaSwag:** Evalúa la inferencia del lenguaje natural (o NLI por sus siglas en inglés) de sentido común cuya clave está en letras o palabras engañosas generadas adversarialmente para completar historias de sentido común, triviales para humanos ($> 95\,\%$ de acierto) pero difíciles para los modelos (Zellers et al., 2019).

d) **WinoGrande:** Dataset de 44 K problemas inspirado en el Winograd Schema Challenge, ampliado mediante un procedimiento de crowdsourcing y reducción sistemática de sesgos con el algoritmo AfLite, para asegurar robustez a asociaciones sesgadas (Sakaguchi et al., 2019).

e) **DROP (*Discrete Reasoning Over Paragraphs*):** Es de comprensión de lectura con 96 K preguntas que exige resolver referencias en el texto y realizar operaciones discretas (suma, conteo, ordenamiento) sobre párrafos, superando las capacidades de modelos que dependen solo de atajos de entidad o parafraseo (Dua et al., 2019).

2.6.4. Beneficios del Feedback Personalizado con LLMs

El uso de LLMs para ofrecer retroalimentación personalizada está revolucionando los enfoques tradicionales de aprendizaje y evaluación. Estas herramientas permiten automatizar la retroalimentación en respuestas abiertas. Esta capacidad fue explorada empíricamente por Wan and Chen (2024), quienes utilizaron GPT-3.5 para generar feedback sobre respuestas a preguntas conceptuales de física. Un hallazgo clave fue que, aunque los estudiantes calificaron la corrección del feedback de la IA y el humano de manera similar, consistentemente calificaron el feedback de la IA como "más útil". Esto sugiere que los LLMs pueden generar comentarios que los estudiantes perciben como más detallados o mejor adaptados a sus respuestas específicas.

Podemos decir que los estudios de Chen and Wan (2024) amplían esta perspectiva al demostrar que los LLMs pueden ser instruidos para generar explicaciones claras y detalladas sobre los resultados de la calificación, lo que aumenta la transparencia del proceso para el estudiante y le permite entender por qué recibió una determinada puntuación. La integración de la retroalimentación humana en el proceso, asegura que los LLMs proporcionen comentarios que resuenen con los estudiantes y que sean más relevantes para su contexto educativo.

2.6.5. Limitaciones y Desafíos Específicos

A pesar de su potencial, los LLMs presentan limitaciones concretas en tareas de calificación, especialmente en dominios matemáticos. En un estudio identificaron que los modelos de IA tienden a cometer errores matemáticos y a "alucinar" (inventar información). Además, mostraron un comportamiento de calificación inconsistente: sin un esquema de calificación, tienden a ser demasiado indulgentes, mientras que con un esquema estricto, pueden volverse demasiado rígidos e incapaces de reconocer soluciones alternativas pero correctas (Mok et al., 2024).

Por su parte, Kortemeyer et al. (2024) destacaron las dificultades operativas. Observaron que los LLMs cometen "errores de conteo" al aplicar rúbricas muy detalladas a problemas largos, perdiendo el rastro de los puntos asignados. También confirmaron la poca fiabilidad de la IA para interpretar y calificar gráficos o diagramas dibujados a mano, una tarea común en los exámenes de ciencias e ingeniería. Estas limitaciones subrayan que, si bien los LLMs son herramientas poderosas, su implementación efectiva requiere supervisión humana y flujos de trabajo cuidadosamente diseñados para mitigar sus debilidades inherentes.

Capítulo 3

Metodología

En este capítulo se describe el procedimiento metodológico empleado para diseñar un modelo matemático cuyo objetivo es optimizar la selección y configuración de los LLM con miras a entregar retroalimentación de alta calidad en cursos de matemáticas a nivel universitario, procurando minimizar los costos asociados.

3.1. Proceso general para la generación de Feedback Personalizado

El flujo de trabajo abarca cinco etapas distribuidas a lo largo de diez días hábiles, correspondiente al periodo de intervención en el curso por cada evaluación que se realice. Para el desarrollo de esta tesis se emplearon únicamente datos históricos y evaluaciones previamente calificadas por la docente responsable de la asignatura. Por lo tanto, el procedimiento de recolección difiere del que se aplicaría durante la intervención en tiempo real, pues desde el inicio se debe fijar la configuración del modelo (incluida la temperatura) que se utilizó en las etapas posteriores.

Figura 3.1.1: Proceso general para la generación de Feedback Personalizado. Fuente: Elaboración propia.

El proceso consta de las siguientes cinco etapas:

1. **Digitalización.** Se escanean los certámenes entregados por las y los estudiantes, generando (i) un archivo PDF que contiene el documento completo y (ii) archivos PNG independientes para cada página.

2. **Transcripción de respuestas.** Cada imagen PNG se introduce en el LLM para obtener una transcripción en formato Markdown segmentada por pregunta. A continuación, un ser humano revisa y corrige manualmente el texto generado hasta garantizar su fidelidad al contenido original. Esta etapa produce una transcripción validada de la respuesta de la o el estudiante.

3. **Análisis técnico.** Al LLM se le proporciona la transcripción validada junto con la pauta de corrección y, de ser aplicable, la rúbrica asociada. El modelo emite un texto detallado de aciertos y omisiones, asignando puntajes parciales según los criterios establecidos. Un ser humano revisa el análisis técnico y lo valida realizando los ajustes que considere pertinentes. El prompt utilizado para esta etapa se encuentra completo en el Apéndice A.3.1.

4. **Feedback Personalizado.** Utilizando el análisis técnico, el LLM genera el Feedback Personalizado para la respuesta de la o el estudiante. Cada Feedback Personalizado se revisa, corrige y evalúa usando la misma rúbrica para asegurar su coherencia y utilidad pedagógica. Esta etapa genera un "feedback prevalidado". El prompt detallado para esta etapa se encuentra en el Apéndice A.3.2.

5. **Entrega a la/el docente.** Finalmente, el "feedback prevalidado" es revisado, corregido y validado por la o el docente, quién luego se lo entrega a sus estudiantes a través de correo electrónico o la plataforma de gestión de aprendizaje institucional.

El proceso completo, desde la recolección de la evaluación hasta la entrega final de la retroalimentación, se detalla visualmente a continuación. El flujo de trabajo se divide en tres partes principales, presentadas en las Figuras 3.1.2, 3.1.3, 3.1.4 3.1.5 y 3.1.6.

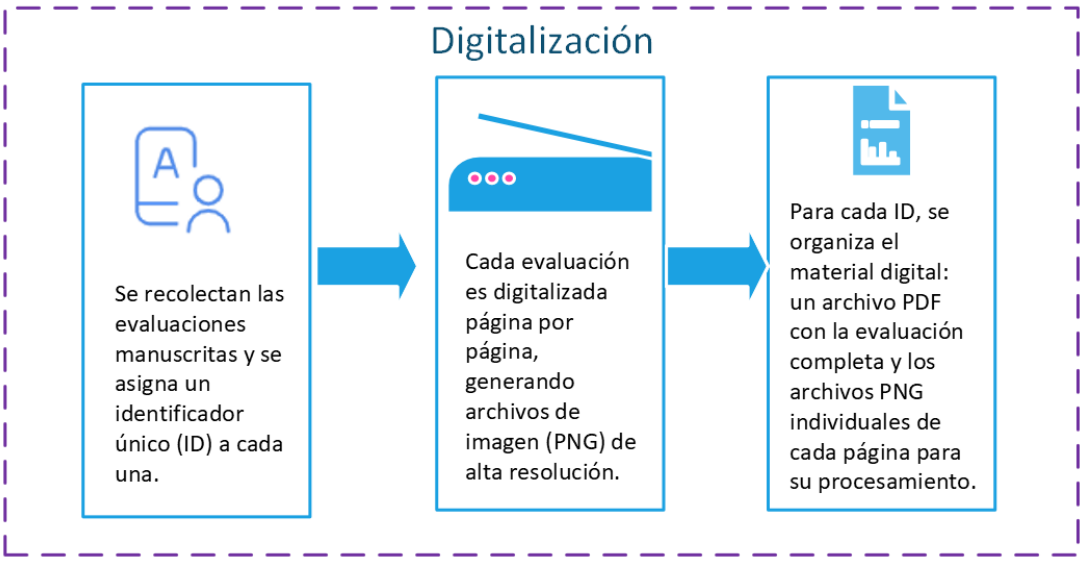


Figura 3.1.2: Etapa 1: Digitalización. Fuente: Elaboración propia.

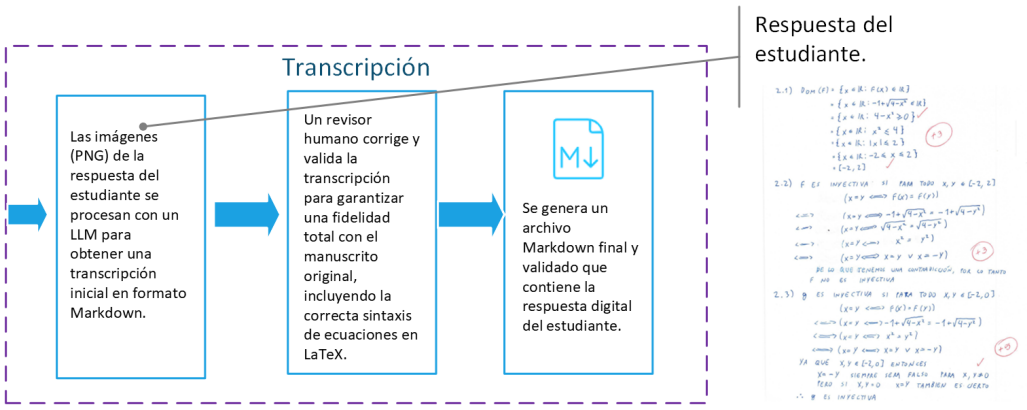


Figura 3.1.3: Etapa 2: Transcripción. Fuente: Elaboración propia.

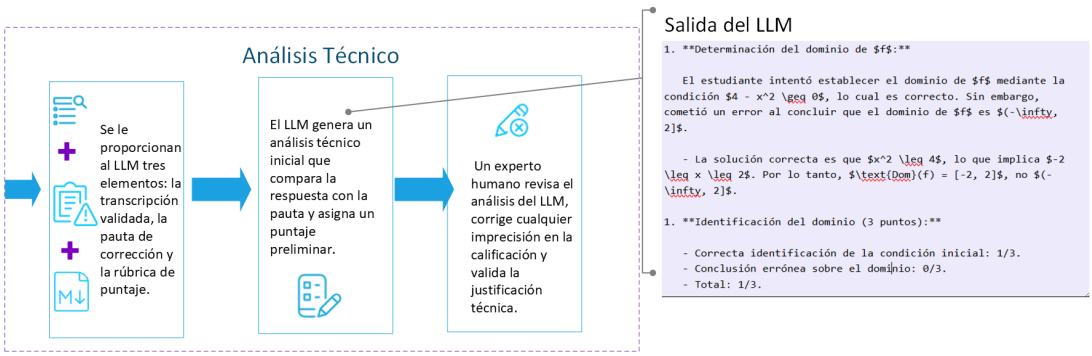


Figura 3.1.4: Etapa 3: Análisis Técnico. Fuente: Elaboración propia.

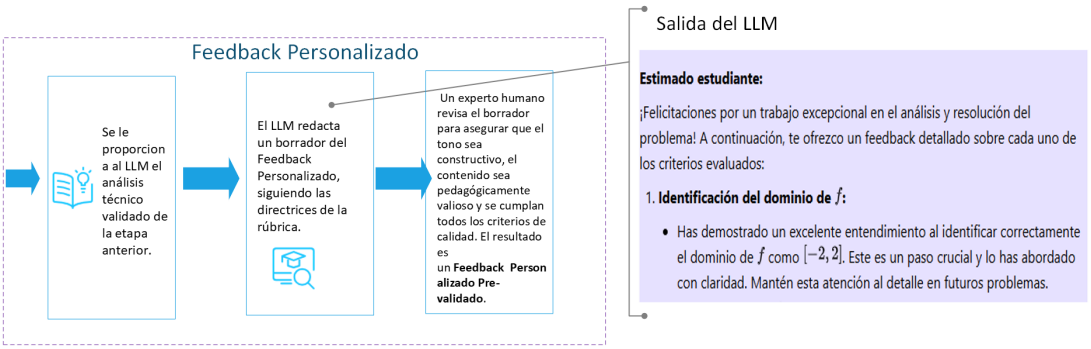


Figura 3.1.5: Etapa 4: Feedback Personalizado. Elaboración propia.

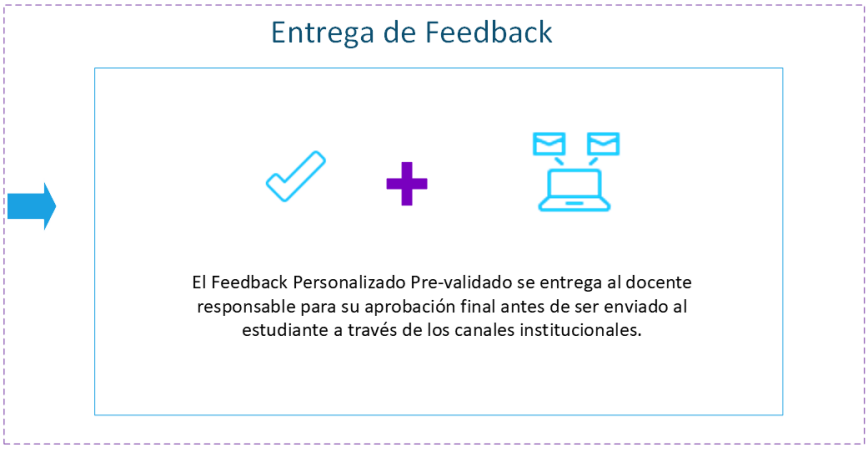


Figura 3.1.6: Etapa 5: Entrega de Feedback. Fuente: Elaboración propia.

3.2. Diseño Experimental

3.2.1. Recolección de la muestra a utilizar.

Para la ejecución del diseño experimental, se utilizó una muestra de datos proveniente de certámenes reales. Específicamente, se seleccionaron aleatoriamente respuestas de estudiantes a la pregunta 2 del segundo certamen del curso de Álgebra I, correspondiente a la sección 11 del primer semestre de 2024 (detallada en el Apéndice A.1).

La calificación de las salidas de los modelos se fundamentó en la experticia de la docente del curso, la Dra. Mónica Selva. Sus calificaciones originales de los certámenes se establecieron como el estándar de referencia para medir la "Calidad del Análisis Técnico". Asimismo, la Dra. Selva fue la responsable de evaluar la

calidad pedagógica de los feedbacks generados por los modelos.

3.2.2. Diseño factorial 3^k en bloques 3^p

Para analizar la calidad de la retroalimentación y optimizar los costos, se ha seleccionado un diseño factorial 3^k en bloques 3^p, siguiendo los lineamientos de Montgomery (2001). Cada uno de los cuatro factores considerados, que son modelo utilizado para el análisis técnico, temperatura del modelo en el análisis técnico, modelo empleado para la generación de la retroalimentación y temperatura del modelo en la generación de la retroalimentación, cuenta con tres niveles (ver 3.2.2).

Este diseño factorial permite examinar simultáneamente el efecto de múltiples factores y sus interacciones, en particular, el impacto del modelo y la temperatura en las fases de análisis técnico y generación de retroalimentación. La incorporación de bloques en la estructura experimental, donde cada bloque representa una respuesta de un estudiante, ayuda a mitigar la variabilidad individual de los estudiantes, mejorando la precisión en la estimación de los efectos de los factores.

Dado que un diseño factorial completo requeriría evaluar las 81 combinaciones en cada respuesta, lo que resultaría inviable por su elevado costo computacional, la inclusión de bloques permite distribuir las combinaciones de manera equitativa. Esto asegura una evaluación representativa sin comprometer la capacidad de estimar los efectos principales y sus interacciones más relevantes. El cuadro 3.2.1 muestra las 81 combinaciones distribuidas en 9 respuestas estudiantiles, denotadas como $R_1, R_2, \ldots, R_9$. Cada columna representa un bloque correspondiente a una respuesta estudiantil, y cada fila contiene una combinación específica de los factores evaluados. Además, se ha introducido un criterio de bloqueo para controlar la variabilidad atribuida a factores ajenos al estudio, como lo es el tipo de respuesta, por ejemplo, si el estudiante resolvió extensamente pero de forma incorrecta, o si no abordó la pregunta en absoluto.

Para minimizar la confusión de interacciones, la asignación de tratamientos dentro de cada bloque se ha realizado de manera balanceada, asegurando que cada combinación factorial sea evaluada en múltiples bloques cuando sea posible. Se ha aplicado un procedimiento de asignación en el que cada bloque recibe un subconjunto de combinaciones factoriales que cubre de manera equitativa los niveles de los factores principales, evitando la sobre-representación de un

nivel específico dentro de un mismo bloque. Además, cada nivel del factor aparece la misma cantidad de veces en cada bloque. Este enfoque permite reducir la influencia de efectos de confusión derivados de la estructura de bloqueo, garantizando una comparación más precisa de los efectos principales y de sus interacciones.

R_1	R_2	R_3	R_4	R_5	R_6	R_7	R_8	R_9
0000	0001	2000	0200	0020	0010	1000	0100	0002
0122	0120	2122	0022	0112	0102	1122	0222	0121
0211	0212	2211	0111	0202	0221	1211	0011	0210
1021	1022	0021	1220	1011	1001	2021	1121	1020
1110	1111	0110	1010	1100	1120	2110	1210	1112
1202	1200	0202	1102	1222	1211	2102	1002	1201
2012	2010	1011	2212	2002	2022	0012	2112	2011
2101	2102	1101	2001	2121	2111	0101	2201	2100
2220	2221	1220	2120	2210	2200	0220	2020	2222

Tabla 3.2.1: Diseño en bloques. Fuente: (Montgomery, 2001, p. 378)

En la tabla, cada configuración de cuatro dígitos codifica de manera compacta tanto el modelo como la temperatura utilizados en las dos fases del proceso:

Clave	Modelo Análisis Técnico	Temperatura Análisis Técnico	Modelo Feedback	Temperatura Feedback
0	gpt-4o-2024-08-06	0	gpt-4o-2024-08-06	0
1	llama-3.3-70b-versatile	0,5	llama-3.3-70b-versatile	0,5
2	gemini-2.0-flash-001	0,8	gemini-2.0-flash-001	0,8

Tabla 3.2.2: Niveles correspondiente a la configuración en cada etapa. Fuente: Elaboración propia.

De este modo, por ejemplo, la configuración “1201” indica que para el análisis técnico se empleó Llama-3.3-70b-versatile con temperatura 0,8 y para el feedback GPT-4o-2024-08-06 con temperatura 0,5.

3.3. Definición de Factores y Niveles

Dado el interés en configurar distintos LLM para el análisis, se consideran los siguientes factores (variables de decisión), cada uno con tres niveles:

- **Selección del LLM**: El primer factor es la elección del LLM. Para este factor los tres niveles corresponde a utilizar los modelos M_1: gpt-4o-2024-08-06, M_2: Llama-3.3-70b-versatile y M_3: Gemini-2.0-flash-001.
- **Temperatura**: El segundo factor corresponde a la temperatura. No todos los modelos tienen el mismo rango del valor en que puede variar su temperatura, pero se tomaron los valores que sean equivalentes al trabajar. Los distintos niveles de temperatura seleccionado para las etapas de Análsis Técnico y Feedback Personalizado son 0, 0,5 ó 1 y 0,8 ó 1,6. Se escogen dichas temperaturas para variar las respuestas de los modelos de más deterministas a más aleatorias.

3.4. Definición de Variables

3.4.1. Variables de Respuesta

Las métricas que se obtienen tras ejecutar el modelo para cada combinación son:

- **Calidad del Análisis Técnico** (A_i): Grado con que el modelo describe y analiza la respuesta del estudiante según la pauta y el puntaje asignado por cada ítem. Para determinar esta calidad, se realizará una comparación con el puntaje asignado en cada ítem por el modelo y el puntaje asignado por la docente. Esto se calculará mediante la siguiente ecuación:

$$A_i = 1 - \Delta X_i, \quad \text{donde } \Delta X_i = \frac{\sum_{j=1}^{N} |s_j^{\mathrm{H}} - s_j^{\mathrm{LLM}}|}{\sum_{j=1}^{N} m_j} \tag{3.4.1}$$

  Donde cada i es la calidad que tendrá la respuesta del modelo en la configuración correspondiente. Además, s_i^{H} es la calificación de la docente que corrigió dicha evaluación, s_i^{LLM} es calificación asignado por el LLM y

m_i es el máximo de puntos posibles en el item i, N es la cantidad de items de la pregunta.

Así, se tiene que $A_i = 1$ si el LLM califica exactamente como la docente y $A_i = 0$ si el LLM comete el máximo error posible.

- **Calidad del *Feedback Personalizado*** (F_i): Calidad de la retroalimentación generada, evaluada con una rúbrica especializada (ver Apéndice A.2 con los criterios y descriptores completos). La rúbrica evalúa 9 criterios, estos son:

Criterio	**Muy Bueno**	**Bueno**	**Suficiente**	**Insuficiente**
Claridad				
Especificidad				
Feedback Constructivo				
Se explicita Desempeño Esperado				
Reconocer respuestas correctas del desempeño del estudiante				
Mostrar Errores del desempeño del estudiante				
Consejos para Mejorar				
Feedback respetuoso				
Extensión y Complitud del feedback				

Tabla 3.4.1: Esquema de rúbrica para evaluar el Feedback Personalizado.

La ecuación para determinar la calidad del Feedback Personalizado está dada por:

$$F_i = \frac{\sum_{j=1}^{9} \bar{c}_j}{900} \text{ con } \bar{c}_j = \frac{c_{j1} + c_{j2}}{2} \tag{3.4.2}$$

donde c_{i1} y c_{i2} son el puntaje asignado por dos evaluadores humanos por cada criterio, $\bar{c}_i$ es el puntaje promedio por criterio y 900 es la puntuación máxima posible.

- **Costo del Análisis Técnico** (C_A^i): Para determinar el costo del análisis técnico se tomará el número de tokens utilizados al momento de generar esta salida, los tokens son la unidad básica de texto que los modelos procesan

para generar una respuesta. La fórmula es la siguiente:

$$C_A^i = T_{in} \cdot \text{precio}_{in} + T_{out} \cdot \text{precio}_{out} \tag{3.4.3}$$

Donde T_{in} son los Tokens de entrada, $precio_{in}$ es el precio por token de entrada, T_{out} son los tokens de de salida y $precio_{out}$ es el precio por token de salida.

- **Costo del *Feedback*** (C_F^i): Al igual que en análisis técnico, se determinará el costo del Feedback Personalizado en cuanto a la entrada y salida de tokens generados.

$$C_F^i = T_{in} \cdot \text{precio}_{in} + T_{out} \cdot \text{precio}_{out} \tag{3.4.4}$$

Donde T_{in} son los tokens de entrada, $precio_{in}$ es el precio por token de entrada, T_{out} son los tokens de de salida y $precio_{out}$ es el precio por token de salida.

3.5. Formulación del Modelo de Optimización

3.5.1. Función Objetivo

Sean α, λ, β, y γ los pesos que reflejan la importancia relativa de la calidad del Análisis Técnico, la calidad del Feedback Personalizado, el costo del Análisis Técnico y el costo del Feedback Personalizado, respectivamente. El problema de maximización es:

$$\max_{x \in \mathcal{X}} \quad Z(x) = \alpha\, A(x) \;+\; \lambda\, F(x) \;-\; \beta\, C_A(x) \;-\; \gamma\, C_F(x),$$

donde $\mathcal{X}$ es el conjunto de las 81 combinaciones posibles definidas en la tabla 3.2.1. El objetivo es encontrar la combinación

$$x^* = \arg\max_{x \in \mathcal{X}} Z(x).$$

Dado que los pesos $\alpha, \lambda, \beta, \gamma$ determinan la ponderación relativa de cada componente en la función objetivo, y no están previamente definidos, se exploraron distintos conjuntos de valores para evaluar su impacto en la optimización.

Capítulo 4

Análisis

4.1. Análisis Exploratorio de Métricas de Rendimiento y Costo

Antes de construir y analizar formalmente el índice compuesto Z, es imperativo realizar un análisis exploratorio de las variables individuales que lo componen. El propósito de esta sección es caracterizar el comportamiento de los modelos. Esta exploración nos permitirá comprender los compromisos inherentes a cada modelo y sentará las bases para la agregación de estas métricas.

4.1.1. Análisis de Rendimiento

Los puntajes del Análisis Técnico y el Feedback Personalizado evalúan la calidad de las respuestas generadas por los modelos.

Modelo	**Min**	**Q1**	**Mediana**	**Promedio**	**Q3**	**Max**	**Desv. Est.**
M_1	0,6	0,75	0,825	0,816	0,85	1	0,0948
M_2	0,5	0,712	0,75	0,8	0,95	1	0,136
M_3	0,6	0,775	0,8	0,817	0,875	0,95	0,0835

Tabla 4.1.1: Estadísticas de Puntaje de Análisis Técnico por Modelo. Fuente: Elaboración Propia.

Modelo	Min	Q1	Mediana	Promedio	Q3	Max	Desv. Est.
M_1	0,472	0,562	0,634	0,637	0,714	0,85	0,105
M_2	0,424	0,547	0,62	0,620	0,703	0,878	0,108
M_3	0,509	0,549	0,661	0,653	0,715	0,853	0,106

Tabla 4.1.2: Estadísticas de Puntaje del Feedback Personalizado por Modelo. Fuente: Elaboración Propia.

La Tabla 4.1.1 y la Tabla 4.1.2 detallan las estadísticas descriptivas de los puntajes de Análisis Técnico y Feedback, respectivamente, por cada modelo LLM.

En el Análisis Técnico, 4.1.1, el modelo M_3 presenta el desempeño más consistente y elevado, con una mediana de 0,8 y un promedio de 0,817, acompañado de la menor desviación estándar 0,0835, lo que indica una mayor estabilidad en sus resultados. M_1 también muestra un buen rendimiento con una mediana de 0,825 y promedio de 0,816, pero con una desviación estándar ligeramente superior de 0,0948. Por otro lado, M_2 tiene un promedio y mediana más bajos, 0,8 y 0,75 respectivamente, y la mayor desviación estándar 0,1364, sugiriendo una mayor variabilidad.

En cuanto al Feedback Personalizado, 4.1.2, M_3 continúa liderando con una mediana de 0,661 y un promedio de 0,653, nuevamente con la desviación estándar más baja 0,106. M_1 sigue de cerca con una mediana de 0,634 y promedio de 0,637, y una desviación estándar similar 0,105. M_2 muestra nuevamente el rendimiento más bajo en promedio 0,620 y mediana 0,62, con la desviación estándar más alta 0,108, lo que sugiere que, si bien su rendimiento es competitivo en algunos aspectos, es menos consistente que los otros modelos en la tarea de Feedback.

4.1.2. Análisis de Eficiencia: Costos y Consumo de Tokens

Modelo	Min	Q1	Mediana	Promedio	Q3	Max	Desv. Est.
M_1	0,00787	0,00987	0,0103	0,0105	0,0114	0,0130	0,00133
M_2	0,00231	0,00275	0,00286	0,00291	0,00303	0,00364	0,000294
M_3	0,000496	0,000726	0,000841	0,000844	0,000948	0,00122	0,000166

Tabla 4.1.3: Estadísticas de Costo en USD del Análisis Técnico por Modelo. Fuente: Elaboración propia.

Modelo	Min	Q1	Mediana	Promedio	Q3	Max	Desv. Est.
M_1	0,00555	0,00658	0,00700	0,00706	0,00744	0,00879	0,000801
M_2	0,00140	0,00175	0,00197	0,00204	0,00231	0,00305	0,000427
M_3	0,000303	0,000454	0,000536	0,000518	0,000581	0,000683	0,000104

Tabla 4.1.4: Estadísticas de Costo en USD del Feedback por Modelo. Fuente: Elaboración propia.

Modelo	Min	Q1	Mediana	Promedio	Q3	Max	Desv. Est.
M_1	2882	3357	3506	3515	3639	4286	378
M_2	3007	3501	3640	3642	3753	4431	394
M_3	2875	3379	3568	3546	3652	4401	417

Tabla 4.1.5: Estadísticas de Tokens de Entrada del Análisis Técnico por Modelo. Fuente: Elaboración propia.

Modelo	Min	Q1	Mediana	Promedio	Q3	Max	Desv. Est.
M_1	832	1109	1197	1222	1377	1597	200
M_2	669	805	911	960	1074	1403	220
M_3	934	1634	1923	1925	2238	2963	481

Tabla 4.1.6: Estadísticas de Tokens de Salida del Análisis Técnico por Modelo. Fuente: Elaboración propia.

Modelo	Min	Q1	Mediana	Promedio	Q3	Max	Desv. Est.
M_1	1082	1232	1594	1755	2111	3438	581
M_2	1117	1436	1941	1904	2172	3302	587
M_3	1054	1575	1656	1849	2140	3065	557

Tabla 4.1.7: Estadísticas de Tokens de Entrada del Feedback por Modelo. Fuente: Elaboración propia.

Modelo	Min	Q1	Mediana	Promedio	Q3	Max	Desv. Est.
M_1	778	899	970	974	1032	1194	105
M_2	841	1018	1148	1163	1234	1554	191
M_3	706	1003	1276	1263	1498	1867	329

Tabla 4.1.8: Estadísticas de Tokens de Salida del Feedback por Modelo. Fuente: Elaboración propia.

Un alto rendimiento debe ser contextualizado por su costo operativo. Al analizar las estadísticas de costo, como se detalla en la Tabla 4.1.3 y la Tabla 4.1.4, se observa que M_3 es consistentemente el modelo más económico, con promedios de 0,000844 USD para el Análisis Técnico y 0,000518 USD para el Feedback Personalizado, demostrando una notable eficiencia en costo. Le sigue M_2, que aunque más costoso que Gemini, mantiene precios competitivos, 0,00291 USD en Análisis Técnico y 0,00204 USD en Feedback Personalizado. En contraste, M_1 es el modelo con los costos más elevados en ambas etapas, promediando 0,0105 USD para Análisis Técnico y 0,00706 USD para el Feedback Personalizado. La baja desviación estándar en los costos para todos los modelos sugiere una consistencia en sus tarifas. En cuanto al consumo de tokens, los modelos presentan patrones distintos. Para los tokens de entrada, el Análisis Técnico Tabla 4.1.5 muestra promedios similares entre los modelos M_1 con 3515, M_2 con 3642 y M_3 con 3546, mientras que en el Feedback Tabla 4.1.7, el consumo de tokens de entrada es generalmente menor, con M_3 1849 y M_2 1904 ligeramente por encima de M_1 1755. La variabilidad en los tokens de entrada es mayor en la etapa de Feedback Personalizado. Respecto a los tokens de salida, Tabla 4.1.6 y Tabla 4.1.8, M_3 tiende a generar el mayor volumen en ambas etapas, destacándose también por la mayor desviación estándar en tokens de salida para el Análisis Técnico. En resumen, M_3 es la opción más atractiva en términos de costo, aunque su volumen de tokens de salida puede ser mayor, posicionando a M_1 como la alternativa más costosa y a M_2 como una opción equilibrada en términos de eficiencia.

4.2. Elección de las constantes

La construcción del índice compuesto $Z = \alpha X_1 + \lambda X_2 - \beta X_3 - \gamma X_4$ busca integrar el Puntaje del Análisis Técnico (X_1), el Puntaje del Feedback (X_2), el Costo del Análisis Técnico (X_3) y el Costo del Feedback (X_4) en una única métrica. El objetivo es que Z represente de la mejor manera posible la calidad o valor global, donde puntajes altos son deseables y costos bajos también. Por ello, X_1 y X_2 suman. Para que los costos X_3 y X_4 operen restando en el índice, fueron transformados a sus beneficios invertidos ($X_3' = -X_3$ y $X_4' = -X_4$) antes de iniciar el análisis.

Para determinar los pesos $\alpha, \lambda, \beta, \gamma > 0$ bajo la restricción $\alpha + \lambda + \beta + \gamma = 1$, se

adoptó adoptado un enfoque basado en el Análisis de Componentes Principales (ACP). Este método es idóneo porque busca capturar la máxima variabilidad presente en los datos originales con el menor número de componentes. Al seleccionar el vector de pesos $\mathbf{w} = (\alpha, \lambda, \beta, \gamma)$ que maximiza la varianza de Z, $\text{Var}(Z) = \mathbf{w}^\top \Sigma' \mathbf{w}$, donde Σ' es la matriz de covarianzas de las variables transformadas (X_1, X_2, X_3', X_4'), nos aseguramos de que el índice Z sea una representación robusta y distintiva de las diferencias entre las observaciones. La solución a este problema es, por definición, el primer vector propio ($\mathbf{v}_1$) de la matriz de covarianzas Σ', ya que este vector apunta en la dirección de máxima varianza de los datos. Finalmente, las cargas de esta primera componente principal se tomaron en valor absoluto y se normalizaron para que su suma sea 1, asegurando pesos positivos y coherentes con la estructura deseada del índice.

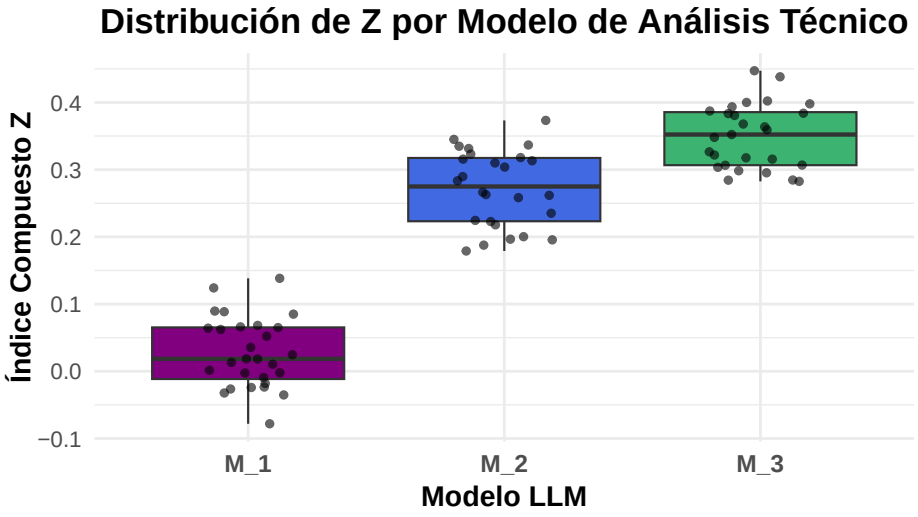
Numéricamente, los pesos obtenidos son:

$$(\alpha,\ \lambda,\ \beta,\ \gamma)\ \approx\ (0{,}12; 0{,}44; 0{,}41; 0{,}03)$$

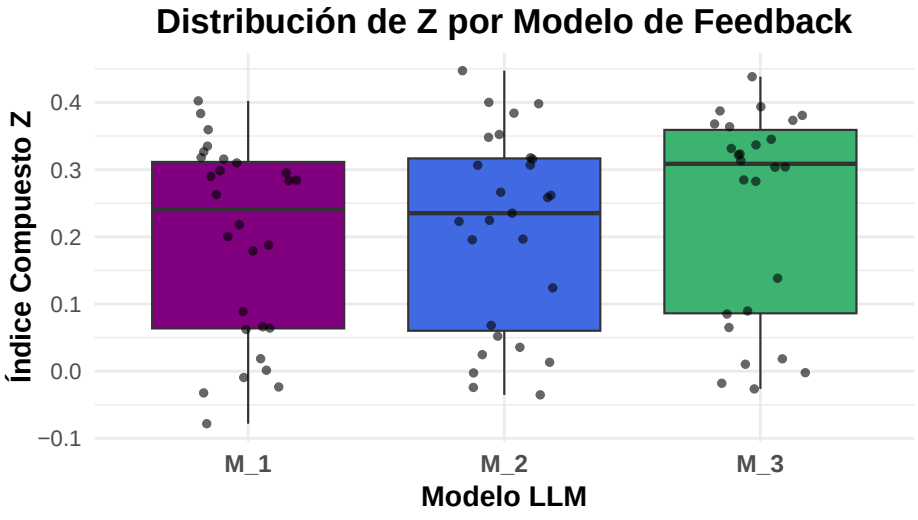
- $\lambda \approx 0{,}44$ (Puntaje del Feedback, X_2): Este es el peso más alto, lo que indica que el Feedback es el factor que más contribuye a la variabilidad total del índice Z. Esto sugiere que las diferencias en los puntajes de feedback tienen una influencia dominante en la forma en que Z discrimina entre distintas observaciones.
- $\beta \approx 0{,}41$ (Costo del Análisis Técnico, X_3): El segundo peso más alto corresponde al costo del análisis técnico. Esto resalta que las variaciones en este costo también son un factor principal en la diferenciación del índice Z.
- $\alpha \approx 0{,}12$ (Puntaje del Análisis Técnico, X_1): Aunque es una contribución positiva, su peso es considerablemente menor que el del feedback o el costo del análisis técnico. Esto podría indicar que, si bien el análisis técnico es importante, su variabilidad tiene un impacto menos distintivo en Z en comparación con otros factores.
- $\gamma \approx 0{,}03$ (Costo del Feedback, X_4): Este es el peso más bajo, sugiriendo que las variaciones en el costo del feedback tienen el menor impacto en la variabilidad de Z.

4.2.1. Síntesis: Visualización del Índice Compuesto Z

El análisis de las métricas individuales revela un claro compromiso, el modelo con el mejor rendimiento no es necesariamente el más eficiente. Para evaluar este balance de forma holística, se construye el índice compuesto Z, cuya metodología se detallará en la siguiente sección. La Figura 4.2.1 y las Tablas 4.2.1 y 4.2.2 presentan una visualización y cuantificación de este índice, mostrando su distribución en función de los modelos del Análisis Técnico y Feedback.



(a) Distribución de Z por Modelo de Análisis Técnico.



(b) Distribución de Z por Modelo de Feedback.

Figura 4.2.1: Comportamiento del índice compuesto Z en función de los modelos. Fuente: Elaboración propia.

Modelo	Min	Q1	Mediana	Promedio	Q3	Max	Desv. Est.
M_1	−0,0781	−0,0115	0,0185	0,0276	0,0653	0,138	0,0522
M_2	0,179	0,223	0,275	0,273	0,317	0,373	0,0567
M_3	0,283	0,307	0,352	0,350	0,386	0,447	0,0481

Tabla 4.2.1: Estadísticas Descriptivas del Valor Z por Modelo de Análisis Técnico. Fuente: Elaboración propia.

Modelo	Min	Q1	Mediana	Promedio	Q3	Max	Desv. Est.
M_1	−0,0781	0,0637	0,240	0,193	0,311	0,402	0,147
M_2	−0,0352	0,0602	0,235	0,211	0,317	0,447	0,147
M_3	−0,0264	0,0862	0,309	0,239	0,359	0,438	0,155

Tabla 4.2.2: Estadísticas Descriptivas del Valor Z por Modelo de Feedback. Fuente: Elaboración propia.

Como se observa en la Tabla 4.2.1 y la Figura 4.2.1a, cuando actúan como modelo de Análisis Técnico, M_3 sobresale con un promedio de Z de 0,350 y una mediana de 0,352, indicando un balance superior entre rendimiento y eficiencia. Le sigue M_2 con un promedio de Z de 0,273 y una mediana de 0,275, mientras que M_1 obtiene el valor más bajo, con un promedio de Z de 0,0276 y una mediana de 0,0185. La baja desviación estándar para estos modelos en esta configuración, entre 0,0481 y 0,0567, sugiere una consistencia en sus resultados.

Para la configuración de Feedback, como se ilustra en la Tabla 4.2.2 y la Figura 4.2.1b, el patrón se mantiene, aunque con una menor distinción y mayor variabilidad entre los modelos. M_3 conserva el promedio de Z más alto, de 0,239. Sin embargo, las desviaciones estándar son notablemente mayores en esta etapa, desde 0,147 a 0,155, lo que implica una dispersión más amplia en los valores de Z comparado con la etapa de Análisis Técnico.

4.3. Análisis de Resultados

4.3.1. Especificación del Modelo ANOVA Anidado para el Índice Z

Para investigar cómo los diferentes modelos de análisis técnico, las temperaturas asociadas, los modelos de feedback y sus respectivas temperaturas influyen en

el índice compuesto Z, se ajustó un Modelo Lineal Anidado (ANOVA Anidado). La estructura anidada se eligió para reflejar la relación jerárquica del diseño experimental, donde los niveles de temperatura son específicos y solo tienen sentido dentro de cada modelo de LLM.

El modelo matemático que representa esta relación se formula de la siguiente manera:

$$Z_{ijklm} = \mu + M_{AT_i} + M_{FD_j} + T_{AT(i)k} + T_{FD(j)l} + B_m + \epsilon_{ijklm} \tag{4.3.1}$$

Donde Z_{ijklm} es el valor observado del índice Z, μ es la media global, M_{AT_i} y M_{FD_j} son los efectos principales de los modelos de Análisis Técnico y Feedback, $T_{AT(i)k}$ y $T_{FD(j)l}$ son los efectos anidados de la temperatura dentro de cada modelo, B_m es el efecto aleatorio del bloque (estudiante), y ϵ_{ijklm} es el error aleatorio residual. La inclusión del término de bloque B_m es crucial para controlar la variabilidad inherente entre las distintas respuestas de los estudiantes, permitiendo una estimación más precisa de los efectos de interés.

4.3.2. Resultados del Análisis de Varianza (ANOVA)

Los resultados del ANOVA anidado, que incluyen la magnitud del efecto cuantificada por el Eta Cuadrado Parcial (η_p^2), se presentan en la Tabla 4.3.1.

Fuente de Variación	**gl**	**Suma Cuad.**	**Cuadrado Medio.**	**F**	**valor-p**	η_p^2
Estrato: Bloque						
LLM An. Técnico	2	0,0217	0,0108	0,880	0,500	0,37
LLM Feedback	2	0,0108	0,0054	0,437	0,681	0,23
Temperatura Feedback	1	0,0032	0,0032	0,260	0,645	0,08
Residuals	3	0,0370	0,0123			
Estrato: Intra-Bloque						
LLM An. Técnico	2	1,5564	0,7782	407.37	$< 2 \times 10^{-16}$ ***	0,94
LLM Feedback	2	0,0178	0,0089	4,66	0,0134 *	0,14
Temperatura Modelo	6	0,0076	0,0013	0,66	0,6796	0,07
Temperatura Feedback	6	0,0140	0,0023	1,22	0,3091	0,12
Residuals	56	0,1070	0,0019			

Tabla 4.3.1: Resultados del ANOVA Anidado para el Índice Z. Fuente: Elaboración propia.

$^{*}\ p < 0{,}05,\ ^{***}\ p < 0{,}001$

La interpretación de los resultados se centra en el estrato *Intra-Bloque*, que evalúa

los efectos de los factores una vez controlada la variabilidad entre estudiantes.

- `LLM An. Técnico`: Se observa un efecto principal abrumadoramente significativo ($F(2, 56) = 407{,}37, p < {,}001$). El tamaño del efecto, medido por el Eta Cuadrado Parcial ($\eta_p^2 = 0{,}94$), es extremadamente grande, indicando que la elección del modelo de Análisis Técnico explica el 94 % de la varianza en el índice Z. Este es, con diferencia, el factor más determinante del rendimiento global.
- `LLM Feedback`: El modelo de Feedback también presenta un efecto estadísticamente significativo ($F(2, 56) = 4{,}66, p = {,}013$), aunque con una magnitud de efecto considerablemente menor ($\eta_p^2 = 0{,}14$).
- **Interacciones Anidadas**: Ninguna de las interacciones con la temperatura fue significativa ($p > {,}05$), lo que sugiere que, en el rango estudiado, el efecto de la temperatura no varía de manera relevante entre los distintos modelos.

Adicionalmente, en el estrato de *Bloque*, ningún factor fue significativo ($p > {,}05$). Este es un resultado positivo, pues indica que el efecto de los modelos es consistente a través de los diferentes estudiantes, lo que apoya la generalización de los hallazgos.

4.3.3. Diagnóstico y Validación de los Supuestos del Modelo

La validez de las inferencias del ANOVA depende del cumplimiento de tres supuestos sobre los residuos del modelo: normalidad, homocedasticidad (homogeneidad de varianzas) e independencia.

4.3.3.1. Normalidad de los Residuos

La normalidad de los residuos se evaluó mediante la prueba de Shapiro-Wilk y análisis gráfico.

- **Prueba de Shapiro–Wilk:** Arrojó un estadístico $W = 0{,}9886$ con un p-valor de 0,697. Al ser este valor muy superior a 0.05, no se rechaza la hipótesis nula de normalidad.
- **Análisis Gráfico:** El histograma de los residuos (Figura 4.3.1) muestra una distribución que, aunque no es perfectamente simétrica, es compatible con la normalidad.

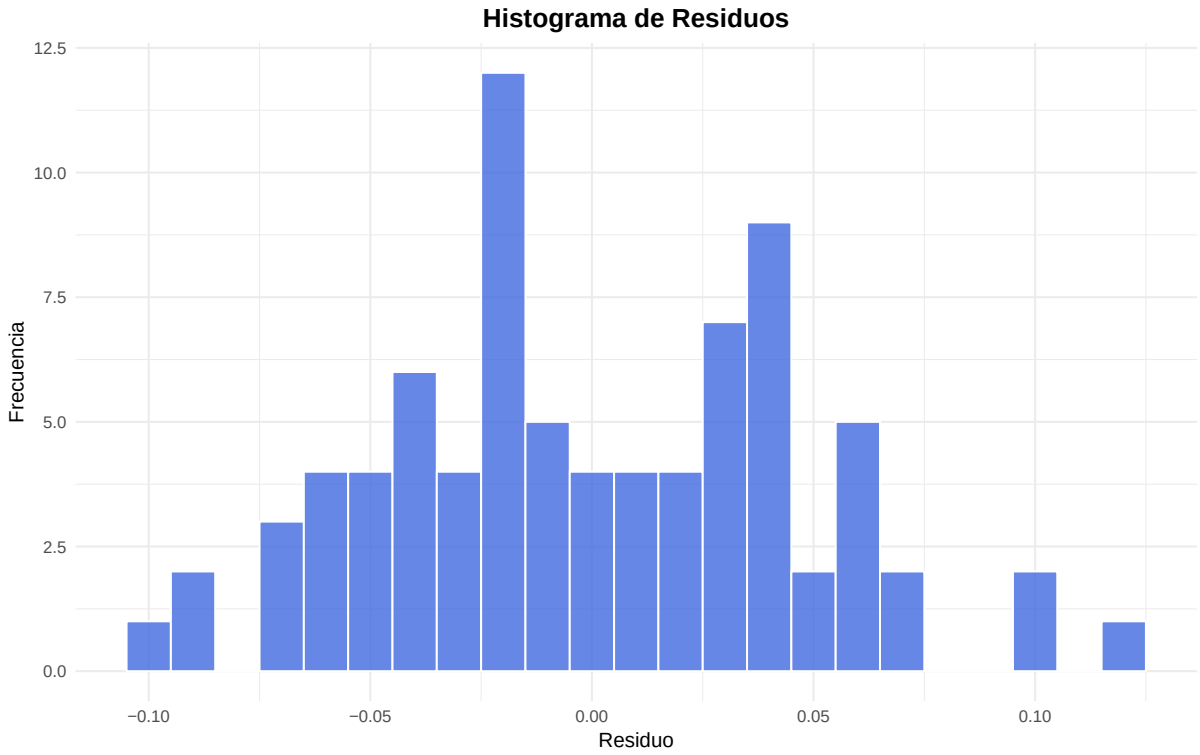


Figura 4.3.1: Histograma. Fuente: Elaboración propia.

4.3.3.2. Homocedasticidad (Homogeneidad de Varianzas)

Se utilizó la prueba de Levene para verificar si la varianza de los residuos es constante a través de los niveles de cada factor. Los resultados indican que se cumple el supuesto para todos los factores y sus combinaciones anidadas: `LLM An. Técnico` ($F(2, 78) = 0{,}58, p = {,}56$), `LLM Feedback` ($F(2, 78) = 0{,}02, p = {,}98$), y las combinaciones de modelo-temperatura de AT ($F(8, 72) = 0{,}24, p = {,}98$) y de FD ($F(8, 72) = 0{,}10, p > {,}99$). La constancia de las varianzas a través de todos los grupos fortalece la validez de los resultados del ANOVA.

4.3.3.3. Independencia de los Residuos

El supuesto de independencia se evaluó mediante la prueba de Durbin-Watson.

- **Prueba de Durbin-Watson:** El test arrojó un estadístico D-W de 1,327 con un p-valor $<.001$.

Discusión: Este resultado indica que se rechaza la hipótesis nula de no autocorrelación, sugiriendo la presencia de una autocorrelación positiva en los residuos. Esta es una violación del supuesto de independencia, posiblemente debida

al orden en que se procesaron los datos. A pesar de esta limitación, la magnitud del efecto de `LLM An. Técnico` ($\eta_p^2 = 0{,}94$) es tan extremadamente grande que la conclusión sobre su significancia se considera robusta. No obstante, este hallazgo se reconoce como una limitación del análisis.

4.3.4. Validación de la Configuración Óptima

Con el fin de confirmar la robustez de los hallazgos y la fiabilidad de la configuración recomendada, se llevó a cabo una etapa de validación externa. Esta fase consistió en aplicar la configuración óptima identificada, el modelo `Gemini-2.0-flash-001` con una temperatura de 1.0, a un conjunto de 19 nuevas respuestas de estudiantes que no formaron parte del experimento factorial original.

Debido al significativo costo en tiempo y esfuerzo humano que implica la evaluación detallada de la calidad del feedback, esta etapa de validación se centró exclusivamente en la métrica de Calidad del Análisis Técnico.. La validación del proceso completo, incluyendo la generación y evaluación del feedback con la configuración óptima, queda propuesta como un trabajo futuro. El objetivo de esta fase fue, por tanto, evaluar si el rendimiento técnico del modelo se mantenía consistente en un nuevo escenario.

Los resultados descriptivos de esta validación se comparan con el rendimiento del mismo grupo de control del experimento original en la Tabla 4.3.2. Se observa una notable consistencia en el rendimiento: la media del A(x) en el conjunto de validación 0,8342 es prácticamente idéntica a la del experimento original 0,8344, lo que sugiere una alta replicabilidad.

Grupo	**Media**	**Desv. Est.**	**Mediana**
Experimento Original	0,8344	0,1093	0,850
Datos de Validación	0,8342	0,1237	0,800

Tabla 4.3.2: Comparación del rendimiento de A(x) entre el grupo de control del experimento original y los datos de validación. Fuente: Elaboración propia.

Para formalizar esta comparación, se realizó una prueba t de Welch para muestras independientes. El resultado de la prueba, $t(14{,}89) = -0{,}003, p = 0{,}997$, indica que no existe una diferencia estadísticamente significativa entre las medias de rendimiento de ambos grupos. Este p-valor, al ser extremadamente alto, confirma

con contundencia que el rendimiento técnico del modelo en los nuevos datos es indistinguible del observado en el experimento inicial.

En conclusión, la evidencia descriptiva e inferencial válida de manera robusta la elección de `Gemini-2.0-flash-001` con temperatura 1 como la configuración recomendada para el análisis técnico. Se ha demostrado que su alto rendimiento no fue una casualidad del conjunto de datos original, sino que es consistente y replicable, lo que respalda su potencial para una implementación práctica y fiable en esta etapa del proceso.

Capítulo 5

Discusión

El análisis cuantitativo de las 81 configuraciones de Grandes Modelos de Lenguaje (LLMs) ha identificado a `Gemini-2.0-flash-001` como el modelo que ofrece el balance óptimo entre calidad pedagógica y eficiencia económica para la generación de feedback en la asignatura de Álgebra I. En esta sección, se discutirán las implicaciones de este y otros hallazgos en el contexto de la literatura existente y las decisiones metodológicas clave del estudio.

5.0.1. El Rol de la Ingeniería de Prompts

A lo largo de este estudio, se aplicó de manera consistente la técnica de *Chain-of-Thought* para estructurar los prompts, tanto en el análisis técnico como en la generación de feedback (ver Apéndice A.3). La elección de CoT se fundamenta en la evidencia robusta de la literatura, donde investigaciones como la de (Wei et al., 2022) demuestran que esta técnica mejora drásticamente el razonamiento de los LLMs en tareas complejas y de múltiples pasos, como es la evaluación matemática.

Introducir el tipo de prompt como un factor adicional habría aumentado exponencialmente la complejidad y el costo del estudio, haciendo inviable un análisis factorial completo.

5.0.2. Implicaciones para la Retroalimentación Educativa

Más allá de la optimización técnica, el hallazgo central de este trabajo tiene profundas implicaciones para la práctica de la retroalimentación en la educación

superior. La capacidad de generar Feedback Personalizado, detallado y oportuno a gran escala, utilizando un modelo costo-efectivo como `Gemini-2.0-flash-001`, aborda directamente el desafío de la alta carga de trabajo docente.

Como señalan autores de referencia en el campo de la evaluación formativa, un feedback efectivo debe ser específico, orientado a la tarea y entregado a tiempo para que el estudiante pueda actuar sobre él. El proceso automatizado en esta tesis permite precisamente esto, ofreciendo a cada estudiante un análisis detallado de sus errores y aciertos, algo que es muy difícil de realizar manualmente en cursos masivos. Por lo tanto, esta investigación no solo valida una herramienta tecnológica, sino que habilita una práctica pedagógica de alto impacto que puede mejorar significativamente la autorregulación y el aprendizaje del estudiante.

Capítulo 6

Conclusión

Esta investigación culminó exitosamente con el desarrollo de un modelo matemático diseñado para optimizar la selección y configuración de LLMs en la generación de Feedback Personalizado para cursos de matemática universitaria. De esta manera, se ha cumplido el objetivo general de la tesis, proporcionando un marco cuantitativo que equilibra la calidad educativa con la eficiencia técnica y económica.

El cumplimiento del objetivo general fue posible gracias a la consecución de los tres objetivos específicos planteados:

1. Se definieron métricas de calidad y coste para evaluar de manera objetiva las salidas de los LLMs. La "Calidad del Análisis Técnico" y la "Calidad del Feedback" permitieron cuantificar el valor pedagógico, mientras que los costos asociados anclaron la evaluación en la viabilidad práctica.

2. Se diseñó un proceso general de cinco etapas para la generación de feedback, estableciendo un flujo de trabajo estructurado y replicable que puede servir como guía para futuras implementaciones institucionales.

3. Se evaluó el rendimiento y costo de los LLMs mediante un diseño experimental factorial anidado. El análisis de los datos reales de la asignatura de Álgebra I permitió una comparación rigurosa, revelando de manera concluyente que el modelo Gemini-2.0-flash-001 ofrece el equilibrio en riqueza pedagógica y eficiencia económica. Este hallazgo demostró además que la elección del modelo es el factor más determinante para el éxito del proceso.

La adopción práctica de esta solución, no obstante, requiere una visión integral

que vaya más allá del rendimiento del modelo. La implementación sostenible debe considerar los costos operativos asociados al personal y la gestión de recursos humanos especializados en cada fase del proceso. Por ello, se recomienda complementar el modelo matemático aquí propuesto con un estudio de costos operativos que incluya tarifas de mano de obra y el desarrollo de programas de formación para el personal.

Finalmente, este trabajo no solo resuelve un problema práctico inmediato, sino que también sienta las bases para una integración informada y estratégica de los LLMs en la educación superior. Abre, además, caminos claros para futuras investigaciones: extender el análisis a otras asignaturas y tipos de evaluación, incorporar métricas de latencia y experiencia de usuario, y evaluar el desempeño longitudinal de los modelos en ciclos académicos completos. Con ello, se avanza firmemente hacia una retroalimentación automatizada, precisa y asequible, que potencie de manera significativa el aprendizaje de la matemática a nivel universitario.

Capítulo 7

Trabajo Futuro

A partir de los hallazgos y las limitaciones identificadas en la investigación, se proponen las siguientes líneas de trabajo futuro para expandir y profundizar el conocimiento sobre la aplicación de LLMs en la educación matemática:

- **Validación Integral del Proceso de Feedback:** Realizar una validación completa que no solo mida la calidad técnica, sino también la percepción y utilidad del feedback por parte de los estudiantes. Esto implicaría recolectar datos cualitativos a través de encuestas o entrevistas para entender cómo interactúan con la retroalimentación generada por los LLMs y si esta impacta positivamente en su aprendizaje.
- **Análisis Longitudinal y de Experiencia de Usuario:** Implementar la solución en un ciclo académico completo para evaluar su desempeño a lo largo del tiempo. Esto permitiría estudiar si la calidad del modelo es consistente en distintas evaluaciones y si se pueden incorporar métricas adicionales, como la experiencia de usuario de las y los docentes y ayudantes que interactúan con la herramienta.
- **Optimización del Flujo de Trabajo y Costo Humano:** Añadir la automatización de las etapas previas al análisis, como la digitalización y transcripción de respuestas manuscritas complejas. Desarrollar un modelo de costo total que no solo incluya las tarifas de uso de los tokens, sino también el costo del capital humano involucrado en la supervisión y validación, para obtener un panorama completo de la viabilidad institucional.

Estas futuras investigaciones permitirán construir sobre las bases sentadas en este trabajo, avanzando hacia una integración cada vez más efectiva, sostenible y pedagógica mente valiosa de la inteligencia artificial en la educación superior.

Bibliografía

Aida Akavova, Zarema Temirkhanova, and Zarina Lorsanova. Adaptive learning and artificial intelligence in the educational space. *E3S Web of Conferences*, 451:06011, 2023. ISSN 2267-1242. doi: 10.1051/e3sconf/202345106011. URL https://www.e3s-conferences.org/10.1051/e3sconf/202345106011. Publisher: EDP Sciences.

Ali Al-Kaswan, Maliheh Izadi, and Arie Van Deursen. Traces of Memorisation in Large Language Models for Code. In *Proceedings of the IEEE/ACM 46th International Conference on Software Engineering*, pages 1–12, Lisbon Portugal, April 2024. ACM. doi: 10.1145/3597503.3639133. URL https://dl.acm.org/doi/10.1145/3597503.3639133.

Ahmad Albarqawi. Exploring the opportunities and limitations of large language models in biomedical sector. December 2022. doi: 10.21203/rs.3.rs-2401418/v1. URL https://www.researchsquare.com/article/rs-2401418/v1.

Salima Aldazharova, Gulnara Issayeva, Samat Maxutov, and Nuri Balta. Assessing AI's problem solving in physics: Analyzing reasoning, false positives and negatives through the force concept inventory. *Contemporary Educational Technology*, 16(4):ep538, November 2024. ISSN 1309-517X. doi: 10.30935/cedtech/15592. URL https://www.cedtech.net/article/assessing-ais-problem-solving-in-physics-analyzing-reasoning-false-positives-and-negatives-through-155 Publisher: Bastas Publications.

Carmen Andrés Jiménez and Rebeca González Zúñiga. El efecto de la retroalimentación en el rendimiento y motivación de los estudiantes de Métodos de investigación para ciencias de la salud en la UNED, Costa Rica. *UNED Research Journal*, 8(2):189–194, October 2016. ISSN 1659-441X, 1659-4266. doi: 10.22458/urj.v8i2.1560. URL https://revistas.uned.ac.cr/index.php/cuadernos/article/view/1560. Publisher: Universidad Estatal a Distancia.

Will Beckford. Sobre el futuro de los modelos de lenguaje preentrenados para la revolución educativa. February 2023. doi: 10.31219/osf.io/dtbmu. URL https://osf.io/dtbmu_v1.

Ivan Belcic and Cole Stryker. RAG vs. fine tuning | IBM. August 2024. URL https://www.ibm.com/mx-es/think/topics/rag-vs-fine-tuning.

Y. Bengio, A. Courville, and P. Vincent. Representation Learning: A Review

and New Perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828, August 2013. ISSN 0162-8828, 2160-9292. doi: 10.1109/tpami.2013.50. URL http://ieeexplore.ieee.org/document/6472238/. Publisher: Institute of Electrical and Electronics Engineers (IEEE).

Christy K. Boscardin, Brian Gin, Polo Black Golde, and Karen E. Hauer. ChatGPT and Generative Artificial Intelligence for Medical Education: Potential Impact and Opportunity. *Academic Medicine*, 99(1):22–27, January 2024. ISSN 1040-2446. doi: 10.1097/acm.0000000000005439. URL https://journals.lww.com/10.1097/ACM.0000000000005439. Publisher: Ovid Technologies (Wolters Kluwer Health).

Víctor Raúl Burga Vargas, Mónica Ysabel Ortega Cabrejos, and Bertila Hernández Fernández. Retroalimentación formativa en el desempeño docente. *Horizontes. Revista de Investigación en Ciencias de la Educación*, 7(27):99–112, February 2023. ISSN 2616-7964. doi: 10.33996/revistahorizontes.v7i27.500. URL https://revistahorizontes.org/index.php/revistahorizontes/article/view/817. Publisher: Centro de Investigacion y Desarrollo Ecuador.

Zhongzhou Chen and Tong Wan. Using Large Language Models to Assign Partial Credit to Students' Explanations of Problem-Solving Process: Grade at Human Level Accuracy with Grading Confidence Index and Personalized Student-facing Feedback. December 2024. doi: 10.48550/arXiv.2412.06910. URL http://arxiv.org/abs/2412.06910. arXiv:2412.06910 [physics].

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training Verifiers to Solve Math Word Problems. November 2021. doi: 10.48550/arXiv.2110.14168. URL http://arxiv.org/abs/2110.14168. arXiv:2110.14168 [cs].

Data Science Dojo Staff. Llama 3.1: All You Must Know About Meta's Most Capable LLM. URL https://datasciencedojo.com/blog/meta-llama-3-1/.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. DROP: A Reading Comprehension Benchmark Requiring Discrete Reasoning Over Paragraphs. April 2019. doi: 10.48550/arXiv.1903.00161. URL http://arxiv.org/abs/1903.00161. arXiv:1903.00161 [cs].

C. de Agustín Durán. Fine-Tuning: ¿qué es? *FOQUM*, 2023. URL https://foqum.io/blog/termino/fine-tuning/.

Bruno Fernandes. El papel y el futuro de los modelos lingüísticos preentrenados en la industria de la educación. February 2023. doi: 10.31219/osf.io/x8kdw. URL https://osf.io/x8kdw_v1.

L. Fernández, A. Mena, M. Magaña, M. Magaña, and M. Fernández. Inteligencia artificial en la educación: modelo de lenguaje de gran tamaño (llm) como recurso educativo. *IPSUMTEC*, 7:157–164, 2024. doi: 10.61117/ipsumtec.v7i2.321.

G Gartlehner, L Kahwati, R Hilscher, I Thomas, S Kugley, K Crotty, M Viswanathan, B Nussbaumer-Streit, G Booth, N Erskine, A Konet, and R Chew. Data Extraction for Evidence Synthesis Using a Large Language Model: A Proof-of-Concept Study. October 2023. doi: 10.1101/2023.10.02.23296415. URL http://medrxiv.org/lookup/doi/10.1101/2023.10.02.23296415.

Google DeepMind. Gemini. *Google DeepMind*, 2025. Último acceso: 9 de julio de 2025.

Jin Guo, Jinghui Cheng, and Jane Cleland-Huang. Semantically Enhanced Software Traceability Using Deep Learning Techniques. In *2017 IEEE/ACM 39th International Conference on Software Engineering (ICSE)*, pages 3–14, Buenos Aires, May 2017. IEEE. doi: 10.1109/icse.2017.9. URL http://ieeexplore.ieee.org/document/7985645/.

Ximena Gutierrez-Vasques and Víctor Germán Mijangos De La Cruz. De las ideas verdes incoloras hasta ChatGpt: los grandes modelos del lenguaje. *TIES, Revista de Tecnología e Innovación en Educación Superior*, (10):12–23, June 2024. ISSN 2683-2968. doi: 10.22201/dgtic.26832968e.2024.10.18. URL https://www.ties.unam.mx/index.php/ties/article/view/18. Publisher: Universidad Nacional Autonoma de Mexico.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring Massive Multitask Language Understanding. January 2021. doi: 10.48550/arXiv.2009.03300. URL http://arxiv.org/abs/2009.03300. arXiv:2009.03300 [cs].

Laura S. Hernández Gutiérrez, Juan Andrés Trejo, and Yolanda Marín Campos. Diseño de un ECOE para evaluar habilidades clínicas en neurología en estudiantes del quinto año. *Investigación en Educación Médica*, 6(24):248–254, October 2017. ISSN 2007-5057, 2007-865X. doi: 10.1016/j.riem.2017.01.002. URL http://riem.facmed.unam.mx/index.php/riem/article/view/212. Publisher: Universidad Nacional Autonoma de Mexico.

huongnguyen253. DeepSeek R1 Emerges as Low-Cost Challenger in Global AI Race. *Eastgate Software*, April 2025. URL https://eastgate-software.com/deepseek-r1-emerges-as-low-cost-challenger-in-global-ai-race/. Section: Tech Enthusiast.

Philip C. Jarrett, Jared Hill, Marshall Howell, Kristen Grabow Moore, Joby J. Thoppil, Laura Vargas Ortiz, Samuel T. Parnell, D. Mark Courtney, Samuel A. McDonald, Deborah B. Diercks, Andrew R. Jamieson, and Dazhe Cao. Temperature-Driven Variability in Emergency Diagnostic Accuracy by a Leading Language Model. June 2025. doi: 10.1101/2025.06.04.25328288. URL http://medrxiv.org/lookup/doi/10.1101/2025.06.04.25328288.

Yudai Kaneda, Ryo Takahashi, Uiri Kaneda, Shiori Akashima, Haruna Okita, Sadaya Misaki, Akimi Yamashiro, Akihiko Ozaki, and Tetsuya Tanimoto. Assessing the Performance of GPT-3.5 and GPT-4 on the

2023 Japanese Nursing Examination. *Cureus*, August 2023. ISSN 2168-8184. doi: 10.7759/cureus.42924. URL https://www.cureus.com/articles/173248-assessing-the-performance-of-gpt-35-and-gpt-4-on-the-2023-japanese-nursing-examina Publisher: Springer Science and Business Media LLC.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large Language Models are Zero-Shot Reasoners. 2022. doi: 10.48550/ARXIV.2205.11916. URL https://arxiv.org/abs/2205.11916. Version Number: 4.

Gerd Kortemeyer, Julian Nöhl, and Daria Onishchuk. Grading Assistance for a Handwritten Thermodynamics Exam using Artificial Intelligence: An Exploratory Study. June 2024. doi: 10.48550/arXiv.2406.17859. URL http://arxiv.org/abs/2406.17859. arXiv:2406.17859 [physics].

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, May 2015. ISSN 0028-0836, 1476-4687. doi: 10.1038/nature14539. URL https://www.nature.com/articles/nature14539. Publisher: Springer Science and Business Media LLC.

Jinsook Lee, Yann Hicke, Renzhe Yu, Christopher Brooks, and René F. Kizilcec. The life cycle of large language models in education: A framework for understanding sources of bias. *British Journal of Educational Technology*, 55(5):1982–2002, September 2024. ISSN 0007-1013, 1467-8535. doi: 10.1111/bjet.13505. URL https://bera-journals.onlinelibrary.wiley.com/doi/10.1111/bjet.13505. Publisher: Wiley.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks, April 2021. URL http://arxiv.org/abs/2005.11401. arXiv:2005.11401 [cs].

Huan Ma, Changqing Zhang, Yatao Bian, Lemao Liu, Zhirui Zhang, Peilin Zhao, Shu Zhang, Huazhu Fu, Qinghua Hu, and Bingzhe Wu. Fairness-guided Few-shot Prompting for Large Language Models. 2023. doi: 10.48550/ARXIV.2303.13217. URL https://arxiv.org/abs/2303.13217. Version Number: 3.

Roy Malhotra. What is Meta LLAMA 3: The Ultimate Language Model Guide. *ValueCoders | Unlocking the Power of Technology: Discover the Latest Insights and Trends*. URL https://www.valuecoders.com/blog/ai-ml/what-is-meta-llama-3-large-language-model/.

L. Rebeca Mateos Morfín and Carlos J. Flores Aguirre. Frecuencia y tipos de retroalimentación sobre la precisión del responder en una tarea de discriminación condicional. *Revista Mexicana de Análisis de la Conducta*, December 2022. ISSN 0185-4534, 2007-0802. doi: 10.5514/rmac.v48.i2.84463. URL https://www.revistas.unam.mx/index.php/rmac/article/view/84463. Publisher: Sociedad Mexicana de Analisis de la Conducta.

Adam C. Mater and Michelle L. Coote. Deep Learning in Chemistry. *Journal of Chemical Information and Modeling*, 59(6):2545–2559, June 2019. ISSN 1549-9596, 1549-960X. doi: 10.1021/acs.jcim.9b00266. URL https://pubs.acs.org/doi/10.1021/acs.jcim.9b00266. Publisher: American Chemical Society (ACS).

Yiqun Miao, Yuan Luo, Yuhan Zhao, Jiawei Li, Mingxuan Liu, Huiying Wang, Yuling Chen, and Ying Wu. Performance of GPT-4 on Chinese Nursing Examination: Potentials for AI-Assisted Nursing Education Using Large Language Models. *Nurse Educator*, 49(6):E338–E343, November 2024. ISSN 1538-9855, 0363-3624. doi: 10.1097/nne.0000000000001679. URL https://journals.lww.com/10.1097/NNE.0000000000001679. Publisher: Ovid Technologies (Wolters Kluwer Health).

Ryan Mok, Faraaz Akhtar, Louis Clare, Christine Li, Jun Ida, Lewis Ross, and Mario Campanelli. Using AI Large Language Models for Grading in Education: A Hands-On Test for Physics. November 2024. doi: 10.48550/arXiv.2411.13685. URL http://arxiv.org/abs/2411.13685. arXiv:2411.13685 [physics].

Douglas C. Montgomery. *Design and Analysis of Experiments*. John Wiley Sons, New York, NY, 5th edition, 2001.

Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. Capabilities of GPT-4 on Medical Challenge Problems. 2023. doi: 10.48550/ARXIV.2303.13375. URL https://arxiv.org/abs/2303.13375. Version Number: 2.

Andrijana Pavlova, Branislav Gerazov, and Anabela Barreiro. Large Language Models and OpenLogos: An Educational Case Scenario. *Open Research Europe*, 4:110, June 2024. ISSN 2732-5121. doi: 10.12688/openreseurope.17605.1. URL https://open-research-europe.ec.europa.eu/articles/4-110/v1. Publisher: F1000 Research Ltd.

Majdi Quttainah, Vinaytosh Mishra, Somayya Madakam, Yotam Lurie, and Shlomo Mark. Cost, Usability, Credibility, Fairness, Accountability, Transparency, and Explainability Framework for Safe and Effective Large Language Models in Medical Education: Narrative Review and Qualitative Study. *JMIR AI*, 3:e51834, April 2024. ISSN 2817-1705. doi: 10.2196/51834. URL https://ai.jmir.org/2024/1/e51834. Publisher: JMIR Publications Inc.

Mahlatse Ragolane, Shahiem Patel, and Pranisha Salikram. AI Versus Human Graders: Assessing the Role of Large Language Models in Higher Education. *Asian Journal of Education and Social Studies*, 50(10):244–263, October 2024. ISSN 2581-6268. doi: 10.9734/ajess/2024/v50i101616. URL https://journalajess.com/index.php/AJESS/article/view/1616. Publisher: Sciencedomain International.

Avita Rath. Empowering Future Dentists: A Comprehensive Mixed-Methods Exploration of Artificial Intelligence in Personalizing Year 3 Clinical Dental Practice Education. *Journal of Dental Education*, May 2025. ISSN 0022-0337,

1930-7837. doi: 10.1002/jdd.13939. URL https://onlinelibrary.wiley.com/doi/10.1002/jdd.13939. Publisher: Wiley.

Jimmy Ronald Riojas Rivera, Marilyn Aurora Buendía Molina, and Jorge Ysaac Angles Camacho. The management of the director and teachers' feedback, in a public university of Peru. In *Proceedings of the 21th LACCEI International Multi-Conference for Engineering, Education and Technology (LACCEI 2023): "Leadership in Education and Innovation in Engineering in the Framework of Global Transformations: Integration and Alliances for Integral Development"*. Latin American and Caribbean Consortium of Engineering Institutions, 2023. doi: 10.18687/laccei2023.1.1.1170. URL https://laccei.org/LACCEI2023-BuenosAires/meta/FP1170.html.

Kylie Robison. Why everyone is freaking out about DeepSeek. *The Verge*, January 2025. URL https://www.theverge.com/ai-artificial-intelligence/598846/deepseek-big-tech-ai-industry-nvidia-impac.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. WinoGrande: An Adversarial Winograd Schema Challenge at Scale. November 2019. doi: 10.48550/arXiv.1907.10641. URL http://arxiv.org/abs/1907.10641. arXiv:1907.10641 [cs].

Jürgen Schmidhuber. Deep learning in neural networks: An overview. *Neural Networks*, 61:85–117, January 2015. ISSN 0893-6080. doi: 10.1016/j.neunet.2014.09.003. URL https://linkinghub.elsevier.com/retrieve/pii/S0893608014002135. Publisher: Elsevier BV.

Sharmin Söderström and Torulf Palm. Feedback in mathematics education research: a systematic literature review. *Research in Mathematics Education*, pages 1–22, September 2024. ISSN 1479-4802, 1754-0178. doi: 10.1080/14794802.2024.2401488. URL https://www.tandfonline.com/doi/full/10.1080/14794802.2024.2401488. Publisher: Informa UK Limited.

J. Tamayo, P. D. V. Navarrete, R. D. T. Riofrio, and L. M. A. Pardo. La ética en la educación superior: abordando desafíos y oportunidades para el aprendizaje inclusivo. *Reincisol.*, 3:890–907, 2024. doi: 10.59282/reincisol.v3(5)890-907.

Pittawat Taveekitworachai, Febri Abdullah, Mury F. Dewantoro, Ruck Thawonmas, Julian Togelius, and Jochen Renz. ChatGPT4PCG Competition: Character-like Level Generation for Science Birds. 2023. doi: 10.48550/ARXIV.2303.15662. URL https://arxiv.org/abs/2303.15662. Version Number: 3.

Gemini Team, Petko Georgiev, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. December 2024. doi: 10.48550/arXiv.2403.05530. URL http://arxiv.org/abs/2403.05530. arXiv:2403.05530 [cs].

Gina Exivia Valencia Mendoza, Rocío Del Lourdes Barragán Merino, Susana Carolina Ledesma Trujillo, and Patricia Moraima Peña. Impacto de la inteligencia artificial generativa en la creatividad de los estudiantes universitarios.

Technology Rain Journal, 3(1):e33, May 2024. ISSN 2953-464X. doi: 10.55204/trj.v3i1.e33. URL https://technologyrain.com.ar/index.php/trj/article/view/33. Publisher: Puerto Madero Editorial Académica.

Tong Wan and Zhongzhou Chen. Exploring generative AI assisted feedback writing for students' written responses to a physics conceptual question with prompt engineering and few-shot learning. *Physical Review Physics Education Research*, 20(1), June 2024. ISSN 2469-9896. doi: 10.1103/physrevphyseducres.20.010152. URL https://link.aps.org/doi/10.1103/PhysRevPhysEducRes.20.010152. Publisher: American Physical Society (APS).

Tong Wan and Zhongzhou Chen. Grading explanations of problem-solving process and generating feedback using large language models at human-level accuracy | Phys. Rev. Phys. Educ. Res. 2025. URL https://journals.aps.org/prper/abstract/10.1103/PhysRevPhysEducRes.21.010126.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-Consistency Improves Chain of Thought Reasoning in Language Models. March 2023. doi: 10.48550/arXiv.2203.11171. URL http://arxiv.org/abs/2203.11171. arXiv:2203.11171 [cs].

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. 2022.

Paul Windisch, Fabio Dennstädt, Carole Koechli, Christina Schröder, Daniel M Aebersold, Robert Förster, and Daniel R Zwahlen. The Impact of Temperature on Extracting Information From Clinical Trial Publications Using Large Language Models. *Cureus*, December 2024. ISSN 2168-8184. doi: 10.7759/cureus.75748. URL https://www.cureus.com/articles/326495-the-impact-of-temperature-on-extracting-information-from-clinical-trial-publications-using-lar
Publisher: Springer Science and Business Media LLC.

Sri Yamtinah, Dimas Gilang Ramadhani, Antuni Wiyarsi, Hayuni Retno Widarti, and Ari Syahidul Shidiq. Leveraging Generative AI for Automatic Scoring in Chemistry Education: A Web Based Approach to Assessing Conceptual Understanding of Colligative Properties. *International Conference on Computers in Education*, November 2024. ISSN 3078-4360. doi: 10.58459/icce.2024.4983. URL https://library.apsce.net/index.php/ICCE/article/view/4983. Publisher: Asia-Pacific Society for Computers in Education.

Lixiang Yan, Lele Sha, Linxuan Zhao, Yuheng Li, Roberto Martinez-Maldonado, Guanliang Chen, Xinyu Li, Yueqiao Jin, and Dragan Gašević. Practical and ethical challenges of large language models in education: A systematic scoping review. *British Journal of Educational Technology*, 55(1):90–112, January 2024. ISSN 0007-1013, 1467-8535. doi: 10.1111/bjet.13370. URL https://bera-journals.onlinelibrary.wiley.com/doi/10.1111/bjet.13370. Publisher: Wiley.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi.

HellaSwag: Can a Machine Really Finish Your Sentence? May 2019. doi: 10.48550/arXiv.1905.07830. URL http://arxiv.org/abs/1905.07830. arXiv:1905.07830 [cs].

Apéndice A

Test

A.1. PREGUNTA 2 - Certamen 2, Álgebra 1 (2024)

Considere la función $f : [-2, 2] \to \mathbb{R}$ definida por $f(x) = -1 + \sqrt{4 - x^2}$.

1. Determine el dominio de f.
2. Determine si f es inyectiva. Justifique.
3. Sea $g : [-2, 0] \to [-1, 1]$ una función definida por $g(x) = -1 + \sqrt{4 - x^2}$. Demuestre que g es inyectiva.
4. Demuestre que g es sobreyectiva.
5. Determine la función inversa de g, $g^{-1}(y)$.

A.2. Rúbrica para evaluar la calidad del Feedback Escrito

Criterios	**Muy Bueno**	**Bueno**	**Suficiente**	**Insuficiente**
Claridad	El feedback es claramente redactado, con un lenguaje accesible y sin términos técnicos innecesarios. La estructura es coherente y fácil de seguir.	El feedback es claro, pero contiene algunos términos técnicos que podrían confundir al estudiante. La redacción es adecuada.	El feedback es mayormente comprensible, pero presenta confusión en algunas partes debido a la redacción o al uso de términos técnicos innecesarias.	El feedback es confuso, mal redactado y contiene numerosos términos técnicos que dificultan la comprensión.
Especificidad	Detalla con precisión los aciertos y errores del estudiante, señalando exactamente dónde éstos se encuentran en el trabajo y cómo deberían haberse abordado.	Señala los errores y aciertos, pero no siempre de manera detallada o específica. Algunas explicaciones podrían ser más claras.	Menciona los errores y aciertos, pero de manera general, sin detalles específicos sobre dónde ocurrieron.	No identifica claramente los errores o aciertos. El feedback es vago o general.
Feedback Constructivo	Ofrece un equilibrio perfecto entre elogios y críticas constructivas, proporcionando tanto refuerzos positivos como sugerencias de mejora concretas.	El feedback es mayormente constructivo, pero podría mejorar en el equilibrio entre críticas y elogios. Las sugerencias de mejora son útiles, pero no siempre concretas.	El feedback es más crítico que constructivo, y las sugerencias de mejora son vagas o insuficientes. El refuerzo positivo es mínimo.	El feedback es predominantemente negativo, sin sugerencias claras de mejora ni refuerzos positivos.

Criterios	**Muy Bueno**	**Bueno**	**Suficiente**	**Insuficiente**
Se explicita Desempeño Esperado	El feedback comienza explicitando claramente lo que se esperaba en la pregunta, proporcionando una explicación completa y precisa sobre cómo debería haberse abordado el problema.	Explica lo que se esperaba en la pregunta, pero de manera más general o con menos detalles.	Menciona el desempeño esperado, pero de manera vaga o poco clara.	No explicita el desempeño esperado en la pregunta.
Reconocer respuestas correctas del desempeño del estudiante	Se explicita y reconoce lo qué es lo que está correcto en la respuesta con detalles precisos y evidencias claras. Se refuerza positivamente al estudiante.	Reconoce lo que está correcto en la respuesta, pero de manera menos detallada o con menor énfasis.	Menciona lo que está correcto, pero de forma genérica o sin profundizar.	No se reconoce lo que está correcto en la respuesta, o el reconocimiento es mínimo.
Mostrar Errores del desempeño del estudiante	Señala con precisión y claridad los errores presentes en la respuesta, explicando dónde se cometieron y cómo afectaron el resultado.	Identifica los errores, pero la explicación es menos detallada o específica.	Menciona los errores de manera general, sin una explicación clara o detallada.	No identifica claramente los errores, o los menciona de manera vaga o inexacta.

Criterios	**Muy Bueno**	**Bueno**	**Suficiente**	**Insuficiente**
Consejos para Mejorar	Proporciona estrategias claras y detalladas para mejorar, incluyendo consejos específicos y recursos adicionales que guíen al estudiante.	Ofrece consejos para mejorar, pero podrían ser más detallados o específicos.	Sugiere mejoras, pero de manera general o con consejos insuficientes.	No ofrece consejos claros para mejorar, o estos son vagos y poco útiles.
Feedback respetuoso	El feedback es empático, evitando símbolos que puedan desmotivar al estudiante. Termina con una frase motivacional que refuerza la confianza del estudiante.	El feedback es mayormente empático, pero podría mejorar en la forma de expresar las críticas. Incluye una frase motivacional.	El feedback es neutral en tono, con poca consideración por el impacto emocional. La motivación es limitada.	El feedback es frío o negativo, utilizando símbolos o lenguaje que pueden ser percibidos como castigadores. No incluye motivación.
Extensión y Complitud del feedback	El feedback es breve pero completo, abordando todos los aspectos necesarios para que el estudiante comprenda y aprenda de sus errores sin abrumarlo.	El feedback es completo, pero podría ser más conciso. Abarca los aspectos principales, pero algunos detalles podrían haberse omitido.	El feedback es adecuado en extensión, pero deja fuera algunos aspectos importantes o se extiende innecesariamente.	El feedback es insuficiente en detalle, omite aspectos clave o es excesivamente extenso, abrumando al estudiante.

A.3. Prompts utilizados

En este apéndice se presentan los prompts completos diseñados para el Análisis Técnico y Generación de Feedback Personalizado. La estructura de los prompts implementa una metodología inspirada en CoT, guiando al modelo a través de un proceso de razonamiento por pasos.

A.3.1. Prompt para el Análisis Técnico

```
ROL: Eres un asistente experto en evaluacion de Algebra I.
   Tienes TRES tareas que cumplir, siguiendo un proceso paso
    a paso.

PRIMERA TAREA (Identificacion):
Accion: Identifica y analiza cada paso que el estudiante
   realiza.
Formato: Escribe estos pasos dentro de las etiquetas: <steps
   > y </steps>.

SEGUNDA TAREA (Revision):
Accion: Escribe explicitamente la pauta de evaluacion y
   comparala con la respuesta del estudiante, se alando
   errores y aciertos.
Pauta de solucion: <solution>{pauta}</solution>
Formato: Escribe tu revision en Markdown dentro de: <
   revision> y </revision>.

TERCERA TAREA (Evaluacion):
Accion: Usando tu revision, evalua el desempeno segun los
   criterios de la rubrica. Menciona el puntaje maximo y el
   obtenido, y luego el total.
Rubrica: <guideline>{rubrica}</guideline>
Formato: Escribe tu evaluacion en Markdown dentro de: <
   evaluation> y </evaluation>.

INSTRUCCIONES ADICIONALES:
- Ecuaciones en formato LaTeX ($ o $$).
- Escribe el nombre de cada criterio de evaluacion.

```

```
RESPUESTA DEL ESTUDIANTE A EVALUAR:
<student>{student}</student>
```

Listing A.1: Prompt para el Análisis Técnico.

A.3.2. Prompt para la Generación de Feedback Personalizado

```
ROL: Eres un tutor de matematicas empatico y pedagogico.
TAREA: Proporcionar un feedback constructivo basandote en
   este analisis tecnico:
<student>{student}</student>

PASO 1 (Analisis y Planificacion - Chain of Thought):
Accion: Revisa el analisis tecnico, identifica fortalezas,
   debilidades y errores.
Piensa en recomendaciones y areas de mejora.
Formato: Escribe tu planificacion interna entre <thinking> y
    </thinking>.

PASO 2 (Redaccion del Feedback):
Accion: Basandote en tu planificacion, redacta el feedback
   para el estudiante.
Formato: Escribe el feedback final entre <feedback> y </
   feedback>.

INSTRUCCIONES ADICIONALES PARA EL FEEDBACK:
1. Abordaje Individual: Proporciona comentarios especificos
   para cada criterio.
2. Tono Constructivo: Comienza con fortalezas, luego errores
    y recomendaciones.
3. Formato Matematico: Ecuaciones en formato LaTeX ($ o $$).
```

Listing A.2: Prompt para la generación de Feedback Personalizado.