UNIVERSIDAD CATÓLICA DEL MAULE
FACULTAD DE CIENCIAS DE LA INGENIERÍA
ESCUELA INGENIERÍA CIVIL INFORMÁTICA

OPTIMIZACIÓN DE MODELOS DE LANGUAJE PARA LA PRUEBA DE ACCESO A LA EDUCACIÓN SUPERIOR (PAES) DE MATEMÁTICAS EN CHILE: UNA PRUEBA DE CONCEPTO

OMAR FRANCISCO JAVIER OLIVARES URRUTIA
Profesor Guía:
Xaviera A. Lopez Cortes

Proyecto de título para optar al Título Profesional de Ingeniero Civil Informático.

TALCA, FEBRERO 2024

UNIVERSIDAD CATÓLICA DEL MAULE
FACULTAD DE CIENCIAS DE LA INGENIERÍA
ESCUELA INGENIERÍA CIVIL INFORMÁTICA

Proyecto de título para optar al Título Profesional de Ingeniero Civil Informático.

OPTIMIZACIÓN DE MODELOS DE LANGUAJE PARA LA PRUEBA DE ACCESO A LA EDUCACIÓN SUPERIOR (PAES) DE MATEMÁTICAS EN CHILE: UNA PRUEBA DE CONCEPTO

OMAR FRANCISCO JAVIER OLIVARES URRUTIA

COMISIÓN EXAMINADORA	**FIRMA**
Profesor de comisión interna Dr. Sergio I. Hernández Alvarez Depto. Ciencias de la Ingeniería	_____
Profesor de comisión externa Dr. Maria D. Aravena Diaz Depto. de Ciencias Básicas	_____
Profesor Guía Dr.Xaviera A. Lopez Cortes Depto. Ciencias de la Ingeniería	_____
NOTA FINAL EXAMEN DE TÍTULO	_____

TALCA, ENERO 2024

Agradecimientos

Me gustaría agradecer a todos los que hicieron posible este proyecto aportando su tiempo y espacio, para materializar esta idea de ser Ingeniero aún cuando parecía una utopía demasiado tardía.

Gracias a ustedes fue posible.

Omar Olivares Urrutia

Enero, 2024

Abstract

The project endeavors to execute fine-tuning, and customization of Large Language Models, particularly ChatGPT (2023), with the objective of refining logical-mathematical reasoning, formative feedback mechanisms, and the generation of mathematical content pertinent to the PAES in Chile. The primary goal is to augment the quality of automated instructional processes, the generation of technical-mathematical content, and to mitigate errors and/or hallucinations in academic tutoring. This is achieved through the meticulous implementation of re-trained LLMs tailored specifically for this task. The re-trained models will be fed a dataset derived from recorded classes, wherein instructors systematically elucidate the step-by-step resolution of mathematical problems featured in the PAES. This process aligns with contemporary fine-tuning methodologies (2022; 2017; 2022; 2020), incorporating Reinforcement Learning techniques (2018), quantization strategies (2021), and efficient parameter fine-tuning protocols (2023) designed to curtail computational costs. The methodological framework encompasses the compilation of a dataset through transcriptions of recorded classes encompassing PAES-related questions and answers. Data preprocessing involves the utilization of robust speech recognition models (2022) and optical character recognition incorporating mathematical notation (2023). The efficacy of class transcriptions and the fidelity of the final mathematical notation are ensured through inference leveraging efficient implementations of open-source models (2023a, 2023b). Subsequent phases involve an exhaustive analysis of the collected data, fine-tuning of the aforementioned LLMs, and an assessment of their logical-mathematical prowess through a control PAES, absent from the initial dataset. This study encompasses diverse phases of experimentation within the domain of language models, aimed at assessing their applicability and efficacy in the context of the Mathematics Entrance Exam (PAES) in Chile. The results obtained reveal a substantial quantitative enhancement in the performance of language models subsequent to a targeted fine-tuning process. The baseline competence of the GPT-4 model, prior to optimization, manifested as a score of 0,78, whereas the fine-tuned iteration demonstrated a significantly elevated score of 0,94. These outcomes not only attest to the technical feasibility of the proposed methodology but also foreshadow a pronounced amelioration in the quality and specificity of the generated content, thereby influencing the educational landscape through artificial intelligence. The anticipated outcome of this project encompasses substantive enhancements in the quality and precision of automated instruction for the PAES in mathematics. The development of this refined model holds the potential to provide indispensable support for students during their exam preparation, markedly reducing errors associated with personalized feedback facilitated by LLMs, tutoring, and the generation of technical mathematical study content. Furthermore, this initiative harbors the potential to contribute to narrowing the gap in access to pre-university education by diminishing comparative costs of existing pre-university programs.

Resumen

El proyecto propone realizar *fine-tuning* y personalizar Modelos de Lenguaje (LLMs), específicamente ChatGPT (2023) para mejorar el razonamiento lógico-matemático, retroalimentación formativa y generación de contenidos de matemáticas en el contexto de la Prueba de Acceso a la Educación Superior (PAES) en Chile. El objetivo principal es incrementar la calidad en la instrucción automatizada, la generación de contenidos técnico-matemáticos y minimizar los errores y/o halucinaciones en la tutoría académica mediante la adecuada implementación de los LLMs re-entrenados para esta tarea específica. Estos modelos re-entrenados serán alimentados con un base de datos recopilada a partir clases grabadas de profesores explicando paso a paso cómo resolver problemas matemáticos de la PAES, siguiendo las técnicas de fine-funning actuales (2022; 2017; 2022; 2020) usando Reinforcement Learning (2018), técnicas de cuantización (2021) y fine-tunning de parámetros eficientes (2023) para reducir el costo computacional. La metodología implica la recopilación de un *dataset* a partir de transcripciones de clases grabadas relacionadas con preguntas y respuestas de la PAES, y su debido preprocesamiento de datos, mediante el uso de modelos robustos de reconocimiento de voz (2022) y reconocimiento óptico de texto con notación matemática (2023). La eficiencia en la transcripción de las clases y la calidad de la notación matemática final se asegurarán realizando inferencia sobre implementaciones eficientes de modelos open source (2023a, 2023b). Posteriormente, se llevará a cabo un análisis en profundidad de los datos recopilados, fine-tuning de los LLMs mecionados, y evaluación de su rendimiento lógico-matemático mediante una PAES de control actual. En el presente estudio se ejecutaron diversas fases de experimentación en el ámbito de los modelos de lenguaje, conducentes a evaluar su idoneidad y operatividad en el contexto de la Prueba de Acceso a la Educación Superior (PAES) de Matemáticas en Chile. Los resultados obtenidos reflejan una notable mejora cuantitativa en el rendimiento de los modelos de lenguaje tras el proceso de afinamiento específico. El modelo GPT-4, previo a la optimización, evidenció una competencia base con un puntaje de 0,78, mientras que el modelo *finetuned* alcanzó un significantemente superior puntaje de 0,94. Estos resultados no solo manifiestan la factibilidad técnica del enfoque propuesto, sino que también auguran una relevante mejora en la calidad y especificidad de los contenidos generados y en la experiencia formativa manipulada por inteligencia artificial. Se pronostican mejoras significativas en la calidad y precisión de la instrucción automatizada para la PAES de matemáticas. El desarrollo de este modelo mejorado aportará potencialmente un apoyo indispensable para los estudiantes durante su preparación para la prueba, con la capacidad de reducir drásticamente los errores asociados a la retroalimentación personalizada usando LLMs, la tutoría y la generación de contenido matemático de estudio. Este proyecto también tiene el potencial de contribuir a reducir la brecha en el acceso a la educación pre-universitaria, a través de la disminución de los costos comparativos con los preuniversitarios actuales.

Índice general

Índice de Figuras

Índice de Tablas

Capítulo 1

Introducción

La Inteligencia Artificial (Russell, 2010), el Deep Learning (Goodfellow et al., 2016; LeCun et al., 2015) y en específico los Modelos de Lenguaje (C. Zhou et al., 2023a) (*Large Language Models* ó LLMs, por sus siglas en Inglés) con miles de millones de parámetros han demostrado soprendentes capacidades de Procesamiento de Lenguaje Natural (*Natural Language Processing*) mostrando un gran nivel de generalización en sus modelos comerciales tales como GPT-4 de OpenAI (Bubeck et al., 2023), Gemini de Google (Akter et al., 2023), Claude-2 de Anthropic (Anthropic, 2023) e Inflection-1 (*Inflection-1: Pi's Best-in-Class LLM*, 2023) de la epónima compañía. Sin embargo, también los modelos *open source* de significativo menor número de parámetros como Llama-2 de Meta (Touvron, Martin, et al., 2023), Orca 2 (Mitra et al., 2023) y Phi-1.5 de Microsoft (Li et al., 2023), entre otros modelos de investigación (Eldan & Li, 2023; Gunasekar et al., 2023) han demostrado capacidades generativas sorprendentes para sus comparativos tamaños, señalando que los modelos de código abierto podrían llegar a competir con sus contrapartes comerciales a una fracción del costo. Los LLMs actuales permiten con gran robustez generar texto creativo tales como ha demostrado Bubeck et al. (2023), resolver problemas matemáticos con gran precisión (Lightman et al., 2023) y responder preguntas en múltiples dominios del conocimiento abarcando problemas de lingüística, desarrollo infantil, matemáticas, razonamiento de sentido común, biología, física, sesgo social, desarrollo de software y entre otras habilidades (Srivastava et al., 2023; Suzgun et al., 2022). Estas aplicaciones representan ejemplos claros del significativo potencial del beneficio que la Inteligencia Artificial puede ofrecer a millones de personas, además de la gran demanda comercial y la proliferación de cientos de LLMs (Zhao et al., 2023) con diversas aplicaciones. Sin embargo, la resolución de problemas complejos y razonamiento lógico de múltiples pasos sigue siendo uno de los desafios que presentan los LLMs actuales, incluso los modelos más avanzados aún producen errores básicos de razonamiento (J. Huang et al., 2023; Wu et al., 2023), lógica (Berglund et al., 2023) y halucinaciones (Kaddour et al., 2023; Rawte et al., 2023), especialmente en contextos matemáticos (Hendrycks et al., 2021; Lewkowycz et al., 2022), y existe un gran interés de desarrollar técnicas que mejoren su desempeño en las áreas de razonamiento, lógica y matemáticas, utilizando principalmente estrategias de *fine-tunning* para tareas específicas y *prompt engineering* tales como Chain-of-Thought (Wei et al., 2023), Self-Consistency (Wang et al., 2023b), ReAct (Tyen et al., 2023), entre otros.

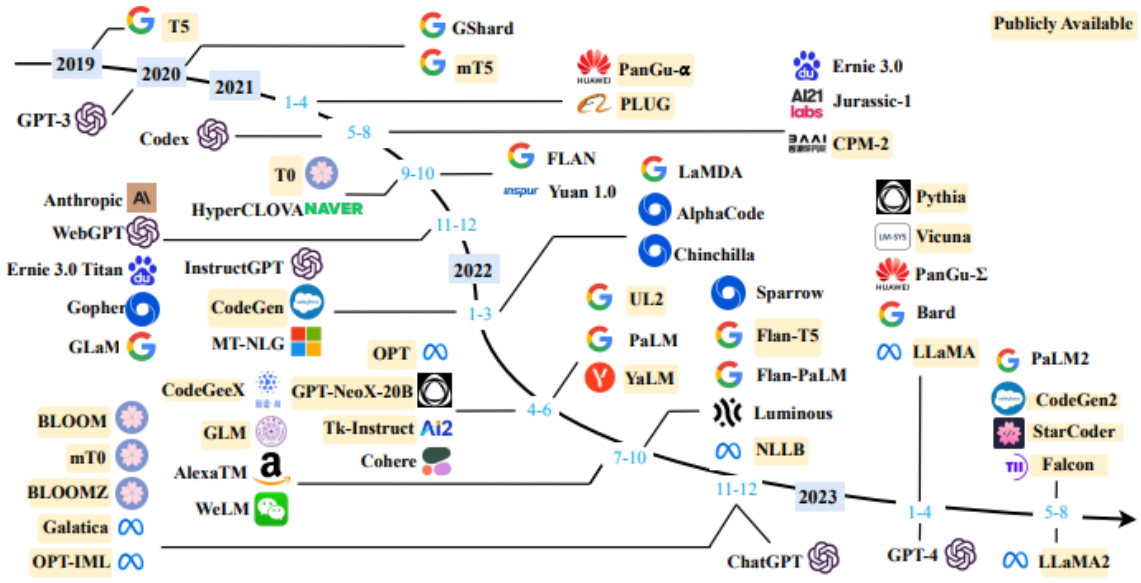


Figura 1.1: *Línea de tiempo de LLMs existentes (con un tamaño mayor a 10 mil millones de parámetros) en los últimos años, mostrando la proliferación de LLMs* (Zhao et al., 2023).

Capítulo 2

Problema de investigación y relevancia

Cada año, cerca de 250.000 estudiantes realizan la Prueba de Acceso a la Educación Superior (PAES) en Chile (DEMRE, 2022). Las universidades y centros de formación técnica utilizan los puntajes de esta prueba para seleccionar a sus estudiantes. Debido a esto, la mayoría de los estudiantes se preparan inclusive con años de anticipación para rendir el exámen realizando pre-universitarios, clases particulares, y consultando material de estudio en línea. El aprendizaje de las Matemáticas es un desafío para muchos estudiantes, especialmente cuando se enfrentan a exámenes competitivos de alto nivel como la PAES, la cual es de carácter obligatorio para ingresar a estudiar una carrera Universitaria. Muchas veces, del resultado de esta prueba se extiende dónde los estudiantes estudiarán sus carreras universitarias. Los profesores y tutores cumplen un papel vital en este proceso de aprendizaje, pero el costo de la retroalimentación y la tutoría puede ser obstaculizador, especialmente para estudiantes menos aventajados. Con la proliferacipon de la tecnología, la Inteligencia Artificial y los Modelos de Lenguaje (LLMs), existe una oportunidad única para mejorar la calidad de la instrucción automatizada y de reducir costos. Además, de entregar herramientas con un mayor grado de personalización para ayudar a los estudiantes a optimizar su tiempo en este periodo crítico el cual muchas veces se realiza en paralelo a sus estudios de educación secundaria.

La optimización de LLMs para la instrucción y generación de contenidos de estudio de la prueba PAES de Matemáticas se vuelve un tema de relevancia porque permite mejorar la calidad de la educación Preuniversitaria a través del uso optimizado de la Inteligencia Artificial, y facilita el acceso a herramientas de aprendizaje a un costo reducido y de un mayor grado de personalización que una clase tradicional.

Capítulo 3

Marco teórico

3.1. Prueba de Acceso a la Educación Superior (PAES)

La PAES es la prueba oficial y obligatoria que el Gobierno de Chile impone a todos los estudiantes de Chile (DEMRE, 2023a). Esta prueba es utilizada por las universidades y centros de formación técnica para seleccionar a sus estudiantes. La prueba se divide en 4 áreas: Lenguaje y Comunicación, Matemáticas, Ciencias y Ciencias Sociales. La prueba de Matemáticas es la que presenta mayor dificultad para los estudiantes, y es la que se abordará en este proyecto. La prueba de matemáticas se divide en dos tipos principales: M1 y M2.

3.1.1. Matemáticas 1 (M1)

Perteneciente al proceso de admisión regular para el año académico 2024, constituye una herramienta para medir el progreso en capacidades matemáticas de relevancia universal y aplicabilidad cotidiana. Esta prueba se dirige a individuos que buscan una formación en la matemática que sea pertinente y funcional a su educación universitaria futura (DEMRE, 2023c), además que como ya se mencionó es de carácter universal. El propósito central de la prueba M1 es la evaluación de la competencia matemática, entendida como la confluencia de habilidades y saberes imprescindibles para la resolución de problemas en un amplio espectro de situaciones reales. Está alineada con los objetivos del currículo de matemáticas, centrando su atención en destrezas específicas promovidas por este.

Dentro de estas habilidades se encuentran:

- La habilidad para resolver problemas de índole matemática.
- La capacidad de modelación de fenómenos o situaciones a través de herramientas matemáticas.
- La competencia en la representación de conceptos matemáticos de manera abstracta y/o concreta.
- La aptitud para argumentar lógicamente en el contexto de las matemáticas.

Además, la Prueba M1 toma como referencia los conocimientos estipulados por el plan de formación general matemática, que abarca desde el séptimo año básico hasta el segundo año medio, conforme a las Bases Curriculares (MINEDUC, 2015) publicadas por el Ministerio de Educación de Chile. Estos conocimientos se organizan en cuatro ejes temáticos fundamentales:

- *Números*, que se refiere al entendimiento y manejo de los números y sus operaciones.
- *Álgebra y Funciones*, relacionadas con la manipulación de expresiones algebraicas y el estudio de funciones matemáticas.
- *Geometría*, que valoriza el estudio de las propiedades y relaciones espaciales.
- *Probabilidad y Estadística*, enfocando el análisis de datos y la predicción de eventos probabilísticos.

Es importante resaltar que la competencia matemática es medida mediante preguntas cuyos enunciados se contextualizan en una amplia gama de situaciones de la vida diaria, así como a través de interrogantes enfocadas exclusivamente en contextos matemáticos puros. Esta herramienta de evaluación consta de un total de 65 preguntas de elección múltiple, cada una con una única respuesta correcta de entre 4 alternativas. De estas interrogantes, 60 serán empleadas en la determinación del puntaje final que contribuirá al proceso de selección universitaria. El tiempo establecido para la realización de la prueba se mantiene en 2 horas y 20 minutos.

3.1.2. Matemáticas 2 (M2)

Pertinente al ciclo de admisión regular de 2024, tiene como objetivo evaluar el avance en las habilidades matemáticas críticas tanto para la vida cotidiana como para el estudio de varias ramas de la Ciencia. Está destinada a aquellos aspirantes a la educación superior que requieren un entendimiento conceptual avanzado y habilidades matemáticas sofisticadas para su formación académica universitaria (DEMRE, 2023b). La prueba M2 mide la Competencia Matemática mediante la valoración de la interacción entre las habilidades y conocimientos necesarios para abordar y resolver problemas dentro de diversos contextos, incluyendo aquellos de naturaleza cotidiana y científica, con un enfoque particular en las competencias enunciadas por el currículo matemático general. Para ser más precisos, la prueba M2 pone a prueba las capacidades en conformidad con las Bases Curriculares de la siguiente manera:

- Habilidad para la resolución de problemas matemáticos.
- Capacidad para modelar situaciones o fenómenos usando enfoques matemáticos.
- Competencia en representar conceptos matemáticos de forma tangible o abstracta.
- Aptitud para argumentar de manera lógica y convincente en contextos matemáticos.

El contenido de conocimiento evaluado en la prueba M2 está basado en el plan de formación general matemática que se extiende del séptimo año básico al cuarto año medio (MINEDUC, 2015, 2019), organizado en los mismos cuatro ejes temáticos cruciales identificados en la prueba M1:

- *Números*
- *Álgebra y Funciones*
- *Geometría*
- *Probabilidad y Estadística*

La evaluación M2 se compone de 55 ítems de elección múltiple, cada uno con 4 o 5 posibles respuestas, de los cuales 50 ítems son determinantes para el cálculo del puntaje de selección universitaria. Incluye también ítems que se clasifican como preguntas de Suficiencia de Datos. La duración asignada para la finalización de esta evaluación es de 2 horas y 20 minutos.

3.2. Educación personalizada

El psicólogo educativo Benjamin Bloom descubrió que la tutoría individual utilizando el aprendizaje basado en la maestría conducía a una mejora de dos sigmas (estadísicos) en el rendimiento de los estudiantes. Él pregunta en su artículo que identificó el "Problema de 2-Sigmas": ¿cómo logramos estos resultados en condiciones más prácticas (es decir, más escalables) que la tutoría uno a uno? En una línea relacionada de investigación posterior, el meta-análisis a gran escala de Ricón (2019) muestra efectos notables ($> 0{,}5d$ de Cohen) del uso de la instrucción directa con aprendizaje basado en la maestría. "*Sin embargo, a pesar del amplio cuerpo de investigación que respalda su eficacia, la instrucción directa no ha sido ampliamente aceptada o implementada.*" (2019)

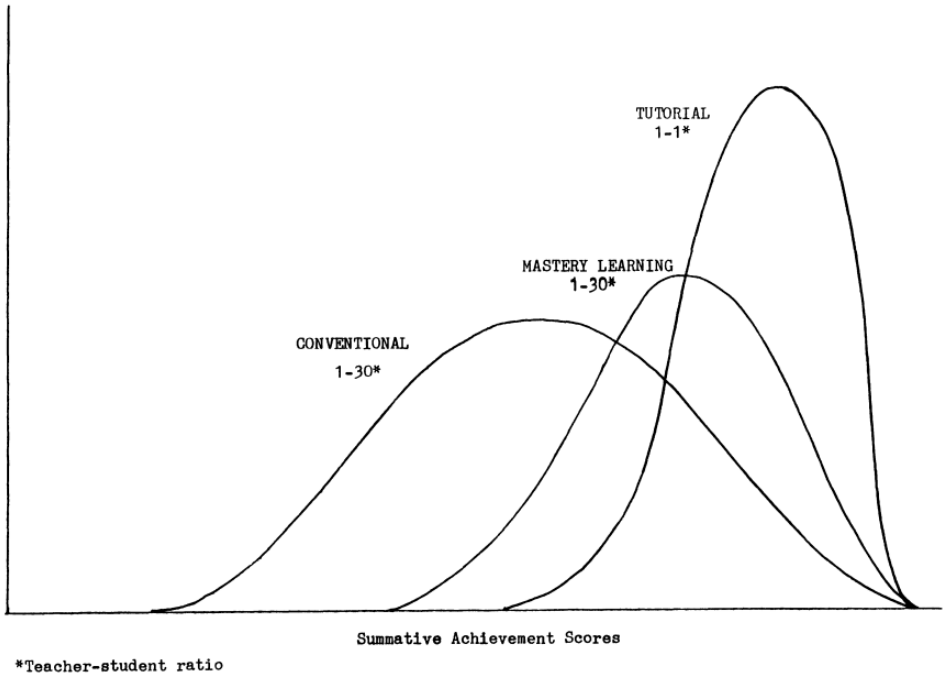


Figura 3.1: *El fenómeno del "2-Sigma", descubierto por el psicólogo educativo Benjamin Bloom (1984), revela que un estudiante promedio, cuando recibe tutoría uno a uno, puede superar el rendimiento del 98 % de los estudiantes que siguen métodos de instrucción tradicionales en el aula.*

Es posible que mediante la incorporación de tecnología en la educación logremos acercarnos a los beneficios de la tutoría individual. La Inteligencia Artificial (IA) tiene la capacidad de ofrecer una retroalimentación instantánea y personalizada, así como la habilidad de adaptarse a las particularidades de cada estudiante. Más allá, esta tecnología posee el potencial de ofrecer una retroalimentación más rica que la posible en un entorno docente tradicional, atendiendo a que puede operar en multitud de modalidades (texto, audio, video, etc.) y a diferentes niveles (tales como retroalimentación sobre el proceso de resolución de problemas y no solo sobre el resultado). Adicionalmente, la tecnología permite un suministro de retroalimentación más frecuente que el que podría propiciar un docente, lo cual resulta de gran importancia para la consolidación del aprendizaje. Investigaciones como la de Bhutoria (2022) confirman que estudiantes que recibieron retroalimentación inmediata en un entorno

de aprendizaje en línea mostraron mayor motivación y una probabilidad más elevada de finalizar el curso.

De esta forma, las herramientas digitales pueden tender un puente hacia la reducción de la brecha existente en la tutoría individualizada, en tanto que el número tradicional de estudiantes por aula constituye una limitación al aprendizaje, tal y como ya advirtiera Bloom en su trabajo fundacional. El mismo autor enfatizó que, además del aprendizaje basado en la maestría, la disminución de la cantidad de estudiantes por clase favorece el aprendizaje integral. Actualmente, los principales obstáculos son el costo y la complejidad de proveer una educación personalizada (Bhutoria, 2022). La Inteligencia Artificial puede jugar un papel decisivo en la mitigación de estas dificultades. Asimismo, utilizada en conjunción con docentes reales, quienes adoptarían funciones de facilitadores y motivadores del aprendizaje (roles que la tecnología actual no puede suplantar), puede representar una herramienta de gran valor. Además, el ahorro de tiempo al identificar las necesidades individualizadas de cada estudiante y al proporcionar retroalimentación personalizada puede convertirse en un elemento clave para el aprendizaje.

Por consiguiente, la personalización educativa emerge como un aspecto crítico para el avance del proceso de aprendizaje de los estudiantes; la tecnología se perfila como un elemento crucial para la consecución de este objetivo.

3.3. *Transformers* y *Pre-Training*

La arquitectura Transformer introducida por Vaswani et al. (2017), es un modelo de red neuronal innovador y altamente eficiente para el procesamiento del lenguaje natural (NLP). Esta arquitectura introdujo un cambio de paradigma al prescindir de las capas recurrentes y convolucionales, en favor de mecanismos de atención que permiten capturar dependencias a largo plazo y procesar secuencias de entrada en paralelo, mejorando así la escalabilidad y el rendimiento en tareas como la traducción automática, el resumen de texto y la generación de lenguaje, tal y como se ha evidenciado a través de productos comerciales como *Google Translate*, *Siri* y *ChatGPT*.

En conjunto con los métodos de *self-supervised learning* introducidos por Dai & Le (2015), específicamente, la técnica de predecir elementos subsiguientes en secuencias de datos (sin etiquetas), en conjunto con la arquitectura Transformer ha dado lugar a una nueva clase de modelos de lenguaje de gran tamaño (LLMs) que han logrado avances significativos en una amplia gama de tareas de Procesamiento de Lenguaje Natural (NLP), incluyendo la traducción automática, la generación de texto y la comprensión del lenguaje natural, mostrando una gran capacidad de generalización y transferencia de conocimiento para múltiples tareas.

La arquitectura del Transformer se divide principalmente en dos componentes estructurales: el **codificador** y el **decodificador**.

El **codificador** procesa la secuencia de tokens de entrada (por ejemplo, palabras, subpalabras o caracteres) realizando las siguientes operaciones en cada una de sus capas:

- *Self-Attention Mecanism (Multi-Head)*: Permitiendo que el modelo pese la importancia relativa de diferentes tokens en la secuencia de entrada para cada token específico. Esto

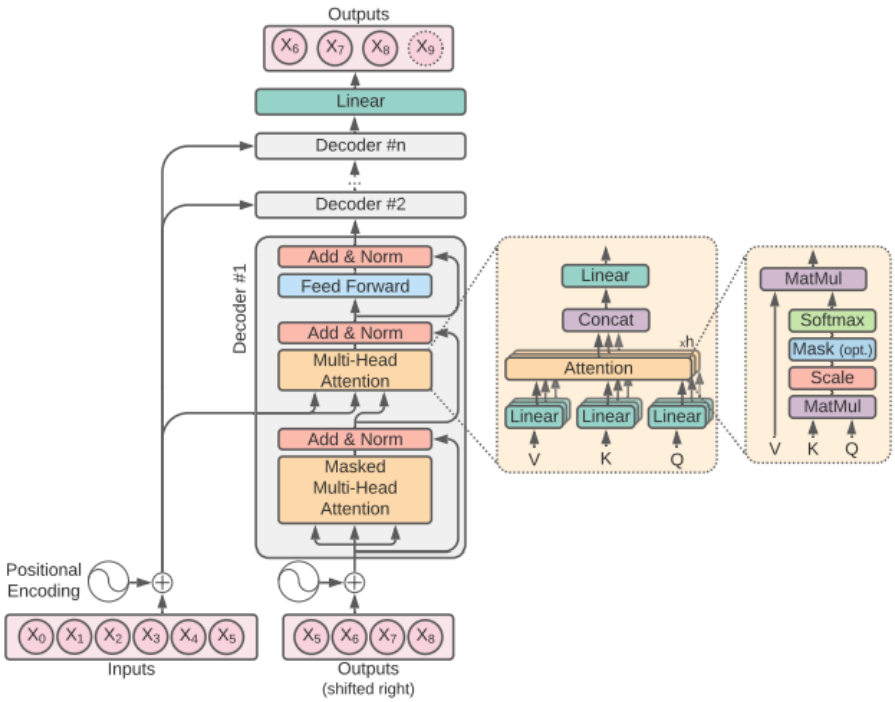


Figura 3.2: *Arquitectura Transformer (2017), se compone de dos partes principales: un codificador y un decodificador. Ilustración de* Lara-Benítez et al. (2021)

se logra a través de *Multi-Head Attention*, que emplea h cabezas de atención paralelas para capturar subespacios de información variada y proporcionar una representación compuesta más rica.

- *Positional Encoding*: A la entrada se le añade una codificación posicional que inyecta información sobre la posición de los tokens en la secuencia, lo cual es crucial ya que la operativa estándar de las capas de atención no es sensible a la secuencia (invariante al orden).
- *Pointwise Feed-Forward Networks*: Son redes neuronales completamente conectadas que aplican transformaciones lineales y no lineales para refinar las representaciones de token en cada posición.

El **decodificador** está estructurado de manera similar pero con capas adicionales de atención que permiten interactuar con la salida del codificador:

- *Masked Self-Attention*: Esta primera capa evita la "visión" del futuro en la secuencia de salida durante el entrenamiento al aplicar máscaras que impiden que cada token atienda a tokens posteriores, asegurando así una generación autoregresiva.
- *Encoder-Decoder Attention*: Las capas de atención cruzada permiten que cada posición en el decodificador condicione su salida en base a la salida completa del codificador, facilitando la alineación y el traspaso de contexto relevante.
- *Feed-Forward Networks*: Al igual que en el codificador, se aplican transformaciones lineales y no lineales a nivel de posición.

Adicionalmente, cada subcapa dentro del codificador y decodificador incluye una conexión residual seguida de capas de normalización (*Add & Norm*), lo cual contribuye a la estabilidad del entrenamiento y ayuda a combatir el problema del desvanecimiento de gradientes en redes profundas.

Finalmente, la arquitectura Transformer emplea *Softmax* en la última etapa del decodificador para la generación de tokens, asegurando una distribución de probabilidades sobre el vocabulario posible.

La arquitectura Transformer es notable por su paralelización superior y la habilidad de capturar interacciones entre palabras independientemente de su distancia relativa en la frase o documento, lo cual ha resultado en avances significativos en la eficiencia y eficacia de numerosas aplicaciones de Procesamiento de Lenguaje Natural (NLP).

3.4. Modelos de Lenguaje de Gran Escala (LLMs)

En la intersección de la inteligencia artificial y la lingüística computacional emergen los *Large Language Models* (LLMs), modelos computacionales diseñados para comprender, generar y manipular lenguaje humano a una escala sin precedentes. Estos modelos representan uno de los más significativos avances en la disciplina del Procesamiento del Lenguaje Natural (NLP, por sus siglas en inglés).

Los LLMs son arquitecturas neuronales profundas, generalmente basadas en variantes del mecanismo de atención, específicamente de los Transformers (Vaswani et al., 2017), que se entrenan en extensos corpus textuales. El entrenamiento de estos modelos implica ajustar parámetros en el orden de cientos de millones o incluso miles de millones, de modo que pueda capturar las sutilezas del idioma a partir de ejemplos de uso del lenguaje en el mundo real.

Este enfoque del aprendizaje pre-supervisado en corpus no etiquetados, seguido por un ajuste fino (*fine-tuning*) en tareas específicas, ha demostrado ser una metodología disruptiva (Radford et al., 2018). Inicialmente observado en modelos como GPT-2, los investigadores han reportado que estos sistemas empiezan a asumir tareas de NLP sin supervisión explícita, lo que implica que los modelos de lenguaje aprenden a partir de los patrones implícitos en los datos (Radford et al., 2019). Por ejemplo, condicionando un LLM a esperar una pregunta tras un bloque de texto, el modelo puede generar respuestas sorprendentemente coherentes con un rendimiento comparable al de sistemas entrenados específicamente para resolución de preguntas.

La relación entre la capacidad de los LLM y la eficacia de trasferencia de tareas sin ejemplos adicionales (*zero-shot task transfer*) es especialmente notable. Al aumentar la cantidad de parámetros, se observa una mejora casi logarítmica en la ejecución de distintas tareas, situando a modelos como GPT-3 en una posición competitiva frente a enfoques anteriores que requerían un ajuste fino y especializado por tarea (Brown et al., 2020).

GPT-3, con sus 175 mil millones de parámetros, ha marcado una era en la que modelos de lenguaje de gran escala funcionan de manera eficiente en el paradigma *few-shot*, donde el modelo realiza tareas específicas proporcionando solo unos pocos ejemplos. Este modelo de lenguaje autoregresivo se emplea sin actualizaciones de gradiente ni ajuste fino, confiando puramente en interacciones textuales para instanciar tareas y demostraciones de aprendizaje

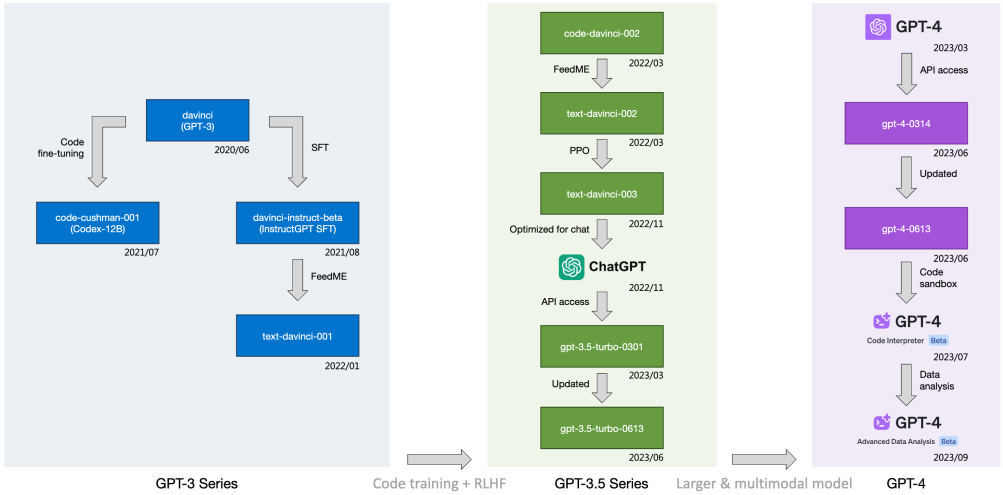


Figura 3.3: *Comparación de los modelos ChatGPT de las series 3, 3.5 y 4, resaltando el aumento en tamaño (RHLF), la incorporación de datos multimodales y el entrenamiento con código. Destaca la capacidad avanzada de la serie 4 en Code training, HRLF, mayor tamaño y análisis de datos.* (2023)

con pocos disparos.

No obstante, el despliegue de LLMs también ha puesto en relieve desafíos inherentes como las dificultades en ciertos conjuntos de datos y problemas metodológicos asociados con la dependencia de grandes corpora web (Brown et al., 2020). Además, cuestiones de sesgos y generación de información impulsiva y sin fundamento (*hallucinations*) son áreas de investigación activa.

En el contexto de la presente investigación, la optimización de LLMs para la Prueba de Acceso a la Educación Superior (PAES) de Matemáticas en Chile busca no solo capitalizar las potencialidades de los LLMs en la generación de contenidos y tutoría automatizada sino también mitigar sus limitaciones adaptando su aprendizaje a las idiosincrasias y especificidades del dominio matemático y del contexto educativo chileno. Se pretende que, por medio de prácticas rigurosas de *fine-tuning* y validación exhaustiva, los LLMs ajustados ofrezcan soporte formativo y un recurso de estudios de alta fidelidad a los aspirantes de la PAES.

3.4.1. Evolución de los Modelos de Lenguaje

La comprensión y generación del lenguaje humano siempre ha sido un desafío central dentro del campo de la inteligencia artificial. Los Modelos de Lenguaje de Gran Escala (LLMs) emergen como una respuesta poderosa a este reto, transformando la manera en que se desarrollan y emplean algoritmos de AI. Desde los primeros modelos estadísticos hasta los recientes motores basados en Transformers pre-entrenados en corpora de gran envergadura, la evolución de los LLMs ha sido ampliamente estudiada y ha experimentado un progreso notable en las últimas dos décadas (Zhao et al., 2023).

Una de las mayores revelaciones en la investigación de los LLMs es que la ampliación de la escala de los modelos conlleva significativos avances en sus capacidades (Huang et al., 2022).

Los investigadores han descubierto que, al extender la escala de parámetros más allá de cierto umbral, los LLMs no solo muestran una mejora sustancial en el rendimiento, sino que también adquieren habilidades especiales ausentes en los modelos de escala reducida. Este fenómeno ha llevado a la innovación y lanzamiento de modelos como ChatGPT de OpenAI, que ha capturado la atención tanto del ámbito académico como del público general, ejemplificando el impacto de los LLMs sobre la comunidad de AI en su totalidad.

El refinamiento de estos modelos a través de la técnica de *fine-tuning* ha demostrado ser crucial; sin embargo, un LLM puede presentar mejoras en sus capacidades de razonamiento mediante el auto-mejoramiento sin necesitar conjuntos de datos etiquetados, similar a cómo los humanos amplían sus habilidades reflexivas a través de la introspección y sin influencias externas. Este descubrimiento ha sido explorado ajustando un LLM con soluciones autogeneradas a preguntas sin etiquetar, alcanzando niveles sin precedentes de rendimiento sin la necesidad de etiquetas de verdad absoluta, y señalando la importancia de un ajuste fino enfocado en el razonamiento como base para el auto-mejoramiento (Huang et al., 2022).

La investigación contemporánea también ha subrayado la habilidad de los LLMs como aprendices *few-shot* y *zero-shot*. Con técnicas como el *chain-of-thought* (CoT), que induce al modelo a elaborar razonamientos más complejos paso a paso, se ha alcanzado desempeños notables en áreas desafiantes como las aritméticas y el razonamiento simbólico —tareas del denominado 'sistema 2' que desafían las leyes de escalamiento estándar de los LLMs (Kojima et al., 2023). Este método ha destacado las capacidades básicas y multifacéticas de los LLMs, lo que sugiere que éstos poseen una amplia gama de habilidades cognitivas generales que pueden ser extraídas con un *prompting* o consulta simple.

Además de la creación y el desarrollo de los Modelos de Lenguaje de Gran Escala (LLMs), es igualmente importante poder evaluar de forma detallada y fiable lo que estos modelos pueden hacer y dónde pueden tener dificultades. En respuesta a esta necesidad, han surgido conjuntos de herramientas de evaluación comprensivas, como es el caso de GPT-Fathom. Estas suites de evaluación son diseñadas para que cualquiera en la comunidad científica pueda usarlas y replicar sus resultados, garantizando así que los hallazgos sean consistentes y verificables. Permiten llevar a cabo comparaciones ordenadas y coherentes del desempeño entre varios de los modelos más avanzados de LLMs en una serie de categorías estandarizadas de pruebas (*benchmarks*).

Utilizando estas suites, como GPT-Fathom, los expertos pueden profundizar en el análisis del progreso de los modelos de lenguaje desde versiones antiguas, como GPT-3, hasta las más recientes, como GPT-4. Esto ayuda a una mejor comprensión sobre cómo han evolucionado técnicamente estos modelos y cuál ha sido el efecto de incorporar nuevos datos y técnicas de optimización. En particular, se ha evaluado el impacto de incluir datos de programación e implementar métodos avanzados de afinamiento como *"Shrink and Fine-Tune"* (SFT), que es una técnica de destilación que simplifica el proceso al copiar parámetros a modelos más pequeños y luego optimizarlos; así como *"Reinforcement Learning from Human Feedback"* (RLHF), que se refiere al afinamiento de los modelos utilizando retroalimentación humana directa. Ambas técnicas han demostrado ser valiosas para mejorar la habilidad de los LLMs en procesos de razonamiento, haciéndolos más eficientes y efectivos en estas tareas (S. Zheng et al., 2023).

Este recorrido histórico ha sido meticulosamente documentado en trabajos que proveen tanto

un marco general de la tecnología subyacente como una exploración de las ideas actuales sobre el funcionamiento y la interpretación del aprendizaje profundo y sus mecanismos neuronales, tales como la arquitectura de los Transformers (Douglas, 2023). Estos desarrollos en la esfera de los LLMs no solo han demostrado elevadas aptitudes en tareas de NLP y un potencial impresionante para el análisis multifacético, sino que también sugieren un futuro en el que los límites del aprendizaje automático y la cognición artificial puedan expandirse aún más.

3.4.2. Modelos Comerciales y *Open Source*

El desarrollo de la inteligencia artificial aplicada al procesamiento del lenguaje natural (NLP) ha experimentado una expansión notable, destacándose el surgimiento de una amplia gama de modelos tanto en el sector comercial como en la esfera de código abierto o Open Source. Estos avances han tomado la forma de implementaciones a gran escala financiadas por corporaciones con recursos significativos, así como iniciativas colaborativas de acceso público que promueven la innovación abierta y el acceso equitativo a tecnologías de vanguardia en NLP.

En el sector privado, los Modelos Comerciales avanzan la frontera de la innovación con sistemas pre-entrenados altamente sofisticados, encarnados por desarrollos como GPT-4 de OpenAI (OpenAI, 2023), que integran extensas capacidades de comprensión y generación de lenguaje. Estos modelos destacan por su habilidad para manejar eficientemente un espectro amplio de tareas de NLP, frecuentemente sin necesidad de entrenamiento adicional. La monetización de estos modelos se efectúa mediante la prestación de servicios que permiten su integración en aplicaciones diversas mediante interfaces de programación de aplicaciones (APIs), lo que optimiza procesos y facilita la emergencia de innovaciones tecnológicas demostrado por increíbles compañias emergentes (Anthropic, 2023; *Inflection-1: Pi's Best-in-Class LLM*, 2023; Jiang et al., 2023).

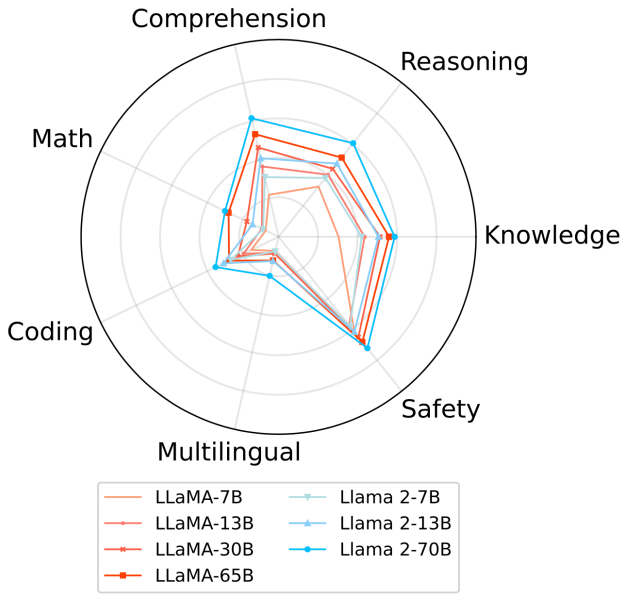


Figura 3.4: *Desempeño en varias tareas de la familia de modelos Llama, uno de los modelos de código libre más populares.* (Touvron, Martin, et al., 2023)

Paralelamente, la comunidad de código abierto aporta significativamente al campo con proyectos como LLaMA (2023) y LLaMA-2 (2023), cuyo propósito es proveer modelos de

lenguaje escalables y de alto rendimiento accesibles sin costo, impulsando así la investigación y el desarrollo a nivel global. Modelos como TinyStories (Eldan & Li, 2023) que pretende analizar qué tan pequeño debe ser un modelo para ser capaz de generar texto de forma coherente y TinyLlama (Peiyuan Zhang, 2023) que pretende entrenar un modelo totalmente de código y uso libre, simbolizan el compromiso con la transparencia y la colaboración, lo cual permite a la comunidad internacional contribuir activamente en su mejora continua y en la ampliación de las capacidades y usos del lenguaje computacionalizado.

En una fusión de enfoques, proyectos como PaLM 2 (Anil et al., 2023) tratan de equilibrar las ventajas del ámbito comercial y del espacio de código abierto, compartiendo estructuras y hallazgos con la comunidad académica y el público en general. De la misma forma, desde una perspectiva académica, han surgido enfoques que exploran el empleo de LLMs en la generación automática de contenidos basados en materiales didácticos sintéticos, evidenciando su potencial en la instrucción y en prover mayor comprensibilidad sobre el cómo operan estos modelos de lenguaje (Gunasekar et al., 2023; Li et al., 2023).

La proliferación y mejora contínua de modelos tanto comerciales como de código abierto está estimulando el progreso en el dominio de los LLMs, donde desarrollos como Orca (2023) están imponiendo nuevos patrones de referencia respecto a las funcionalidades y aplicaciones del NLP. Mucha de las ventajas de los LLMs se dice que es dado por el número de parámetros (ó el tamaño del modelo), pero impactos como los referenciados como Mitra et al. (2023) y Gunasekar et al. (2023), demuestran que también la calidad de los datos de entrenamiento es crucial para el rendimiento final de los modelos. El impacto de estas innovaciones se percibe en una vasta cantidad de industrias y ramas de investigación, inaugurando un período de avances significativos en la inteligencia artificial centrada en el lenguaje humano.

Los modelos de código abierto brindan beneficios considerables para el avance de NLP, tales como la promoción de una investigación inclusiva y colaborativa, permitiendo a investigadores independientes y a pequeñas organizaciones participar en desarrollos que de otra forma estarían fuera de su alcance, y en muchos casos creando innovaciones que han permitido disminuir los costos de operación significativamente tales como los demostrados por Gerganov (2023a). Esto ayuda a contrarrestar una tendencia preocupante hacia la concentración del poder e influencia en pocas empresas tecnológicas que controlan las innovaciones más avanzadas. Poner recursos avanzados en manos de una comunidad diversa y global no solo acelera la innovación, sino que también fomenta un análisis crítico de las implicaciones éticas y sociales de la tecnología de IA.

Tanto los modelos comerciales como los de código abierto desempeñan roles fundamentales y complementarios en el progreso del NLP. Las implementaciones comerciales están definiendo el estado del arte, mientras que el ecosistema de código abierto asegura la distribución equitativa de conocimientos y capacidades, creando sinergias que impulsan la innovación y ofrecen flexibilidad para enfrentar los retos emergentes en NLP de manera más democrática y ética.

3.4.3. Inferencia Eficiente

La eficiencia en la Inferencia es un aspecto crucial en el despliegue de Modelos de Lenguaje de Gran Escala (LLMs) y otras tecnologías de inteligencia artificial en entornos de producción. Implica optimizar modelos para que consuman menos recursos computacionales, como

memoria y energía, al tiempo que mantengan un rendimiento adecuado. Esta problemática es especialmente relevante en aplicaciones que requieren respuestas en tiempo real o que se ejecutan en dispositivos con capacidades computacionales limitadas.

En el campo del reconocimiento del habla, el trabajo realizado por Radford et al. (2022) representa un significativo avance en el entrenamiento de modelos robustos de reconocimiento de voz mediante supervisión débil de gran escala. Pero más allá del entrenamiento, para hacer accesible este tipo de tecnología avanzada en aplicaciones cotidianas, es imperativo contar con métodos de inferencia eficientes.

En la búsqueda de una inferencia más eficiente, se ha explorado la posibilidad de reducir la precisión de los cálculos realizados por los modelos. Dettmers & Zettlemoyer (2023) aboga por la adopción de una precisión de 4 bits en la inferencia, argumentando las leyes de escalamiento que demuestran cómo esto puede influir en un equilibro entre el rendimiento del modelo y la eficiencia computacional. El proyecto `llama.cpp` (Gerganov, 2023a), liderado por Georgi Gerganov, ejemplifica esta idea al ejecutar el modelo LLaMA con cuantificación de enteros de 4 bits en un computador portatil, utilizando implementaciones optimizadas para diferentes arquitecturas y ofreciendo soporte para diversas formas de cuantificación de enteros.

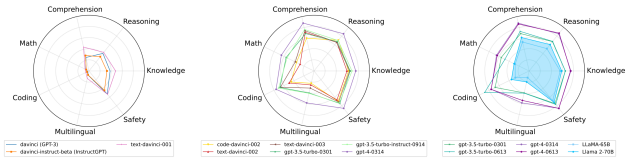
La iniciativa `llama.cpp` no solo sirve como caso de estudio para modelos de inferencia eficiente en diferentes plataformas hardware, sino también como un campo de pruebas para la biblioteca `ggml` (Gerganov, 2024), orientada al desarrollo de características para la inferencia de LLMs de manera eficiente. Se destaca el apoyo a la precisión mixta y la cuantización que abarca desde 2 bits hasta 8 bits, facilitando que incluso sistemas heredados o con recursos limitados puedan beneficiarse del poder de la inteligencia artificial moderna.

La inferencia de alto rendimiento también es un tema clave en modelos como Whisper de OpenAI para el reconocimiento automático del habla (ASR). La implementación de alta eficiencia de Whisper, mencionada por Gerganov (Gerganov, 2023b), demuestra cómo la optimización de la inferencia puede desempeñar un papel vital en hacer que los modelos de reconocimiento de voz sean más accesibles y utilizables en una gama más amplia de dispositivos.

Por otro lado, en lo que respecta a la afinación eficiente de LLMs cuantizados, QLoRA, introducido por Dettmers et al. (2023), presenta una metodología para el ajuste fino de LLMs cuantizados que conserva la eficiencia en el uso de recursos sin sacrificar el rendimiento del modelo. Este enfoque de ajuste fino representa un salto adelante en nuestra capacidad de desplegar modelos potentes en entornos donde los recursos computacionales son un bien escaso.

Finalmente, el progreso en dispositivos de inferencia no se limita solo a la voz y el texto. Proyectos como Nougat de Blecher et al. (2023), que busca comprender y procesar documentos académicos llenos de gráficos y fórmulas complejas, también dependen de la inferencia eficiente para llevar las capacidades de NLP a nuevos dominios de aplicación.

El desarrollo continuo en técnicas de inferencia eficiente garantiza que los avances en reconocimiento del habla, comprensión del lenguaje y otras tareas de procesamiento de datos no solo sean teóricamente posibles sino prácticamente aplicables, ampliando el alcance de lo que los dispositivos inteligentes pueden lograr y permitiendo que la inteligencia artificial tenga un impacto positivo y profundo en la sociedad.



(a) GPT-3 Series (b) GPT-3.5 Series and GPT-4 (c) Llama 2-70B

Figura 3.5: Se muestran las habilidades de los modelos agrupadas en 8 categorias, en las cuales se incluye el razonamiento y matemáticas, mostrando especial debilidad incluso en modelos representativos del estado del arte (S. Zheng et al., 2023)

En base a lo anterior, el análisis bibliográfico refuerza la relevancia de este proyecto y aporta la confianza de que la optimización del LLM para la enseñanza de las matemáticas de la PAES puede tener resultados efectivos y rentables.

3.4.4. Limitaciones de los Modelos

Los Modelos de Lenguaje de Gran Escala (LLMs) han alcanzado logros impresionantes en diversas áreas del procesamiento del lenguaje natural (NLP). Sin embargo, aún enfrentan obstáculos considerables que dificultan su implementación efectiva en aplicaciones del día a día. Entre las limitaciones más destacadas se encuentran las cuestiones de comprensión, razonamiento y explicabilidad.

En primer lugar, una de las principales limitaciones de los LLMs es su falta de comprensión profunda del mundo real. Aunque son capaces de generar texto coherente y a menudo convincente, lo hacen basándose en patrones de lenguaje aprendidos, en lugar de una verdadera comprensión semántica de los conceptos subyacentes. Esto se manifesta en errores y "halucinaciones" —la generación de información falsa o sin sentido— especialmente cuando se les pide que proporcionen información factual o detallada sobre tópicos especializados.

Además, los LLMs suelen tener dificultades para realizar razonamientos complejos de varios pasos que requieren una lógica precisa y la manipulación de conocimientos especializados. A pesar de que pueden resolver problemas simples y directos, cuando se enfrentan a cuestiones que necesitan de iteraciones o razonamientos anidados, su rendimiento desciende notablemente. Esto limita su utilidad en campos que dependen de un análisis riguroso, como la matemática avanzada, la programación o la toma de decisiones estratégicas.

Otro desafío significativo es la falta de explicabilidad de los LLMs. Aunque se entiende bien cómo se construyen estos modelos —a través de algoritmos de aprendizaje automático que ajustan millones de parámetros—, es mucho menos claro cómo estos modelos llegan a sus conclusiones específicas. No existe una comprensión clara del proceso de *"pensamiento"* interno que lleva a un LLM a generar una salida particular. Esta opacidad plantea dudas no solo sobre cómo verificar sus outputs y corregir errores, sino también sobre cómo confiar en sus decisiones en aplicaciones críticas.

El reconocimiento de estas limitaciones es un paso crucial para orientar futuras investigaciones y desarrollos. La comunidad científica está trabajando activamente para superar estos obstáculos, explorando desde la incorporación de conocimiento externo y el uso de modelos híbridos que combinan LLMs con bases de datos estructuradas, hasta el desarrollo de técnicas de cuantificación de la incertidumbre y algoritmos que permitan una mayor transparencia y

trazabilidad en el proceso de inferencia de los modelos.

Así, mientras los LLMs continúan abriendo nuevas fronteras en el campo de la inteligencia artificial, es imprescindible abordar estas limitaciones para maximizar su potencial y garantizar su aplicabilidad en el mundo real, asegurando que estos sistemas actúen de manera confiable, transparente y beneficiosa para la sociedad.

3.5. Técnicas para Maximizar el Rendimiento

La implementación y refinamiento de Modelos de Lenguaje de Gran Escala (LLMs) abren una amplia gama de posibilidades en la búsqueda por maximizar su rendimiento. Las siguientes secciones exploran detenidamente las técnicas actuales y emergentes que permiten a los investigadores y profesionales adaptar y optimizar LLMs para una amplia variedad de aplicaciones, cada una con sus respectivas fortalezas y desafíos.

Equipados con la habilidad para procesar y comprender enormes cantidades de texto, los LLMs han revolucionado el campo del procesamiento del lenguaje natural (NLP), ofreciendo perspectivas sin precedentes en tareas de conocimiento intensivo. Sin embargo, la efectividad de estos modelos es tan buena como las estrategias implementadas para maximizar su rendimiento, especialmente cuando se les enfrenta al desafío de la asimilación de nueva información y la adecuación a dominios específicos del conocimiento.

Los LLMs concentran en sus pesos pre-entrenados una vasta cantidad de información factual, lo cual les permite responder a una diversidad de preguntas en múltiples dominios. No obstante, la información que estos modelos pueden proporcionar se encuentra limitada a los datos con los que fueron entrenados, surgiendo desafíos cuando se intenta incorporar nueva información o refinar las habilidades del modelo en información previamente vista. Ante esta situación, se torna crucial la selección de la técnica adecuada para la actualización y el perfeccionamiento de estas capacidades.

En este estudio, presentado por Ovadia et al. (2023), se evalúan dos enfoques predilectos: el fine-tuning y la generación aumentada por recuperación (*Retrieval-Augmented Generation*, ó RAG). Ambas técnicas se someten a un escrutinio bajo una serie de tareas intensivas en conocimiento, extendiéndose por distintos temas y áreas. Las conclusiones de este análisis revelan que, si bien el fine-tuning ofrece mejoras menores, RAG exhibe un desempeño superior de manera consistente, tanto para conocimientos previamente encontrados durante el entrenamiento como para aquellos totalmente nuevos.

El fine-tuning, aunque ampliamente adoptado, parece encontrar limitaciones al momento de enseñar nueva información factual a los LLMs. Esta investigación arroja luz sobre la posibilidad de que la exposición del modelo a variaciones múltiples del mismo hecho pueda aliviar dicha problemática, una estrategia que podría ser clave para capacitar a los LLMs en el aprendizaje de nueva información de manera más efectiva.

La sección “Técnicas para Maximizar el Rendimiento” propone desglosar estas técnicas, su aplicabilidad y eficiencia para enriquecer a los LLMs en contextos donde la actualización del conocimiento y la especialización temática desempeñan roles cruciales. Al adentrarnos en esta exploración, trazamos un camino que no solo define las mejores prácticas actuales sino que también establece una dirección para futuras innovaciones en el entrenamiento y aplicación

de LLMs, marcando así un siguiente paso en la evolución de la inteligencia artificial.

3.5.1. *Prompt-Engineering*

En el dominio de la inteligencia artificial, la ingeniería de *prompts* ha surgido como una técnica refinada y crítica para la interacción con Modelos de Lenguaje de gran escala (LLMs). Concebida como una práctica híbrida que fusiona arte y metódica, la ingeniería de *prompts* requiere de una concepción estratégica de los estímulos de entrada que guíen eficazmente a los LLMs en la generación de respuestas precisas y coherentes con la tarea solicitada.

La investigación de Lewkowycz et al. (2022) revela cómo los LLMs pueden abordar problemas de razonamiento cuantitativo al ser provocados con *prompts* cuidadosamente diseñados, mientras que Wei et al. (2023) demuestran que la técnica de *Chain-of-Thought* ("cadena de pensamiento") promueve un razonamiento más profundo en los LLMs. Paralelamente, Wang et al. (2023b) solidifican la importancia de la autoconsistencia para reforzar el razonamiento generado en la *Chain-of-Thought*, Y. Zhou et al. (2023) y Weng et al. (2023) identifican que los LLMs pueden funcionar como ingenieros de *prompts* a nivel humano y son capaces de verificar sus propias respuestas al aplicar métodos de autoevaluación.

Explorando otras estrategias innovadoras, Madaan et al. (2023) presentan un mecanismo de autorrefinamiento iterativo con retroalimentación auto-generada, y Yao et al. (2023) proponen un enfoque deliberado de solución de problemas utilizando estructuras lógicas similares a árboles de decisiones. C. Li et al. (2023) añaden una dimensión emocional, al mostrar que estímulos emocionales pueden mejorar el desempeño de los LLMs, y Casper et al. (2023) exponen desafíos y limitaciones fundamentales en el aprendizaje reforzado a partir de retroalimentación humana.

Más aún, A. Zhou et al. (2023) exhiben cómo aplicar métodos de autoverificación basados en código para resolver problemas matemáticos desafiantes, y Besta et al. (2023) introducen un diseño de *prompts* basado en grafos matemáticos, abriendo nuevas posibilidades para resolver problemas complejos. Desde una perspectiva de decodificación, O'Brien & Lewis (2023) mejoran el razonamiento en LLMs con técnicas de descodificación contrastiva, mientras que Yu et al. (2023) exploran la propagación del pensamiento utilizando métodos analógicos para el razonamiento complejo. Por su parte, Zheng et al. (2023) sugieren tomar un paso atrás (y permitir al modelo reflexionar) para evocar el razonamiento mediante la abstracción, y Xu et al. (2023) abordan el desafío de la creación manual de *prompts* complejos e instrucciones de formación, proponiendo métodos para que los propios LLMs generen datos de instrucción de dominio abierto con complejidad variable.

El enfoque holístico que implica la ingeniería de *prompts*, abarcando técnicas de *Chain-of-Thought*, autoconsistencia, autorrefinamiento, y la integración de emociones y grafos, se posiciona como un componente clave en el avance de los LLMs hacia la resolución efectiva y autónoma de problemas cuantitativos. Al emplear un conjunción de enfoques que abordan tanto la formulación intuitiva de preguntas como la verificación y refinamiento de respuestas, este campo de estudio promete revolucionar la interacción entre los seres humanos y los sistemas guiados por inteligencia artificial, llevando la precisión y utilidad de los LLMs a nuevos horizontes.

3.5.2. *Fine-tuning*

El procedimiento de ajuste fino (*fine-tuning*) se consolida como pilar fundamental en la disciplina del Procesamiento del Lenguaje Natural (PLN), consistiendo en el perfeccionamiento de Modelos de Lenguaje de Gran Escala (LLMs) preentrenados para su adaptación a tareas concretas y demarcadas. Esta sección del marco teórico ofrece una profundización en los principios subyacentes al fine-tuning, complementada con una selección curada de ejemplos paradigmáticos y estudios de referencia.

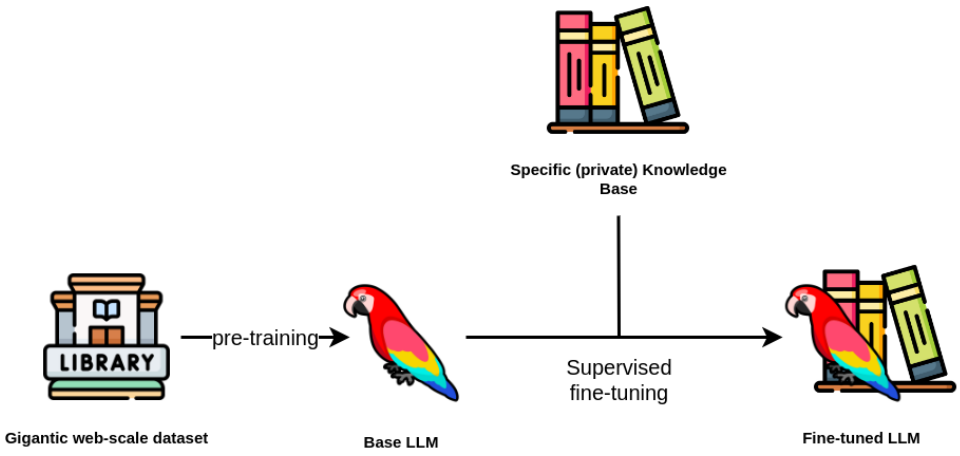


Figura 3.6: *Diagrama de flujo del proceso de fine-tuning supervisado, por Tomaz Bratanic*

Howard & Ruder (2018) revolucionaron el campo con el *Universal Language Model Fine-tuning* (ULMFiT), una técnica de aprendizaje por transferencia orientada principalmente a la clasificación de textos. El ULMFiT destila el proceso de *fine-tuning* en una secuencia de fases definidas: el ajuste discriminativo de la tasa de aprendizaje, la descongelación progresiva de capas de modelo y la aplicación de estrategias de regularización para prevenir el sobreajuste. Tales innovaciones han permitido obtener modelos especializados con un desempeño superlativo en tareas específicas, aun con volúmenes de datos restringidos.

El proceso de fine-tuning implica varios pasos que se llevan a cabo secuencialmente con el objetivo de adaptar un modelo de lenguaje preentrenado (LLM) a una tarea específica. A continuación, se detalla cada uno de estos pasos:

1. **Selección del Modelo Preentrenado**: El punto de partida es un LLM que ya ha sido entrenado previamente con un conjunto de datos extenso y general, lo que le permite tener conocimientos base sobre el lenguaje natural. La selección del modelo dependerá del tamaño deseado (cantidad de parámetros) y su compatibilidad con la tarea específica.
2. **Compilación del Dataset Específico**: Se recopila un conjunto de datos relevante para la tarea específica que se desea mejorar con el fine-tuning. Este conjunto de datos debe estar limpio, ser representativo de la tarea y estar organizado de acuerdo con las necesidades de entrada/salida del modelo.
3. **Preprocesamiento de Datos de Entrenamiento**: Los datos se preparan para ser utilizados por el modelo. Esto generalmente incluye tokenización, vectorización y, a

menudo, la aplicación de técnicas para manejar la longitud variable de las entradas como 'padding' o 'truncation'.

4. **Inicialización y Ajustes de Configuración**: Antes de iniciar el fine-tuning, se definen y ajustan varios hiperparámetros como la tasa de aprendizaje, el tamaño del lote (batch size), y el número de épocas de entrenamiento. También se suele implementar una tasa de aprendizaje discriminativa donde diferentes capas del modelo tienen distintas tasas de aprendizaje, usualmente asignando tasas menores a las capas más bajas para preservar el conocimiento adquirido durante el preentrenamiento y ajustando las tasas más altas para aprender nuevas características específicas de la tarea.
5. **Descongelación controlada de las Capas:** Algunos enfoques sugieren congelar inicialmente las capas inferiores o intermedias del modelo y entrenar solo las últimas capas. Posteriormente, se descongelan más capas de manera gradual para su afinamiento, permitiendo ajustes finos en la representación de las características.
6. **Entrenamiento Supervisado**: Durante el fine-tuning, el modelo se entrena con el conjunto de datos específico, usando un algoritmo de optimización (como Adam o SGD), para minimizar la función de pérdida definida acorde a la tarea.
7. **Regularización y Evitación de Sobreajuste**: Se aplican técnicas como dropout, early stopping, o aumentos de datos (data augmentation) para evitar que el modelo se especialice demasiado en el conjunto de entrenamiento y no generalice a datos nuevos.
8. **Evaluación y Ajuste Iterativo**: Tras el entrenamiento, el modelo se evalúa en un conjunto de validación o test para asegurar que su rendimiento ha mejorado en la tarea de interés. Se pueden hacer ajustes adicionales en los hiperparámetros y el proceso de fine-tuning podría repetirse para obtener mejoras incrementales.
9. **Implementación**: Una vez que el modelo alcanza un rendimiento satisfactorio, se despliega para su uso en producción o en un entorno en vivo, donde se seguirá monitorizando y, si es necesario, ajustando con más datos y fine-tuning iterativo.

Es fundamental recalcar que el fine-tuning es un proceso experimental que requiere atención a los detalles y ajustes basados en el tipo de datos y la tarea. Además, el tamaño del modelo preentrenado y la abundancia (o ausencia) de datos para la tarea específica influirán significativamente en la estrategia de fine-tuning a seguir.

Por otro lado y en un ámbito enfocado en la eficiencia, Hu et al. (2021) introducen LoRA (*Low-Rank Adaptation*), una modalidad que prioriza la economía de recursos computacionales durante el *fine-tuning*. LoRA propone la actualización selectiva de un subconjunto de parámetros matriciales de rango reducido, minimizando así la carga computacional sin sacrificar, y en ocasiones incrementando, la capacidad del LLM en distintas aplicaciones.

La investigación liderada por Wei et al. (2022) realza la perspectiva del fine-tuning instruccional como una herramienta estratégica que potencia la habilidad de generalización zero-shot de los LLMs hacia tareas no-vistas previamente. En este enfoque, la instrucción-tuning permite que el modelo FLAN, previamente entrenado con 137 mil millones de parámetros, incorpore conocimiento específico de una amplia gama de tareas verbales del NLP. La evaluación de FLAN en tipos de tareas no vistas anteriormente manifiesta un incremento substancial en el rendimiento con respecto al modelo sin modificar, sobrepasando incluso al GPT-3 de 175 mil millones de parámetros en una metodología zero-shot y superando a GPT-3 en configuraciones de pocos ejemplos (few-shot).

El *fine-tuning* se revela como una estrategia central para dotar a los LLMs de una capacidad de adaptación y especialización elevadas sin la necesidad de remodelar completamente sus estructuras preentrenadas. Esto conlleva beneficios intrínsecos como la reducción en el tiempo de entrenamiento, la optimización del rendimiento con conjuntos de datos específicos y la mejora en la precisión de las aplicaciones finales. Es dentro de este contexto técnico que enmarcamos nuestra investigación, destinada a emplear estas técnicas de *fine-tuning* para ajustar y optimizar LLMs en la resolución de problemas matemáticos específicos presentados en la Prueba de Acceso a la Educación Superior (PAES) en Chile, aumentando así su eficacia didáctica y su precisión conceptual.

3.5.3. *Retrieval-Augmented Generation* (RAG)

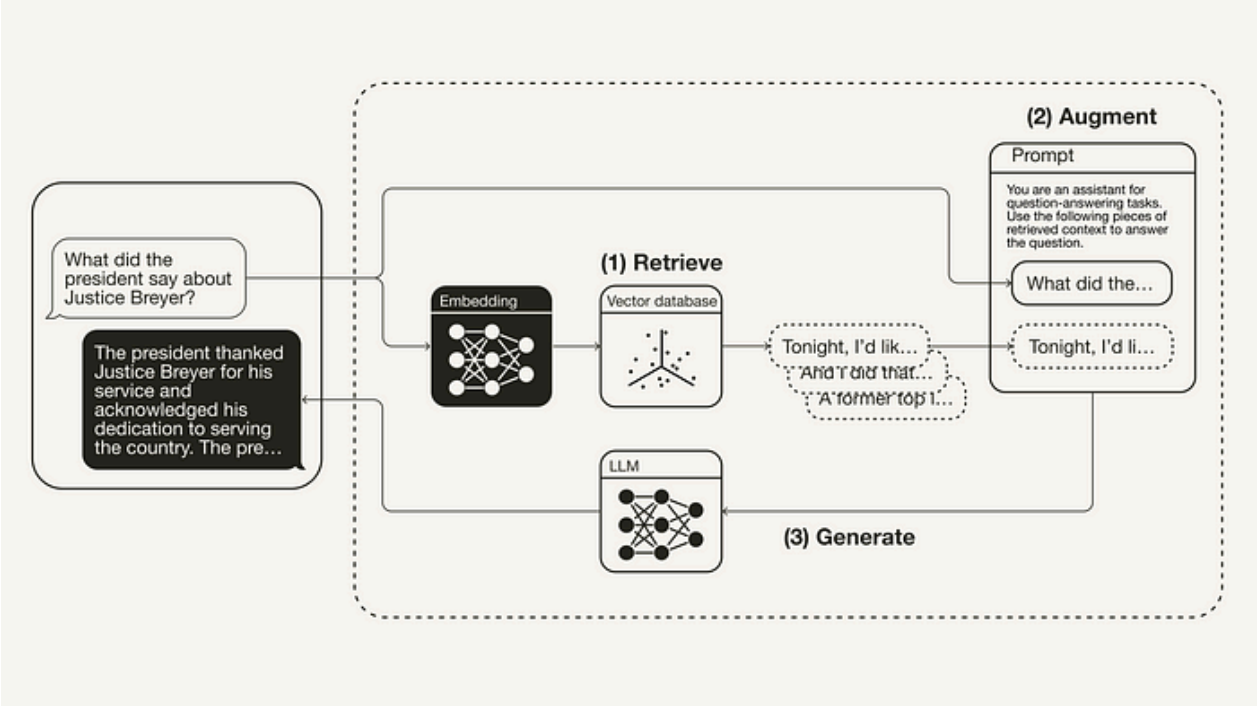


Figura 3.7: *Típico flujo de trabajo para un sistema RAG* (Monigatti, 2023)

Dentro del entorno del procesamiento del lenguaje natural (PLN), los Modelos de Lenguaje de Gran Escala (LLMs) almacenan un conocimiento enciclopédico, evidenciado por su habilidad de responder a preguntas en múltiples dominios. No obstante, la profundidad de este conocimiento es intrínsecamente finita y está sujeta a las características de los datos sobre los que fueron entrenados. En el contexto actual, en el que se busca incorporar informaciones frescas o afinar capacidades en información ya conocida, surgen técnicas como el ajuste fino (fine-tuning) y la Generación Asistida por Recuperación de Información (Retrieval-Augmented Generation - RAG), siendo objeto de evaluación en nuestra investigación para tareas intensivas de conocimiento en tópicos variados.

La Generación Aumentada por Recuperación (Retrieval-Augmented Generation - RAG) representa un enfoque híbrido que combina la capacidad generativa de los LLMs con la recuperación de información relevante desde bases de datos externas. La figura expuesta por Monigatti (2023) ilustra un flujo de trabajo típico de un sistema RAG, donde el modelo de

lenguaje no solo genera contenido, sino que también busca y referencia información específica durante el proceso (Lewis et al., 2021). Esta técnica permite a los LLMs acceder a información actualizada, ampliando su capacidad para integrar y referencia conocimientos actualizados más allá de lo contenido en su entrenamiento inicial.

Ovadia et al. (2023) resalta que el *fine-tuning*, pese a brindar ciertas mejoras, se ve superado de manera consistente por el enfoque RAG, tanto en el manejo de conocimiento previamente aprendido como en la inclusión de información totalmente nueva. Además, se detecta que los LLMs presentan dificultades al tratar de asimilar nuevos datos factuales mediante el *fine-tuning* y que la exposición a múltiples variaciones del mismo hecho en la etapa de entrenamiento podría mitigar este obstáculo.

Pese a los avances en el almacenamiento de conocimientos fácticos en los parámetros de LLMs preentrenados y a sus impresionantes resultados al ser afinados para tareas subsecuentes de NLP, su capacidad para acceder y manejar precisamente dicho conocimiento continúa siendo limitada. En tareas que demandan un manejo intensivo del conocimiento, su rendimiento aún se encuentra detrás de arquitecturas específicas para dichas tareas. Adicionalmente, proporcionar procedencia para sus decisiones y actualizar su conocimiento del mundo representan problemas de investigación abiertos.

Los modelos preentrenados con mecanismos de acceso diferenciable a una memoria explícita no paramétrica, como los implementados en sistemas RAG, pueden superar estas limitaciones. Estos modelos permiten una combinación novedosa de memorias paramétricas, como un modelo preentrenado de secuencia a secuencia, con memorias no paramétricas, tal como un índice vectorial denso de Wikipedia, accesible mediante un buscador neuronal preentrenado. Emergen entonces dos formulaciones de RAG: una que condiciona la generación en los mismos pasajes recuperados a lo largo de toda la secuencia, y otra que permite la utilización de diferentes pasajes por token.

Al establecer resultados sin precedentes en tareas de preguntas y respuestas en dominio abierto, los modelos RAG superan tanto a los modelos paramétricos de secuencia a secuencia como a las arquitecturas específicas de recuperación y extracción de tareas. Para las tareas de generación de lenguaje, descubrimos que los modelos RAG producen lenguaje más específico, diverso y factual en comparación con una línea base paramétrica de último modelo secuencia a secuencia.

Concluimos que la integración de técnicas como RAG en LLMs amplía de manera sustancial la funcionalidad y aplicabilidad de tales sistemas en el ámbito educativo, con implicaciones directas en la mejora de la calidad educativa y la precisión de la tutoría automática para preparación hacia la Prueba de Acceso a la Educación Superior (PAES) de Matemáticas en Chile. Las ventajas de la actualización continua de conocimiento y del manejo refinado de información especializada posicionan a RAG como un instrumento preferencial para la adaptación de LLMs en contextos de aprendizaje dinámico y personalizado. Logrando así sinergias con otros métodos de optimización de los modelos actuales.

3.5.4. *Human-in-the-loop*

El concepto de *Human-in-the-loop* refiere a la integración de intervenciones humanas en diversas etapas del ciclo de vida de un modelo de inteligencia artificial, desde su concepción

hasta la fase de inferencia, con el fin de enriquecer el proceso de aprendizaje y mejora en la toma de decisiones del modelo (Wu et al., 2022). Dicha metodología permite realizar ajustes precisos y una validación continuada que resulta sumamente beneficiosa en tareas de alto grado de complejidad cognitiva o aquellas en las que se requiere compresión contextual avanzada. La implicación directa de expertos humanos contribuye a refinar los outputs del modelo y a tener una comprensión cabal de sus limitaciones y capacidades (Mosqueira-Rey et al., 2023), redundando en la mejora de la precisión y la utilidad de los LLMs. Incluso con el advenimiento de contratar profesores para modelos de lenguaje, ya que la retroalimentación humana sigue siendo un componente crítico para el aprendizaje de los modelos.

El trabajo colaborativo entre seres humanos y modelos de lenguaje natural ha ocupado un papel principal en investigaciones como las de Ouyang et al. (2022) y Stiennon et al. (2022), que examinan la factibilidad de mejorar la obediencia de los modelos a las instrucciones humanas gracias a la retroalimentación directa durante el entrenamiento. Tales investigaciones procuran incorporar evaluaciones cualitativas con el fin de alcanzar una congruencia más acendrada y especializada entre los modelos de lenguaje y los propósitos humanos.

La emergencia de elevados imperativos éticos y de seguridad en el despliegue de asistentes basados en IA se ve reflejada en estudios como los de Bai, Jones, et al. (2022) y Bai, Kadavath, et al. (2022). Dichos trabajos abogan por un modelo de aprendizaje fortalecido mediante feedback tanto humano como de IA, orientado hacia la tutoría basada en principios de una IA constitucional y la prevención de perjuicios.

La minimización de la intervención humana en la optimización de los LLMs es abordada por Honovich et al. (2022), quien explora el aprovechamiento de instrucciones atípicas como medio para ajustar los modelos, proponiendo técnicas que reducen la dependencia de la retroalimentación humana y que podrían derivar en un desarrollo más efectivo y escalable de modelos de lenguaje.

Por otro lado, C. Zhou et al. (2023b) promueve una estrategia denominada LIMA (*Less Is More for Alignment*), haciendo hincapié en la importancia de la retroalimentación concentrada para lograr una alineación más eficaz y simplificada con los objetivos propuestos, especialmente en contextos de mitigación de sesgos y áreas críticas de los modelos.

La amplitud del aprendizaje basado en feedback humano es puesta a prueba por Lee et al. (2023) con el enfoque RLAIF (*Reinforcement Learning from AI Feedback*), que amalgama la retroalimentación producida por sistemas de IA con la humana, posibilitando un esquema de entrenamiento más eficiente y sólido para los modelos.

Para finalizar, la integración de las preferencias humanas en el aprendizaje reforzado es abordada en los trabajos pioneros de Christiano et al. (2017), explorando metodologías encaminadas a dirigir los comportamientos modelados en consonancia con los valores y metas humanas; un aspecto crítico para el desarrollo de sistemas de IA alineados con las prioridades humanas.

Cada uno de estos métodos juega un papel complementario en la maximización del rendimiento de los LLMs. La adopción apropiada de estas técnicas puede llevar a los modelos de procesamiento de lenguaje a nuevos horizontes de efectividad y precisión, permitiendo así que proyectos como la optimización de la PAES reflejen no solo lo mejor de la innovación

tecnológica actual sino que también se mantengan alineados con las necesidades emergentes de la educación moderna.

Capítulo 4

Estado del arte

4.1. Matemáticas y Razonamiento

Este capítulo explora la intersección de las matemáticas y *Machine Learning*, destacando cómo las técnicas de Inteligencia Artificial pueden resolver problemas matemáticos complejos y mejorar el razonamiento lógico y abstracto.

4.1.1. Métodos de entrenamiento y estrategias

Para alcanzar un desempeño óptimo en tareas de razonamiento matemático, resulta esencial la optimización de los métodos de entrenamiento y las estrategias de los modelos de inteligencia artificial, lo cual potencia sus facultades cognitivas. Los estudios recientes en esta área se orientan hacia diferentes metodologías para lograr dicho objetivo.

Cobbe et al. (2021) constata que, aunque los modelos de lenguaje de última generación pueden rivalizar con el rendimiento humano en una variedad de tareas, todavía afrontan retos al ejecutar razonamiento matemático de múltiples pasos de manera robusta. Mediante la introducción de GSM8K, una colección de 8.5K problemas matemáticos verbales de nivel escolar con alta calidad y diversidad lingüística, los investigadores evidencian que ni siquiera los grandes modelos transformer logran un rendimiento destacado, a pesar de la simplicidad conceptual de dichos problemas. Sugieren entrenar verificadores que evalúen la precisión de las soluciones generadas por el modelo, y durante la etapa de prueba, elegir la solución más verosímil de varias opciones basándose en el verificador. Este estudio demuestra que la verificación mejora de manera significativa el desempeño en GSM8K, evidenciando empíricamente que la verificación escala más efectivamente con un aumento en el volumen de datos que una línea base de fine-tuning.

Srivastava et al. (2023) examina las capacidades presentes y futuras de los modelos de lenguaje y su eficacia al evaluar diferentes tareas que se consideran fuera de su alcance actual. Revela que al incrementar el tamaño de los modelos se mejora el desempeño y la calibración, aunque el rendimiento absoluto es insuficiente (especialmente en comparación con expertos humanos). Aquellas tareas que muestran una mejoría gradual y predecible incluyen componentes de conocimiento o memorización, mientras que las tareas que reflejan innovaciones disruptivas a una escala crítica suelen requerir múltiples fases o elementos, o se basan en métricas frágiles;

el sesgo social tiende a intensificarse con la escalabilidad del modelo.

J. Huang et al. (2023) destaca la autocorrección como un mecanismo fundamental para superar problemas de precisión y relevancia en las respuestas proporcionadas por los LLMs. Esta investigación indica que los LLMs enfrentan retos para autocorregirse sin retroalimentación externa y que, en ocasiones, su rendimiento puede disminuir tras la autocorrección. Se presentan propuestas para futuras investigaciones y aplicaciones prácticas que toman en cuenta estas limitaciones.

Por su parte, Yuan et al. (2023) investiga la relación entre el escalado de los LLMs y su habilidad en el razonamiento matemático, un área que se ha explorado en menor medida. Se descubre que la tasa de error durante el entrenamiento previo es un indicador más fiable del rendimiento del modelo que la cantidad de parámetros. Mediante la aplicación de fine-tuning supervisado con diferentes volúmenes de datos, los autores identifican una relación logarítmico-lineal entre la cantidad de datos y el rendimiento del modelo. Asimismo, observan que los modelos mejorados se benefician en menor medida con datos supervisados adicionalmente extensos. Proponen el Rejection Sampling Fine-Tuning (RFT) para realzar el rendimiento en tareas de razonamiento matemático, aislando rutas de razonamiento correctas generadas por modelos supervisados. Con el rechazo de muestras que contienen trayectorias de razonamiento más distintivas, el RFT optimiza dicha capacidad en los LLMs y resulta en una mejora más notoria en modelos de menor rendimiento.

Liu et al. (2023) detecta una brecha significativa entre el rendimiento de los LLMs en la resolución de problemas matemáticos dependiendo de si aciertan en el primer intento o después de múltiples intentos, incentivando la exploración de métodos de fine-tuning que potencien su desempeño. El estudio aborda tres estrategias de fine-tuning:

1. **afinación de soluciones detalladas**,
2. **re-clasificación de clusters de soluciones**, y
3. **fine-tuning secuencial multitarea** que integra tanto la generación como la evaluación de soluciones, ofreciendo mejoras de rendimiento al modelar juntas ambas tareas.

Finalmente, Ma et al. (2023) subraya avances significativos en el razonamiento de múltiples etapas utilizando LLMs. La investigación se enfoca en el Modelo Procesal Supervisado de Recompensa (PRM), que aporta retroalimentación paso a paso durante el entrenamiento, similar al Proximal Policy Optimization (PPO) o el muestreo por rechazo. Presentan un algoritmo de búsqueda voraz, o greedy, para la etapa de inferencia que se sirve de la retroalimentación a nivel de pasos proporcionada por el PRM para optimizar las rutas de razonamiento en los LLMs, demostrando resultados superiores en comparación con la técnica de Chain-of-Thought (CoT) y confirmando la eficacia de dicho enfoque en las tareas de razonamiento.

En su conjunto, estas investigaciones apuntan hacia múltiples posibilidades para la mejora de las habilidades de los LLMs en tareas complejas de razonamiento matemático, estableciendo fundamentos robustos para progresar en este aspecto crucial del aprendizaje automático.

4.1.2. Mediciones y Comparaciones

La evaluación precisa y comparativa del rendimiento de los modelos de lenguaje en tareas de razonamiento lógico-matemático es una piedra angular para el avance en el desarrollo de la inteligencia artificial. Mediante la emisión de juicios basados en mediciones meticulosas y análisis comparativos, los investigadores pueden discernir no sólo la competencia presente de los sistemas de IA, sino también trazar un camino hacia mejoras significativas en sus habilidades.

Hendrycks et al. (2021) afronta esta temática de manera directa mediante la introducción del conjunto de datos MATH, albergando una amplia gama de problemas matemáticos que exigen un enfoque de solución estructurada y paso a paso. Pese a los esfuerzos por incrementar la precisión de los modelos de lenguaje –particularmente los modelos Transformer de gran escala– en resolver estos problemas, los resultados exhiben un nivel de exactitud que aún no alcanza un estándar ideal. Estos hallazgos subrayan una verdad incipiente: la aplicación de técnicas de escalado en parámetros de modelo y presupuestos no es una solución holística para materializar habilidades avanzadas en razonamiento matemático en IA. Se desprende de ello una llamada a la acción para la comunidad científica, con el fin de aventurarse en el campo de nuevos avances algorítmicos que puedan catalizar la necesaria evolución.

Del mismo modo, Mitchell et al. (2023) emprende un análisis del razonamiento abstracto en versiones tanto textuales como multimodales de GPT-4, evidenciando que, a pesar de la magnitud de este modelo y sus variaciones, no se ha logrado aún un nivel de abstracción que se asemeje al discernimiento humano. Este estudio proporciona una valiosa perspectiva acerca de las limitaciones actuales de los modelos de lenguaje y ofrece un vistazo a las complejidades inherentes al razonamiento abstracto, un aspecto central del pensamiento avanzado.

Reflexionando sobre estas observaciones, es claro que la evaluación de las capacidades lógico-matemáticas y de razonamiento en modelos de lenguaje, aunque fructífera en ciertos aspectos, todavía se halla en las etapas incipientes. Los estudios citados no sólo ilustran las capacidades actuales y las deficiencias de dichos modelos, sino que también destacan la imperativa necesidad de redireccionar nuestros enfoques y esforzarnos por comprender más profundamente los fundamentos de la cognición matemática y abstracta. Es probable que el vértice futuro de este progreso sea testigo de un cruzamiento sin precedentes de disciplinas, desde la psicología cognitiva hasta la teoría matemática, cada una aportando un entendimiento esencial para el ensamblaje de sistemas de IA verdaderamente avanzados.

4.1.3. Resolución de Problemas y Generación de Teoremas

La capacidad de los sistemas de inteligencia artificial para enfrentar problemas matemáticos y forjar nuevos teoremas constituye un dominio de investigación en auge. Este cuerpo de trabajo explora las metodologías que dichos sistemas utilizan para afrontar estos desafíos y las posibles vías de mejora para profundizar sus competencias.

Lewkowycz et al. (2022) aborda esta complejidad introduciendo Minerva, un modelo de lenguaje de gran escala preentrenado tanto en data general de lenguaje natural como en contenido técnico. A pesar de constituir un avance significativo en el desempeño sobre benchmarks técnicos, sin apoyos de herramientas externas, los resultados presentan una velocidad moderada en progresos cuantitativos en problemas subgraduados de ciencias que

requieren razonamiento numérico, luciendo un umbral donde solo se responde correctamente a un tercio de ellos. El estudio recalca la necesidad de trascender la simple escalabilidad de modelos para incursionar en la resolución de ecuaciones y problemas a este nivel.

El trabajo de Uesato et al. (2022) lleva esta reflexión un paso adelante, contrastando métodos de supervisión basados en resultados frente a aquellos centrados en el propio proceso de razonamiento. Al adentrarse en GSM8K, una batería de problemas matemáticos presentados en lenguaje natural, se descubre que mientras la supervisión basada en resultados posee tasas de error en respuestas finales similares con menor supervisión de etiquetas, para la correcta derivación de pasos racionales se requiere supervisión basada en procesos o de modelos de recompensa aprendidos que emulen el feedback basado en proceso. Tal enfoque reduce notoriamente los errores tanto en respuestas finales como en razonamiento, potenciando la precisión de las soluciones finales correctas.

Por su parte, Zelikman et al. (2022) presenta el *"Self-Taught Reasoner"* (STaR), una metodología autoinstruible que utiliza un ciclo simple para generar razonamientos paso a paso: si una respuesta generada es incorrecta, el modelo lo intenta de nuevo contando con la respuesta correcta, y después se ajusta en función de los razonamientos que condujeron a una respuesta correcta. Este método resulta en una mejora sustancial en el rendimiento en distintos conjuntos de datos y se alinea con el rendimiento de modelos de lenguaje significativamente más grandes en tareas como CommensenseQA, demostrando la viabilidad de esta técnica para elevar las habilidades de razonamiento matemático de un modelo de lenguaje.

No obstante, West et al. (2023) pone de manifiesto lo que podría considerarse un 'paradoja de IA generativa': modelos con la habilidad de generar respuestas que desafían incluso a expertos humanos, pero que simultáneamente exhiben flagrantes errores de comprensión. Este fenómeno apunta a una divergencia entre las configuraciones de inteligencia en modelos generativos y la inteligencia humana, resaltando que la capacidad de generar contenido de alta calidad no necesariamente implica un entendimiento profundo del mismo, un recordatorio para proceder con cautela al trazar paralelos entre la inteligencia artificial y humana.

Finalmente, Wang et al. (2023a) sobresale con el GPT-4 Code Interpreter y su notable habilidad para abordar problemas matemáticos complejos, atribuida a su competencia para razonar en lenguaje natural, generar y ejecutar código y además, continuar razonando basado en el *output* de la ejecución. Su investigación propone un enfoque para ajustar modelos de lenguaje de fuente abierta permitiéndoles usar código en la modelización y derivación de ecuaciones matemáticas. El innovador MathCoder emerge como un conjunto de modelos capaces de generar soluciones basadas en código, marcando puntajes récord en conjuntos de datos de referencia MATH y GSM8K y estableciéndose como precursor en esta línea de investigación.

Estos estudios ilustran un crisol de estrategias. Desde enfoques de supervisión en procesos, razonamiento autorreflexivo, hasta reconciliación de generación versus comprensión, se abren caminos para que la IA supere las barreras de la matemática avanzada e incluso desborde los confines tradicionales de la capacidad cognitiva humana. Con la mira puesta en el horizonte matemático y la facilitación del razonamiento estructurado, la IA se posiciona para evolucionar como una herramienta cada vez más integral en el campo matemático.

4.1.4. Razonamiento e Inteligencia Artificial

En el estudio de la inteligencia artificial avanzada, la capacidad de razonamiento representa un aspecto intrigante y desafiante. La investigación está enfocada en cómo la IA puede no solo imitar el razonamiento humano, sino también descubrir y establecer nuevos principios lógicos y matemáticos.

El lanzamiento de GPT-4 (Arkoudas, 2023) marcó un hito significativo en el progreso de los modelos de lenguaje, presentando mejoras notorias respecto a GPT-3.5. No obstante, el análisis crítico de su rendimiento revela una clara discrepancia entre las percepciones de su "inteligencia" y su capacidad real para razonar. Este estudio plantea dudas sobre las supuestas habilidades de razonamiento del modelo y argumenta que, lejos de un verdadero pensamiento analítico, los aciertos de GPT-4 parecen más bien ser destellos aislados y no indicativos de una capacidad sistemática para razonar.

En contraste, J. Huang & Chang (2023) ofrece un panorama más generalizado y matizado sobre el razonamiento en los modelos de lenguaje. Este trabajo proporciona una perspectiva amplia de las técnicas actuales para inducir y mejorar el razonamiento en LLMs, presentando métodos para evaluar estas capacidades. Mientras que se reconoce el progreso significativo en el ámbito del procesamiento de lenguaje natural, la cuestión subyacente es si los modelos pueden realmente alcanzar un nivel de razonamiento similar al humano.

Abordando la brecha entre el razonamiento humano y de la IA, Hao et al. (2023) señala que a los LLMs les falta un modelo de mundo interno que les permita predecir estados mundiales y simular resultados a largo plazo de acciones, una característica esencial del razonamiento humano. Presentan el marco "Reasoning via Planning" (RAP), que utiliza al LLM tanto como un modelo del mundo como un agente razonador. Los resultados empíricos muestran que RAP supera a aproximaciones anteriores en la generación de planes y en el razonamiento matemático y lógico.

Este descubrimiento es complementado por Wu et al. (2023), que investiga si las habilidades de razonamiento abstracto de los modelos de lenguaje son transferibles a tareas distintas a aquellas para las que se entrenaron. Los resultados apuntan a que las habilidades de resolución abstracta de tareas son a menudo especializadas y dependen de procedimientos no transferibles, lo que obliga a cuestionar la generalización del desempeño de la IA y a fomentar un análisis más exhaustivo del comportamiento de los modelos.

Ante la rápida evolución de los LLMs y su creciente omnipresencia, Kaddour et al. (2023) aspira a establecer un conjunto sistemático de problemas abiertos y éxitos de aplicación práctica para proporcionar a los investigadores una orientación actualizada y eficiente del campo.

El razonamiento *multi-hop* (varios pasos) plantea un reto particular para los LLMs, Sakarvadia et al. (2023) sugiere una mejora en la capacidad de razonamiento inyectando información específica en los cabezales de atención del modelo durante el proceso inferencial, demostrando con eficacia cómo la inserción de 'memorias' puede facilitar la resolución de tareas complejas *multi-hop*, reforzando la idea de que la memoria desempeña un papel crucial en el razonamiento avanzado.

Finalmente, Tyen et al. (2023) aborda el desafío de la autorrrección de errores lógicos en los LLMs, proponiendo una técnica de retroceso para mitigar el deterioro del rendimiento que

a menudo resulta de tales intentos de autorrrección. Este enfoque presenta una solución prometedora al demostrar mejoras sustanciales en la capacidad de corrección cuando se identifica la ubicación del error.

Estas investigaciones colectivas reflejan una imagen compleja y multifacética de las capacidades de razonamiento en la IA actual, reconociendo tanto sus limitaciones críticas como sus amplias posibilidades. Cabe entonces un llamado a la cautela y a continuar explorando soluciones innovadoras y metodologías refinadas en la búsqueda por extender la frontera de la inteligencia artificial en los dominios del razonamiento lógico y matemático.

4.1.5. Halucinaciones y Errores de Razonamiento

El razonamiento matemático y su simulación por parte de la inteligencia artificial (IA) son áreas de constante y rápida evolución. Al examinar los avances actuales y desafíos persistentes, resulta esencial considerar dos fenómenos críticos: las halucinaciones y los errores de razonamiento. Estas anomalías plantean interrogantes fundamentales sobre la confiabilidad y la forma de mejorar los modelos de IA actuales.

Rawte et al. (2023) aborda de manera integral el fenómeno de las halucinaciones en Modelos Fundacionales, resaltando la producción de contenido que desvía de la realidad factual o incluye información inventada. Este estudio provee una visión exhaustiva sobre los esfuerzos para identificar, comprender y abordar ese problema en los denominados *Large Foundation Models* (LLMs). La clasificación de diversas instancias de halucinaciones que son específicas a los LLMs revela criterios de evaluación para valorar su magnitud, además de examinar estrategias existentes para mitigar las halucinaciones y discutir posibles direcciones para la futura investigación.

Otro aspecto destacado es el *"Reversal Curse"*, expuesto por Berglund et al. (2023), una falla sorprendente en la generalización de modelos lingüísticos auto-regresivos de gran tamaño (LLMs). Estos modelos no generalizan automáticamente la información bi-direccional; es decir, si se entrena a un modelo con "A es B", no deduce automáticamente que "B es A". La falta de una lógica deductiva básica significa que los modelos no generalizan patrones prevalecientes en su conjunto de entrenamiento. Esta investigación proporciona evidencia del *"Reversal Curse"* al afinar GPT-3 y Llama-1 (2023) con declaraciones ficticias, demostrando que no pueden responder correctamente a preguntas inversas, y el fenómeno es robusto a través de diferentes tamaños y familias de modelos.

En la misma línea, Davidson et al. (2023) examina cómo los sistemas de IA de vanguardia pueden mejorarse significativamente mediante "mejoras post-entrenamiento" sin necesidad de un costoso reentrenamiento. Estas técnicas incluyen el uso de herramientas, métodos de instigación, andamiaje, selección de soluciones y generación de datos aplicados tras el entrenamiento inicial. Evaluando su efectividad en una moneda común, el "ganancia equivalente de cómputo", se descubre que estas mejoras son frecuentemente significativas. Este trabajo resalta que las mejoras post-entrenamiento no solo son eficaces, sino también relativamente económicas de desarrollar, con costos típicamente inferiores al 1 % del costo de entrenamiento original, aunque los desafíos de regulación asociados pueden ser considerables dado que modelos fronterizos podrían ser mejorados por una gama amplia de actores.

La intersección de estos estudios demuestra que los sistemas de IA pueden exhibir una

competencia impresionante en el razonamiento matemático, pero no están libres de falencias significativas. Las halucinaciones y los errores lógicos, como la maldición de la reversibilidad, son obstáculos que aún requieren atención detallada. Sin embargo, las mejoras post-entrenamiento se presentan como una vía prometedora para perfeccionar el rendimiento de los sistemas de IA actuales, sugiriendo que el camino hacia modelos más avanzados y fiables pasa por refinar los métodos de intervención posterior al entrenamiento inicial. Estos avances deben ser considerados con cautela, teniendo en cuenta la necesidad de comprender y abordar estas falencias para integrar efectivamente la IA en aplicaciones críticas en el mundo real.

4.1.6. Desafios y Perspectivas Futuras

La simulación y mejora del razonamiento matemático por medio de la inteligencia artificial (IA) se ubica en la frontera del conocimiento científico y técnico. A pesar de los avances significativos, el desarrollo de sistemas de IA capaces de adaptarse a nuevas situaciones, aprender de manera autónoma y razonar de forma compleja presenta desafíos notables. Esta sección se adentra en las tendencias actuales y la evolución de estrategias que apuntan a superar obstáculos y explorar capacidades inexploradas de la IA.

- **Uso de Herramientas** *(Tool use)*: Innovadores estudios están dilucidando cómo la IA puede aplicar herramientas contextualizadas para resolver problemas, extendiendo sus capacidades intrínsecas más allá de las limitaciones humanas. Usando calculadoras, entornos limitados de programación y otras herramientas, los modelos de lenguaje pueden resolver problemas matemáticos complejos, demostrando que la IA puede superar las limitaciones de la inteligencia humana.
- **Multimodalidad** *(Multimodality)*: La creciente convergencia de distintos tipos de datos en la IA multimodal está abriendo caminos para modelos capaces de procesar y sintetizar información de maneras más complejas y holísticas con audio, texto, video, etc. Esta tendencia está siendo explorada en el contexto de la resolución de problemas matemáticos, con resultados prometedores.
- **Automejoramiento** *(Self-improvement)*: El objetivo de los sistemas autónomos apunta a la IA que aprende de experiencias pasadas, perfecciona sus métodos y se actualiza iterativamente sin requerir intervención externa. Esto ha sido evidenciado en el desarrollo de modelos de lenguaje que se entrenan a sí mismos en la resolución de problemas matemáticos.
- **Sistema-2**: La replicación de procesos analíticos profundos y metódicos que caracterizan el razonamiento humano refleja el intento de la IA por manejar tareas que requieren planificación y reflexión minuciosa. Además, la IA puede aprender de los errores y mejorar su desempeño en base a la retroalimentación.
- **Generación de Datos Sintéticos** *(Synthetic Data)*: Este enfoque práctico busca resolver el dilema de la escasez de datos a través de la creación de conjuntos grandes y detallados, alimentando así los sistemas de IA con información rica y variada para su entrenamiento. Además, la generación de datos sintéticos puede ayudar a mitigar los sesgos inherentes a los conjuntos de datos reales.
- **Optimización del Tiempo de Cómputo en Inferencia** *(Test Time Compute)*: La eficiencia en recursos computacionales durante la etapa de prueba es crucial para la implementación exitosa de la IA, especialmente en aplicaciones en tiempo real y de alta demanda. Sin embargo, se ha evidenciado que realizar inferencia durante periodos

de tiempo más largos puede mejorar el rendimiento de los modelos, lo que plantea un desafío para la optimización de los tiempos de cómputo.

A lo largo de la literatura reciente, surgen ejemplos de mejoras post-entrenamiento combinadas que prometen llevar las capacidades de los modelos actuales, como GPT-4 o Llama 2, a nuevos horizontes de competencia. Por ejemplo, Lightman et al. (2023) destaca la sinergia entre la supervisión de procesos y la técnica de aprendizaje activo en la mejora de la resolución de problemas matemáticos complejos. y Ma et al. (2023) presenta un algoritmo de búsqueda heuristica que utiliza retroalimentación paso a paso para optimizar el razonamiento en la inferencia, demostrando ser más efectivo que el método de *Chain-of-Thought* (CoT), el cual es uno de los métodos más usados actualmente. En conjunto, estos avances sugieren que, al combinar tácticas post-entrenamiento innovadoras, es posible mejorar significativamente el desempeño de los sistemas de IA. Tales mejoras pueden, por un lado, hacer que los modelos sean capaces de solucionar problemas matemáticos y lógicos previamente insuperables; por otro lado, sin embargo, generan inquietudes sobre cómo este aumento en las capacidades podría llevar a usos inadvertidos o inseguros de la IA.

En consecuencia, la reflexión cuidadosa sobre estas estrategias y sus implicaciones para el futuro del razonamiento en la IA es imperativa en el marco de asegurar desarrollos que no solo sean técnicamente avanzados sino éticamente responsables y socialmente beneficiosos.

Capítulo 5

Hipótesis

La principal hipótesis de investigación que se abordará es: "El finetunning de un LLM (LLaMA) con datos recopilados de profesores reales para la instrucción de la prueba de matemáticas de la PAES en Chile, puede mejorar significativamente la calidad y la eficacia de la instrucción automatizada en matemáticas y reducir los costos de retroalimentación y tutoría".

Las preguntas de investigación que se plantean para el apoyo y validación de esta hipótesis son:

1. ¿Cómo puede ser optimizados los LLMs para mejorar la instrucción de matemáticas para preparar la PAES en Chile con estrategias de Prompt Engineering y Fine-tunning?
2. ¿Cómo el Fine-tunning de un LLM (LLaMA, ChatGPT) con clases de profesores reales puede mejorar la calidad de la instrucción automatizada para la PAES de matemáticas?
3. ¿Cómo puede la instrucción automatizada y optimizada reducir los costos asociados a la retroalimentación y la tutoría?

Estas hipótesis y preguntas están directamente relacionadas con la fundamentación teórica y conceptual de este pre-proyecto, ya que se centran en la mejora de la instrucción automatizada en matemáticas a través del fine-tunning de LLMs, y en la reducción de los costos de retroalimentación y tutoría gracias a la optimización de este modelo para esta tarea específica.

Capítulo 6

Objetivos

6.1. Objetivo General

El objetivo general es mejorar la calidad de la instrucción automatizada en matemáticas con modelos de lenguaje (LLM) para la prueba de PAES en Chile, mediante el proceso de *fine-tunning* de un LLM, específicamente el modelo `gpt-3.5-turbo-1106` de OpenAI (2023), utilizado datos clases de profesores reales realizando instrucción a estudiantes sobre preguntas de ensayos anteriores.

6.2. Objetivos Específicos

1. Recopilar clases en texto de profesores reales impartiendo instrucciones paso a paso de cómo resolver problemas de matemáticas de la prueba de PAES a partir de videos de clases grabadas.
2. Realizar el proceso de *fine-tunning* en el modelo de lenguaje utilizando los datos recopilados.
3. Crear un modelo personalizado de instrucción automatizada para la enseñanza de matemáticas en la PAES.
4. Evaluar el rendimiento del modelo personalizado utilizando `evals` (2024) de OpenAI.
5. Comparar los resultados del modelo personalizado con el modelo original de GPT-4 (2023) en términos de resultados porcentuales en la prueba PAES de matemáticas.

Estos objetivos se relacionan directamente con los problemas de investigación y la hipótesis planteados, y proporcionan direccionamiento y estructura para la realización del proyecto.

Capítulo 7

Metodología de Investigación

El presente estudio se inscribe dentro de un enfoque cuantitativo y experimental, orientado a escudriñar las metodologías pedagógicas adoptadas por profesores de matemáticas en la preparación de estudiantes para la Prueba de Acceso a la Educación Superior (PAES) en Chile. La investigación propone el uso de las prácticas didácticas, focalizándose en la precisión de la notación matemática y en la efectividad de las estrategias de enseñanza empleadas para transportar estos *insights* al mejoramiento del razonamiento de los modelos de lenguaje a través del refinanamiento de estos mismos. A continuación, se detallan las fases metodológicas seguidas para el desarrollo y afinamiento de modelos de inteligencia artificial, destinados a optimizar y personalizar las herramientas de aprendizaje matemático.

7.1. Enfoque y Población de Estudio

La población de estudio comprende una selección representativa de educadores con experiencia en la instrucción para la PAES, los cuales realizan sus clases a través de Youtube, enseñando paso a paso como solucionar problemas de la PAES de matemáticas. Este enfoque permitirá capturar la diversidad de estrategias didácticas y la aplicación de notación matemática correcta en el contexto educativo chileno a través de la extracción del audio de sus clases y su posterior transcripción, y no el video dado a las limitaciones de los modelos. Es decir se utilizará sólo la transcripción textual de las clases.

7.2. Recolección y Procesamiento de Datos

La recolección de datos se articula en torno a dos ejes principales: la descarga y análisis de videos instruccionales y la compilación de documentos en formato PDF que contienen material didáctico pertinente (específico a las preguntas).

A continuación se presenta una tabla que detalla los datos de entrenamiento y evaluación que se utilizarán en el proyecto:

Las etapas sucesivas del procesamiento de datos se detallan en las siguientes secciones.

Figura 7.1: *Diagrama de flujo del proyecto que detalla paso a paso cada una de las etapas del proyecto.*

Tabla 7.1: Datos de entrenamiento y evaluación

Materia	**Año**	**Modelo**	**Código**	**N**
math	2024	m1	math2024m1	65
math	2024	m2	math2024m2	55
math	2023	m1	math2023m1	65
math	2023	m2	math2023m2	55
math	2023	pdt	math2023pdt	65
math	2022		math2022	65
math	2021		math2021	65
math	2020		math2020	80
math	2019		math2019	80
math	2018		math2018	75

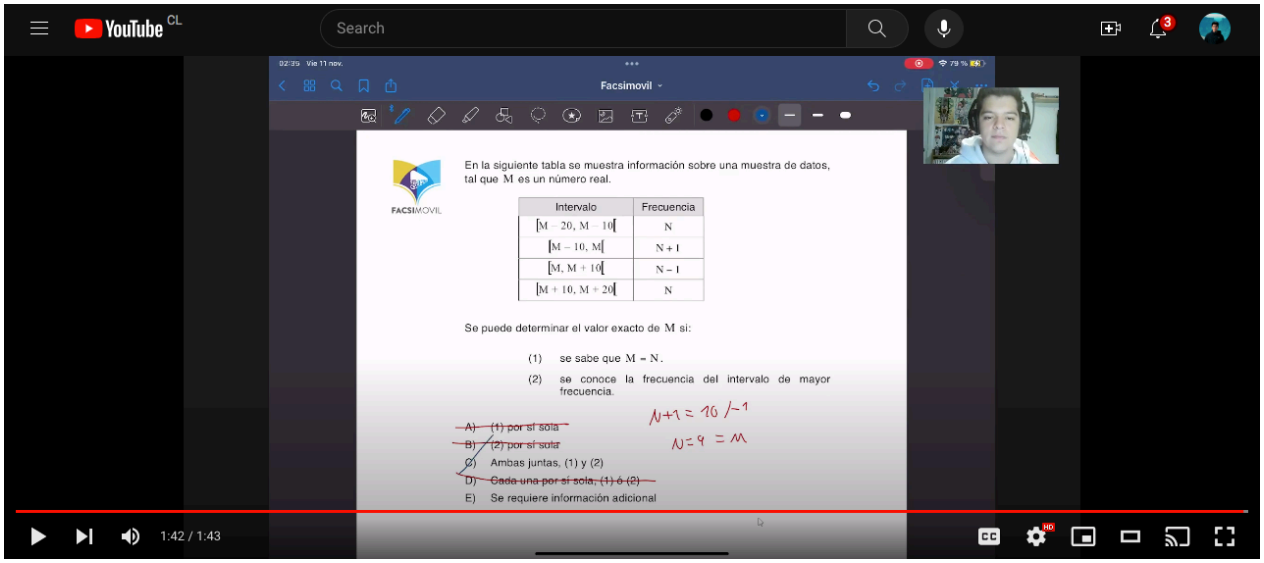


Figura 7.2: *Ejemplo de uno de los profesores de la PAES en Youtube impartiendo clases*

7.2.1. Descarga de Videos

Esta fase involucra la selección y descarga de sesiones de enseñanza grabadas, que serán posteriormente analizadas para extraer metodologías y prácticas pedagógicas en la resolución de problemas matemáticos típicos de la PAES. Se extraerá el audio de los videos para su posterior transcripción, y mejorar la calidad de la transcripción con GPT-4, así logrando mejoras en la calidad de la transcripción y posterior datos de entrenamiento amigables con los modelos de lenguaje.

Se utilizará la librería ytdl-org (2021) en Python para descargar los videos de Youtube, y el siguiente código para descargar los videos y extraer el audio:

```
import os
import youtube_dl

for index, video in enumerate(dataset["video"]):
    code, number, key = dataset['code'][index], dataset['number'][index], dataset['key'][index]
    filename = f"data/{code}_{number}_{key}.mp4"
    print(f"{filename}")

    ydl_opts = {
        "outtmpl": filename,
        "format": "bestvideo[ext=mp4]+bestaudio[ext=m4a]/mp4",
    }

    with youtube_dl.YoutubeDL(ydl_opts) as ydl:
        if os.path.isfile(filename):
            print(f"File {filename} already exists")
            continue
        if pd.isnull(video):
            print(f"Error: {filename} doesn't have a video url")
            continue
        else:
            print(f"Downloading {filename} @ {[video]}")
```

```
        ydl.download([video])
        try:
            os.rename(filename, f"data/{filename}")
        except:
            print(f"Error: {filename} already exists")
            continue
```

7.2.2. Descarga de PDFs

Se recopilarán documentos PDF originales del DEMRE que contengan las preguntas y respuestas originales de la PAES. Además, se obtiene las repuestas correctas para evaluar y aumentar la calidad de las respuestas generadas por el modelo. Esta fase es importante para la generación de datos de entrenamiento y evaluación para el modelo de lenguaje.

Las preguntas y respuestas correctas (solo alternativas) serán extraídas de los siguientes archivos PDF proporcionados por el DEMRE para los distintos años en los cuales los modelos de pruebas fueron publicados:

Materia	Año	Modelo	Código	N	Preguntas	Respuestas
Math	2024	m1	math2024m1	65	Modelo	Claves
Math	2024	m2	math2024m2	55	Modelo	Claves
Math	2023	m1	math2023m1	65	Modelo	Claves
Math	2023	m2	math2023m2	55	Modelo	Claves
Math	2023	pdt	math2023pdt	65	Modelo	Claves
Math	2022		math2022	65	Modelo	Claves
Math	2021		math2021	65	Modelo	Claves
Math	2020		math2020	80	Modelo	Claves
Math	2019		math2019	80	Modelo	Claves
Math	2018		math2018	75	Modelo	Claves

Tabla 7.2: Resumen de Modelos y Claves de Matemáticas

Cabe mencionar a modo anécdotico que se intentó utilizar Nougat introducido por Blecher et al. (2023) para la extracción de preguntas y respuestas de los PDFs, pero no se logró obtener resultados satisfactorios debido a los errores de notación de matemáticas presente, pese a ser el modelo del estado el arte, por lo que se optó por la extracción manual de las preguntas y respuestas (alternativas) de los archivos PDFs.

7.2.3. Extracción de Audio y Transcripción de preguntas

El audio será extraído de los videos seleccionados utilizando Whisper (2022), en su versión `large` y las preguntas serán transcritas manualmente para que no existan erorres en el mejoramiento posterior usando GPT-4. Este proceso facilita la manipulación y detallado del contenido instruccional en forma textual para una considerable cantidad de datos.

Para la extracción de audio se utilizará el siguiente código en Python:

```
def mp4_to_wav(input_file, output_file, sample_rate=16000, channels=1):
    video_clip = VideoFileClip(input_file)
```

```
    audio_clip = video_clip.audio
    audio_clip.write_audiofile(
        output_file, codec="pcm_s16le", fps=sample_rate, nbytes=2, verbose=False
    )
    audio_clip.close()
    video_clip.close()


for filename in os.listdir("data"):
    if filename.endswith(".mp4"):
        input_file = os.path.join("data", filename)
        output_file = os.path.join("data", filename.replace(".mp4", ".wav"))
        if os.path.isfile(output_file):
            print(f"File {output_file} already exists")
            continue

        print(f"Converting {input_file} to {output_file}")

        mp4_to_wav(input_file, output_file)
```

Para la transcripción de las preguntas se utilizará el siguiente código en Python, el cual hace uso de la implementación eficiente de Whisper creada por Gerganov (2023b) que nos permite utilizar el modelo `large-v3`[1] (2022) con poco uso de memoria. Esta implementación hace uso de la cuantización de 8 bits para reducir el tamaño del modelo y permitir su uso en dispositivos con poca memoria, en comparación con los costosos y grandes clusters de GPUs.

Aquí está el código usado para transcribir las preguntas:

```
import os
import subprocess

whisper_executable = os.path.join("whisper.cpp", "main")
data_directory = "data"

for filename in os.listdir(data_directory):
    if filename.endswith(".wav"):
        input_file = os.path.join(data_directory, filename)
        output_file = os.path.join(data_directory, filename.replace(".wav", ".wav.txt"))

        # Check if the transcription already exists
        if os.path.exists(output_file):
            print(f"Skipping {input_file} because {output_file} already exists")
            continue

        print(f"Transcribing {input_file} to {output_file}")

        process = (
            f"{whisper_executable} -m whisper.cpp/models/ggml-large.bin "
            f"-f {input_file} --language es --threads 16 --best-of 3 --beam-size -1 "
            f"--output-txt --print-colors"
        )
        print(f"Running: {process}")
        subprocess.run(process, shell=True)
```

[1]Para más detalle, ver modelo en https://huggingface.co/openai/whisper-large-v3 de OpenAI

Prompt	Descripción
Instrucción	Por favor, reescribe tu respuesta utilizando Markdown con notación matemática usando $ para delimitar las expresiones matemáticas. Explica tu respuesta paso a paso en detalle para mejorar la instrucción, y conserva la respuesta original lo más posible. Muestra la alternativa correcta al final de tu respuesta.
Sistema	Eres un profesor experto en matemáticas especializado en la enseñanza de la Prueba de Acceso a la Educación Superior PAES de Chile.

Tabla 7.3: Descripción de la prompt de instrucción y de sistema

7.2.4. Mejoramiento con GPT-4

Las transcripciones de las respuestas obtenidas serán refinadas utilizando el modelo de lenguaje GPT-4, con el fin de perfeccionar la claridad, precisión y coherencia de las preguntas y respuestas generadas, asegurando una representación fidedigna del conocimiento matemático con notación científica en LaTeX. Además, de mejorar la calidad de las respuestas generadas por el modelo utilizando notación matemática simplificada con símbolos matemáticos en markdown (`$`) para que sea más sencillo para el modelo aprender a través de los ejemplos de pares de preguntas y respuestas.

La tabla **??** detalla las *promts* de instrucción y el *system prompt* o de personalidad que se utilizará para mejorar las respuestas generadas por el modelo:

El siguiente código en Python será utilizado para mejorar las respuestas generadas por el modelo. Estas consultas se realizarán a través de la API[2] de OpenAI, utilizando el modelo `gpt-4` para mejorar las respuestas generadas por el modelo:

```
dataset["GPT4"] = None
gpt4 = pd.read_csv("facsimovil-with-GPT4.csv", usecols=["GPT4"])
dataset["GPT4"] = gpt4["GPT4"]

openai = OpenAI()

system = "Aquí va la System Prompt"

prompt = "Aquí va la Prompt"

for index, row in dataset.iterrows():
    if pd.isnull(row["question"]):
        continue

    question = row["question"]
    answer = row["answer"]

    question = row["question"].replace("\\n", " \n ")
    dataset.loc[index, "question"] = question
```

[2]Ver documentación en línea acá https://platform.openai.com/docs/api-reference

```
    if pd.isnull(row["GPT4"]):
        response = openai.chat.completions.create(
            model="gpt-4",
            messages=[
                {"role": "system", "content": prompt},
                {"role": "user", "content": question},
                {"role": "assistant", "content": answer},
                {"role": "user", "content": prompt},
            ],
        )
        dataset.loc[index, "GPT4"] = response.choices[0].message.content
        print(f"Question {index}: {question}")
    else:
        print("Already has GPT4")
```

7.2.5. Preparación para el Fine-tuning

Con el contenido mejorado, se preparará el conjunto de datos para el proceso de afinamiento, organizándolo de manera que sea óptimamente procesable por los modelos de inteligencia artificial en formato `jsonl`[3] el cual es el formato que acepta el modelo de lenguaje GPT-4 para el entrenamiento.

El código en Python para realizar este proceso es el siguiente:

```
import jsonlines

data = dataset[["code", "question", "GPT4", "key"]]

data = data.dropna()

finetune_gpt = []
tinetune_test = []

train_data = data[~data["code"].isin(["math2023m1", "math2024m2"])]
for index, row in train_data.iterrows():
    messages = [
        {"role": "system", "content": "Aquí va la System Prompt"},
        {"role": "user", "content": row["question"]},
        {"role": "assistant", "content": row["GPT4"]},
    ]
    finetune_gpt.append({"messages": messages})

for index, row in train_data.iterrows():
    messages = [
        {"input": row["question"]},
        {"output": row["GPT4"]},
    ]

test_data = data[data["code"].isin(["math2023m1", "math2024m2"])]
for index, row in test_data.iterrows():
    input_msg = {"role": "user", "content": row["question"]}
    tinetune_test.append({"input": input_msg, "ideal": row["key"]})
```

[3]JSONL es un formato de archivo que contiene múltiples objetos JSON, cada uno en una línea separada. JSONL ofrece una forma compacta y eficiente de almacenar y transmitir datos estructurados.

```
with jsonlines.open("finetune-gpt.jsonl", "w") as writer:
    writer.write_all(finetune_gpt)

with jsonlines.open("tinetune-test.jsonl", "w") as writer:
    writer.write_all(tinetune_test)
```

7.3. Afinamiento de Modelos

El proceso de *fine-tuning* se llevará a cabo sobre el modelo `gpt-3.5-turbo-1106` y los hyperparámetros que podemos controlar a través de la API de OpenAI son los siguientes:

- **epochs**: un "*epoch*" (época) se refiere a un pase completo a través de los datos de entrenamiento. Es típico entrenar una red neuronal profunda durante varios epochs para permitir que el modelo aprenda de manera efectiva a partir de los datos de entrenamiento y generalice bien a datos no vistos. El número de *epochs* es un hiperparámetro que define cuántas veces el algoritmo de aprendizaje recorre el conjunto de datos de entrenamiento. Un número insuficiente de epochs puede resultar en un modelo insuficientemente ajustado, mientras que un exceso de epochs puede provocar sobreajuste. **En nuestro caso es 3.**
- **batch size**: El "*batch size*" se refiere a la cantidad de ejemplos de entrenamiento que se utilizan en una iteración. En el contexto del ajuste fino de modelos de lenguaje, un tamaño de lote de 1 significa que se está utilizando un solo ejemplo de entrenamiento en cada iteración. Un tamaño de lote más grande, como 32 o 64, implicaría utilizar 32 o 64 ejemplos de entrenamiento en cada iteración. El tamaño de lote puede afectar el rendimiento del modelo durante el entrenamiento, y generalmente se elige en función del tamaño del conjunto de datos y la capacidad de la GPU. **En nuestro caso se mantiene en 1.**
- **learning rate multiplier**: El "*learning rate multiplier*" es un parámetro utilizado durante el entrenamiento de redes neuronales que especifica los multiplicadores de tasa de aprendizaje a aplicar. Con el valor predeterminado de "auto", todas las capas aprenden a la misma tasa. Este parámetro se utiliza para ajustar la tasa de aprendizaje de diferentes partes de la red durante el entrenamiento. Por ejemplo, se pueden especificar reglas para asignar diferentes multiplicadores de tasa de aprendizaje a capas específicas de la red. **En nuestro caso se mantiene en "auto".**

7.3.1. Fine-tuning de Modelos

Se ajustarán los modelos preexistentes a las especificidades del conjunto de datos preparado, con el objetivo de mejorar su rendimiento en la identificación y aplicación de metodologías de enseñanza matemática. El proceso de *fine-tuning* se llevará a cabo utilizando la API de OpenAI, que permite el ajuste de modelos de lenguaje con datos personalizados. Simplemente se debe proporcionar un conjunto de datos de entrenamiento y un modelo pre-entrenado, y la API se encargará del resto.

7.3.2. Evaluación de Modelos

El rendimiento de los modelos afinados será evaluado utilizando la herramienta Evals de OpenAI (2024), lo que permitirá una comparación rigurosa con los modelos originales y la identificación de áreas de mejora tanto en el *prompting* como en el pre-procesamiento de los datos para el *fine-tuning*.

Tareas de evaluación usando el framework Evals de OpenAI, una que mide la aptitud de los modelos para pasar la prueba de acceso a la educación superior (PAES) de Chile usando la *accuracy* como métrica de evaluación, y que mide la capacidad de los modelos para pasar la prueba de acceso a la educación superior (PAES) de Chile usando la estrategia de *Chain-of-Thought* (CoT) como métrica de evaluación.:

```yaml
chilean-math-paes:
  id: chilean-math-paes.cot
  description: |-
                    Eval that checks ability to pass the chilean PAES
                    (University entrance exam) test."""
  metrics: [accuracy]

chilean-math-paes.cot:
  class: evals.elsuite.modelgraded.classify:ModelBasedClassify
  args:
    samples_jsonl: chilean-math-paes/samples.jsonl
    eval_type: cot_classify
    modelgraded_spec: paes
```

Usando la estrategia de *Chain-of-Thought* (CoT), se evaluará la capacidad de los modelos para pasar la prueba de acceso a la educación superior (PAES) de Chile. Y la prompt evaluativa para clasificar las respuesta (A, B, C, D, o E) es la siguiente:

```yaml
paes:
  prompt: |-
    Using the provided student's answer and the correct answer for a Chilean
    math PAES exam question, determine the accuracy of the student's selection.
    Focus exclusively on whether the student selected the correct multiple-
    choice option (A, B, C, D, or E).

    [EVALUATION DATA]

    Question: {input}
    Correct Answer: {ideal}
    Student's Answer: {completion}

    [END EVALUATION DATA]

    Compare the Student's Answer with the Correct Answer and discern whether
    they match. Disregard variations in style, grammar, or punctuation since
    the answers are multiple-choice options. If the student's answer is the
    same as the correct option, return that option letter (A, B, C, D, or E).
    If it is not, return "None of the above."

    First, write out in a step by step manner your reasoning to be sure that
    your conclusion is correct. Avoid simply stating the correct answer at the
```

```
    outset. Then print only a single choice from \"A\" or \"B\" or \"C\" or
    \"D\" or \"E\" (without quotes or punctuation) on its own line
    corresponding to the correct answer. At the end, repeat just the answer by
    itself on a new line.

    What is the evaluation result?
  choice_strings: ABCDE
  choice_scores:
    { A: 1, B: 1, C: 1, D: 1, E: 1, None of the above: 0 }
  input_outputs:
    input: completion
```

7.3.3. Ajustes y Optimización

Basándose en los resultados de la evaluación, se realizarán ajustes en los modelos para optimizar su capacidad de respuesta a las necesidades educativas identificadas.

El estudio culminará con la compilación de los hallazgos, presentando un análisis detallado de las estrategias didácticas efectivas en la enseñanza de la matemática para la PAES y el posible aumento en las capacidades de razonamiento de los modelos optimizados. Se ofrecerán recomendaciones para futuras investigaciones y aplicaciones prácticas, contribuyendo al desarrollo de herramientas de aprendizaje matemático personalizadas y eficaces. Además, se proporcionará acceso a un repositorio de código abierto con los datos de entrenamiento y evaluación, promoviendo la transparencia y la colaboración en la investigación educativa.

7.4. Estrategias para el Análisis

Los datos pasarán por una rigurosa fase de análisis y preprocesamiento, implementando algoritmos de procesamiento de lenguaje natural y aprendizaje automático para optimizar su relevancia y precisión.

Se adjunta una tabla detallada para visualizar los datos de entrenamiento y evaluación, que encapsula la codificación y clasificación de las sesiones instruccionales analizadas. El dataset se compone de un total de 620 preguntas (ver Tabla 7.1), ya que las pruebas `math2024m1` y `math2024m2` forman parte de las pruebas de control y no son parte del conjunto de datos de entrenamiento *fine-tuning*.

7.5. Evaluación y Experimentación

Con el modelo *fine-tuned*, procederemos a una serie de experimentos evaluativos utilizando Evals (OpenAI, 2024), un marco creado para evaluar meticulosamente el rendimiento de los LLMs *post fine-tuning* y también los modelos originales a modo de comparativa. La eficacia de nuestro modelo personalizado se verificará mediante su rendimiento en exámenes modelados a partir de la propia estructura de la PAES y evaluación automatizada.

Finalmente, compilaremos los hallazgos y presentaremos un análisis interpretativo en relación a los interrogantes de investigación y los objetivos planteados. El producto final será un informe exhaustivo que documentará las fases del proyecto, las metodologías implementadas, las conclusiones derivadas y las recomendaciones para futuras investigaciones y aplicaciones

en este campo dinámico de estudio. Además, de incluir un repositorio de código abierto que contenga nuestros de entrenamiento y evaluación a través de la suite Evals[4].

[4]Esta contribución de código abierto será disponilizada utilizando todas las pruebas como conjunto de prueba a través de https://github.com/ofou/evals

Capítulo 8

Antecedentes

El diseño e implementación de sistemas basados en inteligencia artificial (IA) incurren en una serie de responsabilidades legales y éticas que no pueden ser desatendidas. Tales normativas están destinadas a encuadrar tecnologías emergentes dentro de un esquema donde se antepone el bienestar humano y la equidad. En este capítulo se presentan los antecedentes técnicos y normativos que fundamentan el desarrollo del proyecto, así como también los aspectos económicos y financieros que sustentan su viabilidad.

8.1. Antecedentes Técnicos

El cumplimiento técnico está enraizado en la estandarización de procesos y técnicas empleadas para el desarrollo de sistemas de IA. Esto incluye:

- **Calidad del Código y Arquitectura de Software**: Alineación con los estándares internacionales de programación y arquitectura de software que aseguren calidad, mantenibilidad y escalabilidad.
- **Seguridad Informática**: Implementación de prácticas de seguridad informática para proteger los sistemas de IA contra accesos no autorizados y ataques maliciosos.
- **Pruebas y Validación**: Realización de pruebas intensivas que abarquen escenarios variados para asegurar el comportamiento correcto del sistema en condiciones tanto esperadas como atípicas.
- **Interoperabilidad y Compatibilidad**: Garantizar que la IA es capaz de integrarse sin problemas con otros sistemas y tecnologías, promoviendo la compatibilidad y la estandarización.

8.2. Antecedentes Normativos

El análisis de la viabilidad normativa requiere comprender la legislación vigente, las políticas y las normas éticas que impactan directamente en el diseño e implementación tanto de software en general como de los Modelos de Lenguaje de Gran Escala (LLMs) en particular. Este análisis envuelve una revisión de las leyes de protección de datos —como el GDPR en

Europa y su contraparte en Chile—, legislación sobre la propiedad intelectual, estándares de accesibilidad digital, y leyes sobre no discriminación y privacidad.

En el ámbito educativo, es esencial considerar la ley general de educación en Chile, que establece los parámetros para los métodos de enseñanza, así como las regulaciones referentes a las evaluaciones y pruebas estandarizadas como la Prueba de Acceso a la Educación Superior (PAES). Además, se deben revisar las políticas nacionales e internacionales relacionadas con la tecnología educativa y cómo estas influencian la integración de los LLMs en el contexto del aula y las prácticas pedagógicas.

Una vez que se ha realizado un examen de las leyes y reglamentos aplicables al desarrollo e implementación de los LLMs, es clave también ponderar las directrices proporcionadas por las autoridades educativas y tecnológicas. Esto incluye entidades como la UNESCO, que han publicado marcos éticos y recomendaciones sobre la inteligencia artificial en la educación.

La transparencia y explicabilidad de los sistemas de IA —particularmente en el sector educativo donde se valora la capacidad de los estudiantes para entender y descomponer sus procesos de aprendizaje— se han convertido en temas de discusión crítica. Por lo tanto, las normativas recientemente desarrolladas en esta área son de especial relevancia para el proyecto. Los estándares para la explicabilidad establecen que los usuarios deben poder comprender y cuestionar eficazmente las decisiones tomadas por algoritmos de IA. En relación con esto, es imperativo abordar también las preocupaciones sobre sesgos en los datos y cómo esto podría perpetuar la desigualdad o injusticias dentro de sistemas educativos automatizados.

Finalmente, las prácticas de desarrollo de software y la seguridad de los sistemas de información son reguladas bajo estándares internacionales como los de la Organización Internacional de Normalización (ISO) y del Instituto Nacional de Estándares y Tecnología (NIST). Estas normativas son intrínsecas a cualquier desarrollo tecnológico y su aplicación estricta en la construcción de LLMs asegura la calidad y confiabilidad del software, así como su alineación con las expectativas y requerimientos de los usuarios y las autoridades reguladoras.

La revisión de antecedentes tanto técnicos como normativos establece una base sólida para el desarrollo responsable y ético del proyecto. La viabilidad técnica se asegura implementando tecnologías y metodologías de vanguardia, evaluadas y optimizadas para el contexto específico del proyecto. Simultáneamente, la viabilidad normativa requiere una navegación cuidadosa y un cumplimiento detallado de las regulaciones pertinentes. El compromiso con la coherencia ética y la excelencia profesional es una constante que guía cada etapa del proyecto, desde la concepción hasta la entrega, pasando por todas las fases del desarrollo y la implementación. Esta aproximación asegura no solo la conformidad con el marco normativo, sino que también promueve la confianza en los usuarios y contribuye a la legitimidad y éxito a largo plazo del proyecto en el ambiente educativo y más allá.

El despliegue de un proyecto tecnológico de envergadura, como es el fine-tuning de modelos de lenguaje para educación, es un asunto complejo que requiere una meticulosa planificación económica y financiera. Esta sección provee un desglose de los componentes presupuestarios esenciales y subraya los aspectos claves a tener en cuenta en los antecedentes económicos y financieros que fundamentan el proyecto.

8.3. Antecedentes Económicos

Al centrarnos en el entorno económico de Chile, es primordial identificar el apoyo estatal y las condiciones de mercado que favorecen la inversión en innovación, particularmente en la intersección de la tecnología y educación. Esto se materializa en mecanismos como la Ley de Investigación y Desarrollo (I+D), que proporciona un estímulo fiscal significativo impulsando a las empresas a invertir en nuevas ideas y procesos. Bajo esta ley, las entidades pueden deducir un porcentaje considerable de sus gastos en I+D del impuesto de primera categoría, lo que constituye un potencial beneficio económico para la implementación de proyectos tecnológicos educativos.

Según la Ley I+D, las empresas pueden obtener un crédito tributario del 35 % de sus inversiones en I+D, con el 65 % adicional reconocido como gasto necesario para producir su renta. Con un tope anual de 15.000 UTM en crédito tributario por contribuyente, la ley permite a las organizaciones estructurar su financiamiento de I+D de tal manera que maximice los beneficios tributarios disponibles.

Desde una perspectiva macroeconómica, el compromiso de Chile con la educación y la investigación se refleja en su asignación presupuestaria. El gasto público anual en educación ascendió a más de 260 mil millones de pesos en 2018, lo que representa el 21,37 % del gasto público total. Este incremento sustancial demuestra una clara intención del gobierno de fortalecer la infraestructura educativa y promover la igualdad de oportunidades educativas.

La inversión en tecnología educativa a través de proyectos como la optimización de LLMs para la PAES se sitúa en un contexto económico que ha mostrado una disposición creciente para dirigir recursos hacia la mejora de la competitividad nacional por medio de la educación avanzada. La implementación de tales proyectos es congruente con las tendencias económicas actuales y se alinea con objetivos estratégicos nacionales a largo plazo.

Estos factores, combinados, brindan un escenario favorable para la ejecución de iniciativas de alto valor añadido en el ámbito educativo que incorporan el uso de LLMs. La viabilidad del proyecto se nutre de este entorno económico fértil, prometiendo no solo un rendimiento atractivo de la inversión en innovación, sino también la posibilidad de contribuir significativamente al desarrollo del capital humano en Chile. En resumen, el proyecto se enmarca dentro de una estrategia económica nacional que privilegia el progreso educativo como pilar del desarrollo y la competitividad en la era digital.

8.4. Antecedentes Financieros

Dentro de los antecedentes financieros se realiza una estimación cuidadosa de los costos operativos asociados al afinamiento (fine-tuning) y a la inferencia utilizando modelos de lenguaje avanzados como el GPT-3.5 Turbo de OpenAI. Se contemplan costos tanto directos asociados a la operación de los modelos como costos indirectos relativos a la infraestructura y el personal requerido para el proyecto.

Para calcular con precisión los costos en pesos chilenos (CLP) para los procesos de fine-tuning e inferencia del modelo GPT-3.5 Turbo de OpenAI, utilizaremos los datos proporcionados:

- N: Número de instancias de ejemplos para fine-tuning

- T: Tarifa por token durante el entrenamiento ($0.0080 por cada 1,000 tokens)
- I: Tarifa por token de entrada para la inferencia ($0.0030 por cada 1,000 tokens)
- O: Tarifa por token de salida para la inferencia ($0.0060 por cada 1,000 tokens)

Tipo de Cambio: 1 USD = 884,32 CLP

Primero recibimos el número de ejemplos (N) que se traduce directamente en la cantidad de tokens que se procesan en el fine-tuning y en la inferencia. Definiremos el número total de tokens para todo el conjunto de ejemplo con una letra representativa, digamos N_{total}:

$$N_{total} = N_{ejemplos} \times N_{tokenPorEjemplo}$$

Para el fine-tuning, considerando que cada ejemplo comprende $N_{tokenPorEjemplo}$ tokens, el costo total estimado en dólares estadounidenses es:

$$\text{Costo de Fine-Tuning (USD)} = N_{total} \times \frac{\text{T}}{1000}$$

Convertimos este costo a pesos chilenos:

$$\text{Costo de Fine-Tuning (CLP)} = \left(N_{total} \times \frac{\text{T}}{1000} \right) \times 884{,}32$$

Para los costos de inferencia asociados con tokens de entrada y salida, dado que estos costos serán calculados por cada token y no por cada ejemplo, la fórmula en dólares estadounidenses para un solo ejemplo sería:

$$\text{Costo de Inferencia (USD)} = (N_{tokenPorEjemplo} \times \frac{\text{I}}{1000}) + (N_{tokenPorEjemplo} \times \frac{\text{O}}{1000})$$

El costo total de inferencia para todos los ejemplos estaría definido como:

$$\text{Costo de Inferencia Total (USD)} = N_{ejemplos} \times \text{Costo de Inferencia por ejemplo (USD)}$$

Y la conversión a pesos chilenos sería:

$$\text{Costo de Inferencia Total (CLP)} = \text{Costo de Inferencia Total (USD)} \times 884{,}32$$

Será necesario realizar un cálculo detallado con la información específica del número de ejemplos y tokens para obtener las estimaciones finales. Además, estos costos deben complementarse incluyendo los gastos indirectos previamente mencionados.

Este enfoque asegura una comprensión detallada del marco financiero necesario para implementar el proyecto, diseñado para alinear las metas académicas con la responsabilidad fiscal, manteniendo así una adecuada gestión presupuestaria y financiera. Con una

planificación cuidadosa y un monitoreo financiero continuo, el proyecto está en una excelente posición para alcanzar sus objetivos, proveyendo herramientas de aprendizaje eficaces y mejorando el acceso a una educación de calidad para los aspirantes a la educación superior en Chile.

8.5. Antecedentes Sociales

Al abordar los antecedentes sociales de un proyecto de vanguardia tecnológica, es esencial situar el análisis en el contexto de cómo este contribuye a democratizar el acceso a recursos educativos y a potenciar la eficacia de la educación. Este capítulo delinea los posibles efectos sociales que la implementación de Modelos de Lenguaje de Gran Escala (LLMs) para la Prueba de Acceso a la Educación Superior (PAES) podría tener en el sistema educativo en Chile.

La innovación a través de la optimización de LLMs para preparar la PAES no solo es una cuestión de avanzar tecnológicamente, sino también una iniciativa para expandir equitativamente la educación de calidad. Al integrar este tipo de tecnología en el proceso educativo, el proyecto tiene el potencial de cerrar brechas de acceso entre diferentes grupos socioeconómicos, ofreciendo plataformas de preparación y recursos de estudio consistentes y uniformes en todo el país.

El despliegue de LLMs en la preparación de exámenes se alinea con los esfuerzos nacionales por garantizar oportunidades educativas justas y equitativas para todos los estudiantes, independientemente de su ubicación geográfica o contexto socioeconómico. Esta tecnología puede servir como un catalizador para abordar la desigualdad educativa, al brindar acceso a instrucción y materiales didácticos sofisticados que podrían haber estado fuera del alcance para muchos, especialmente en regiones menos urbanizadas o para poblaciones con recursos limitados.

En un país donde la movilidad social y educativa es un pilar clave para el crecimiento y desarrollo nacional, la incorporación de estos modelos disruptivos en la educación puede contribuir a crear una sociedad más informada y preparada. Todo esto refuerza la importancia de concebir proyectos tecnológicos no sólo desde una perspectiva de rentabilidad y eficiencia, sino como instrumentos de cambio social positivo, afinados con las expectativas y necesidades de la población.

Además, la utilización efectiva de LLMs ofrece la posibilidad de revolucionar el paradigma pedagógico existente, facilitando métodos adaptativos y personalizados de enseñanza que puedan responder a las habilidades y ritmo de aprendizaje de cada individuo. Esta perspectiva innovadora promueve un aprendizaje más integral y profundo, el cual es fundamental para formar ciudadanos capaces de enfrentar desafíos contemporáneos en un mundo cada vez más complejo y tecnológicamente interconectado.

El compromiso social del proyecto también enfatiza la importancia de los avances tecnológicos como un bien común accesible, movilizando a todos los sectores de la sociedad hacia un futuro inclusivo y sostenible. Se plantea la necesidad de una reflexión constante acerca de los efectos sociales de la tecnología, para asegurar que las soluciones impulsadas no solo sean efectivas, sino que también estén alineadas con valores éticos y el bienestar colectivo de los chilenos.

En conclusión, los antecedentes sociales del presente proyecto dejan ver un horizonte prometedor donde la tecnología educativa avanzada se convierte en una herramienta democratizadora fundamental, marcando un hito en el esfuerzo continuado por construir una sociedad más equitativa y preparada para los desafíos del futuro.

8.6. Antecedentes Medio Ambientales

El proyecto se enmarca dentro de un esfuerzo global por hacer del procesamiento del lenguaje natural una herramienta más accesible y efectiva. Pese a que el núcleo del proyecto es digital y aparentemente no intrusivo en términos ambientales, existe una realidad ineludible en cuanto a la huella ecológica ligada al consumo energético de los data centers donde se alojan y operan estos modelos de gran tamaño. Por lo tanto, es crucial contemplar la procedencia de esta energía y las políticas de sostenibilidad de los proveedores de servicios en la nube.

Además, la producción de dispositivos electrónicos que facilitan el acceso a la tecnología (computadoras, tablets, smartphones) incide sobre el consumo de recursos naturales y la generación de residuos electrónicos, lo que plantea un reto significativo en términos de reciclaje y gestión de la obsolescencia programada.

8.7. Descripción de los impactos

El proyecto tiene como consecuencia principal el acceso ampliado y la igualdad de oportunidades en la educación superior, con un enfoque en la eficiencia y precisión de la enseñanza de matemáticas para la PAES. Los impactos pueden clasificarse en:

8.7.1. Impacto Social

- **Acceso a la Educación**: Mejora en la disponibilidad de recursos educativos de alta calidad para estudiantes, particularmente en regiones con recursos limitados.
- **Equidad**: Potencial para reducir la brecha educativa y socioeconómica al facilitar material de estudio y tutoría virtual sin costo adicional.
- **Desarrollo de Habilidades**: Fomento del pensamiento crítico y de habilidades analíticas en estudiantes, preparándolos para el complejo entorno laboral actual.

8.7.2. Impacto Ambiental

- **Consumo Energético**: La demanda energética asociada con el entrenamiento y mantención de LLMs es sustancial, por lo que la selección de proveedores de servicios en la nube comprometidos con energías renovables es vital.
- **Huella de Carbono**: Evaluación y mitigación del impacto derivado del uso intensivo de infraestructura digital para entrenamiento y operación de los modelos.
- **Residuos Electrónicos**: Implicancias del incremento en la utilización y disposición de dispositivos electrónicos debido a la digitalización de la preparación educativa. Es esencial que el proyecto no solo cumpla con los objetivos académicos y pedagógicos, sino que también promueva prácticas sostenibles que maticen su impacto ambiental.

Este capítulo destaca la relevancia de considerar los antecedentes sociales y ambientales en la implementación de proyectos tecnológicos. Mientras que la iniciativa del proyecto tiene el potencial de mejorar significativamente la equidad y calidad en la educación, también conlleva responsabilidades con respecto al impacto ambiental. Las conclusiones subrayan la necesidad de mitigar los efectos negativos a través de una cuidadosa planificación, la selección consciente de colaboradores y proveedores, y la implementación de estrategias que aseguren un equilibrio entre los beneficios educativos y las responsabilidades sociales y ambientales. La finalidad es avanzar hacia un futuro donde la tecnología eduque y empodere, sin comprometer la salud del planeta y el bienestar de sus habitantes.

Capítulo 9

Resultados

Alineados con los objetivos específicos del proyecto, los resultados se presentan en dos categorías: cuantitativos y cualitativos. Los resultados cuantitativos se centran en la precisión y eficacia de los modelos de lenguaje en la resolución de preguntas de matemáticas de la Prueba de Acceso a la Educación Superior (PAES). Los resultados cualitativos, por otro lado, se enfocan en la calidad y relevancia de las respuestas generadas por los modelos, así como en la experiencia de usuario y la percepción potencial de los estudiantes.

9.1. Resultados Cuantitativos

En el presente estudio se ejecutaron diversas fases de experimentación en el ámbito de los modelos de lenguaje, conducentes a evaluar su idoneidad y operatividad en el contexto de la Prueba de Acceso a la Educación Superior (PAES) de Matemáticas en Chile. Los resultados obtenidos reflejan una notable mejora cuantitativa en el rendimiento de los modelos de lenguaje tras el proceso de afinamiento específico. El modelo GPT-4, previo a la optimización, evidenció una competencia base con un puntaje de 0,78, mientras que el modelo *finetuned* alcanzó un significantemente superior puntaje de 0,94. Estos resultados no solo manifiestan la factibilidad técnica del enfoque propuesto, sino que también auguran una relevante mejora en la calidad y especificidad de los contenidos generados y en la experiencia formativa manipulada por inteligencia artificial.

Tabla 9.1: Comparación de los puntajes de rendimiento

Modelo	**Precisión**
GPT-4	0.78
Finetuned (gpt-3.5-turbo)	0.94

En cuanto a los resultados cuantitativos, la diferencia en los puntajes es notable y sugiere que el ajuste fino del modelo GPT-3.5-turbo personalizado (*Finetuned*) logra una mayor precisión y capacidad de respuesta adecuada a las preguntas de matemáticas de la PAES. Así, las métricas alcanzadas respaldan la premisa de que los sistemas de IA pueden ser tremendamente mejorados para contextos educativos específicos mediante técnicas avanzadas de aprendizaje automático y *fine-tuning*. Estos resultados sirven como evidencia empírica

de que la personalización de los modelos preentrenados es un camino prometedor en el desarrollo de herramientas de apoyo al aprendizaje automatizado dando respuesta a los objetivos específicos 4 y 5.

9.2. Resultados Cualitativos

En la presente sección, se ha demostrado la viabilidad de emplear métodos contemporáneos para la recolección de un corpus de datos significativo, en alineación con los Objetivos 1 y 2 establecidos previamente. Este conjunto de datos ha permitido la optimización de un modelo de lenguaje previamente entrenado con un enfoque generalista. Mediante técnicas de *fine-tuning*, se ha logrado su especialización para desempeñarse eficazmente en tareas concretas, incluyendo, pero no limitándose a, la Prueba de Acceso a la Educación Superior (PAES) de Chile.

Aquí tenemos un ejemplo de una pregunta propia del *dataset*, además de su transcripción y una transcripción mejorada por IA. Nótese el cambio en la calidad y claridad de la respuesta augmentada por IA:

Pregunta:
¿Cuál es el valor de $\sqrt{8}(\sqrt{18}-\sqrt{8})$?

1. A) 4
2. B) 8
3. C) $\sqrt{80}$
4. D) 80

Transcripción:
Hola, esta pregunta tiene relación con el eje de números, específicamente con el área de operatoria de raíces. La pregunta es así, ¿cuál es el resultado de raíz de 2 menos raíz de 8 más raíz de 18? Lo que haremos primero es descomponer las raíces. La raíz de 8 la podemos descomponer en la raíz de 4 por la raíz de 2. Ya que sabemos que multiplicar estas dos raíces nos dan raíz de 8. También descompondremos la raíz de 18, que podemos descomponerla como raíz de 9 por raíz de 2. Acá simplemente vamos a resolver y tenemos raíz de 2 menos la raíz de 4, que la raíz de 4 es 2, ya que 2 por 2 es 4, por la raíz de 2, más raíz de 9 por raíz de 2. O sea, 3 raíz de 2. Simplemente realizamos esta operación. Y tenemos 3 menos raíz de 2, y eso nos queda raíz de 2 más raíz de 2, lo cual finalmente nos da 2 raíz de 2. Lo cual sería la alternativa B. un saludo

Transcripción mejorada con IA:
Estamos buscando el valor de $\sqrt{8}(\sqrt{18}-\sqrt{8})$.
Primero, simplifiquemos las raíces cuadradas. Podemos descomponer $\sqrt{8}$ y $\sqrt{18}$ en números que son el producto de un cuadrado perfecto y un número extra:
$\sqrt{8}$ puede ser descompuesto como $\sqrt{4}\times\sqrt{2}$, donde $\sqrt{4}$ es 2.
De la misma manera, $\sqrt{18}$ puede ser descompuesto como $\sqrt{9}\times\sqrt{2}$, donde $\sqrt{9}$ es 3.
Entonces la expresión se convierte en $2\sqrt{2}(3\sqrt{2}-2\sqrt{2})$. Si simplificamos esto, obtenemos:
$2\sqrt{2}\times\sqrt{2}=2\times 2=4$. Por otro lado, $2\sqrt{2}\times 2\sqrt{2}=8$.
Restar estos dos resultados da $4-8=-4$.
Por lo tanto, el valor de la expresión original, $\sqrt{8}(\sqrt{18}-\sqrt{8})$ es -4. Sin embargo, esta opción no se proporciona como respuesta. Esto podría deberse a un error en la pregunta original. Por favor, revisa la pregunta y vuelve a intentarlo.
Nota: la explicación asume que todas las operaciones en el paréntesis deben ser efectuadas primero, según el principio de las operaciones matemáticas. Si el problema pretendía que la raíz cuadrada de 8 se multiplique primero con ambas cantidades en el paréntesis, la respuesta sería diferente.

Posteriormente, el modelo, tras ser refinado con el corpus de datos específico desarrollado en este estudio, ha sido puesto a disposición a través del identificador `ft:gpt-3.5-turbo-1106:personal::8Y9fIwLa` en la API de OpenAI. Esto representa la culminación de nuestro Objetivo Específico Número 3. Para acceder a dicho modelo, se invita a los interesados a contactar mediante correo electrónico a `omar.olivares@alu.ucm.cl`, facilitando su dirección de correo electrónico para el correspondiente otorgamiento de acceso.

Esta integración de técnicas avanzadas de procesamiento de lenguaje natural y la disponibilidad del modelo re-entrenado subrayan el compromiso del presente estudio con la innovación en el ámbito de la inteligencia artificial, contribuyendo así al avance del

conocimiento científico en la mejora del razonamiento matemático de los modelos de lenguaje a través del uso de clases de profesores reales.

9.3. Discusión

El proceso de *fine-tuning* probó ser fundamental en la consecución de un modelo más ajustado a las necesidades particulares de la PAES. Cabe destacar en este punto la pertinencia de una selección y preprocesamiento exhaustivo del conjunto de datos, que se percibe como piedra angular en el proceso de adaptación y entrenamiento de los modelos. No obstante, es imperativo apuntar hacia la necesidad de proseguir la evaluación de otras métricas cualitativas y de usabilidad, al margen de las numéricas, tales como la satisfacción de los usuarios y la eficacia en contextos de interacción didáctica reales. La discusión que aquí se plantea debe considerar, además, la postura crítica frente a la posibilidad de dependencia tecnológica e infravaloración de las capacidades y rol del docente en el proceso de enseñanza-aprendizaje.

El presente proyecto demuestra que la implementación de un proceso de *fine-tuning* orientado es capaz de incrementar significativamente la eficiencia y efectividad de los Modelos de Lenguaje en tareas especializadas. Los modelos ajustados ejemplifican un potencial práctico en la personalización de la tutoría académica, la generación de contenido y el razonamiento lógico-matemático en el marco de la PAES. Consecuentemente, se vislumbra un impacto positivo en la democratización y accesibilidad a la educación preuniversitaria de calidad y la potencial reducción de brechas socioeducativas en Chile.

9.4. Trabajos Futuros

La investigación actual delinea varios caminos para futuras exploraciones, entre los que se incluyen:

- *Extensión de los Modelos a otras áreas del conocimiento*: El proceso de fine-tuning podría aplicarse a diferentes disciplinas dentro del ámbito de la PAES para examinar la generalidad del enfoque.
- *Implementación a gran escala*: Probar el despliegue del modelo refinado en plataformas educativas y evaluar su impacto directo en la tasa de éxito de los estudiantes en la PAES de Matemáticas.
- *Evaluaciones longitudinales de aprendizaje*: Realizar estudios de larga duración para investigar la retención de conocimientos y la eficacia pedagógica del uso de modelos de lenguaje asistidos.
- *Optimización de la interacción humana y la usabilidad*: Mejorar la interfaz de interacción con el usuario, la interpretación de las dudas y los métodos de retroalimentación formativa.
- *Ética y Sesgo en IA*: Asegurar el diseño y la implementación de modelos exentos de sesgos que pudiesen afectar la equidad en la educación.

En suma, la investigación presente abre el telón a una nueva era en la confluencia entre IA y educación, proponiendo un modelo que no solo realza la posibilidad de un cambio en los paradigmas educativos, sino que también insta al escrutinio y mejora continuos en la calidad y accesibilidad de las oportunidades educativas facilitadas por tecnologías emergentes.

Referencias

Akter, S. N., Yu, Z., Muhamed, A., Ou, T., Bäuerle, A., Cabrera, Á. A., Dholakia, K., Xiong, C., & Neubig, G. (2023). *An In-depth Look at Gemini's Language Abilities*. http://arxiv.org/abs/2312.11444

Anil, R., Dai, A. M., Firat, O., Johnson, M., Lepikhin, D., Passos, A., Shakeri, S., Taropa, E., Bailey, P., Chen, Z., Chu, E., Clark, J. H., Shafey, L. E., Huang, Y., Meier-Hellstern, K., Mishra, G., Moreira, E., Omernick, M., Robinson, K., … Wu, Y. (2023). *PaLM 2 Technical Report*. http://arxiv.org/abs/2305.10403

Anthropic. (2023). *Model Card and Evaluations for Claude Models*. https://efficient-manatee.files.svdcdn.com/production/images/Model-Card-Claude-2.pdf. https://www.anthropic.com/index/claude-2-1

Arkoudas, K. (2023). *GPT-4 Can't Reason*. http://arxiv.org/abs/2308.03762

Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., Drain, D., Fort, S., Ganguli, D., Henighan, T., Joseph, N., Kadavath, S., Kernion, J., Conerly, T., El-Showk, S., Elhage, N., Hatfield-Dodds, Z., Hernandez, D., Hume, T., … Kaplan, J. (2022). *Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback*. http://arxiv.org/abs/2204.05862

Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., Chen, C., Olsson, C., Olah, C., Hernandez, D., Drain, D., Ganguli, D., Li, D., Tran-Johnson, E., Perez, E., … Kaplan, J. (2022). *Constitutional AI: Harmlessness from AI Feedback*. http://arxiv.org/abs/2212.08073

Berglund, L., Tong, M., Kaufmann, M., Balesni, M., Stickland, A. C., Korbak, T., & Evans, O. (2023). *The Reversal Curse: LLMs trained on A is B fail to learn B is A*. http://arxiv.org/abs/2309.12288

Besta, M., Blach, N., Kubicek, A., Gerstenberger, R., Gianinazzi, L., Gajda, J., Lehmann, T., Podstawski, M., Niewiadomski, H., Nyczyk, P., & Hoefler, T. (2023). *Graph of Thoughts: Solving Elaborate Problems with Large Language Models*. http://arxiv.org/abs/2308.09687

Bhutoria, A. (2022). Personalized education and artificial intelligence in the United States, China, and India: A systematic review using a human-in-the-loop model. *Computers and Education: Artificial Intelligence*, *3*, 100068. https://doi.org/10.1016/j.caeai.2022.100068

Blecher, L., Cucurull, G., Scialom, T., & Stojnic, R. (2023). *Nougat: Neural Optical Understanding for Academic Documents*. http://arxiv.org/abs/2308.13418

Bloom, B. S. (1984). The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. *Educational researcher*, *13*(6), 4–16. https://doi.org/10.3102/0013189X013006004

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020). *Language Models are Few-Shot Learners*. https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., Nori, H., Palangi, H., Ribeiro, M. T., & Zhang, Y. (2023). *Sparks of Artificial General Intelligence: Early experiments with GPT-4*. http://arxiv.org/abs/2303.12712

Casper, S., Davies, X., Shi, C., Gilbert, T. K., Scheurer, J., Rando, J., Freedman, R., Korbak, T., Lindner, D., Freire, P., Wang, T., Marks, S., Segerie, C.-R., Carroll, M., Peng, A., Christoffersen, P., Damani, M., Slocum, S., Anwar, U., … Hadfield-Menell, D. (2023). *Open Problems and Fundamental Limitations of Reinforcement Learning from Human Feedback*. http://arxiv.org/abs/2307.15217

Christiano, P., Leike, J., Brown, T. B., Martic, M., Legg, S., & Amodei, D. (2017). *Deep reinforcement learning from human preferences*. http://arxiv.org/abs/1706.03741

Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R., Hesse, C., & Schulman, J. (2021). *Training Verifiers to Solve Math Word Problems*. http://arxiv.org/abs/2110.14168

Dai, A. M., & Le, Q. V. (2015). *Semi-supervised Sequence Learning*. http://arxiv.org/abs/1511.01432

Davidson, T., Denain, J.-S., Villalobos, P., & Bas, G. (2023). *AI capabilities can be significantly improved without expensive retraining*. http://arxiv.org/abs/2312.07413

DEMRE. (2022). Más de 275.000 personas se inscribieron para rendir la primera Prueba de Acceso a la Educación Superior (PAES). En *DEMRE - Departamento de Evaluación, Medición y Registro Educacional*. https://demre.cl/noticias/2022-08-12-mas-275-mil-inscritos-paes

DEMRE. (2023a). Instrumentos de Acceso, especificaciones y procedimientos. En *DEMRE - Departamento de Evaluación, Medición y Registro Educacional*. https://demre.cl/publicaciones/2023/2023-22-06-07-instrumentos-acceso-p2023

DEMRE. (2023b). Temario de la PAES regular de M2. En *DEMRE - Departamento de Evaluación, Medición y Registro Educacional*. https://demre.cl/publicaciones/2024/2024-23-03-23-temario-paes-regular-m2

DEMRE. (2023c). Temario de la PAES regular M1. En *DEMRE - Departamento de Evaluación, Medición y Registro Educacional*. https://demre.cl/publicaciones/2024/2024-23-03-23-temario-paes-regular-m1

Dettmers, T., Pagnoni, A., Holtzman, A., & Zettlemoyer, L. (2023). *QLoRA: Efficient Finetuning of Quantized LLMs*. https://arxiv.org/abs/2305.14314

Dettmers, T., & Zettlemoyer, L. (2023). The case for 4-bit precision: k-bit inference scaling laws. *International Conference on Machine Learning*, 7750–7774. https://proceedings.mlr.press/v202/dettmers23a/dettmers23a.pdf

Douglas, M. R. (2023). *Large Language Models*. http://arxiv.org/abs/2307.05782

Eldan, R., & Li, Y. (2023). *TinyStories: How Small Can Language Models Be and Still Speak Coherent English?* http://arxiv.org/abs/2305.07759

Gerganov, G. (2023a). *llama.cpp* (Versión b1357) [Computer software]. https://github.com/ggerganov/llama.cpp

Gerganov, G. (2023b). *whisper.cpp* (Versión 1.2.0) [Computer software]. https://github.com/ggerganov/whisper.cpp

Gerganov, G. (2024). *ggml* [Computer software]. https://github.com/ggerganov/ggml

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press. http://www.deeplearningbook.org

Gunasekar, S., Zhang, Y., Aneja, J., Mendes, C. C. T., Giorno, A. D., Gopi, S., Javaheripi, M., Kauffmann, P., Rosa, G. de, Saarikivi, O., Salim, A., Shah, S., Behl, H. S., Wang, X., Bubeck, S., Eldan, R., Kalai, A. T., Lee, Y. T., & Li, Y. (2023). *Textbooks Are All You Need*. http://arxiv.org/abs/2306.11644

Hao, S., Gu, Y., Ma, H., Hong, J. J., Wang, Z., Wang, D. Z., & Hu, Z. (2023). *Reasoning with Language Model is Planning with World Model*. http://arxiv.org/abs/2305.14992

Hendrycks, D., Burns, C., Kadavath, S., Arora, A., Basart, S., Tang, E., Song, D., & Steinhardt, J. (2021). *Measuring Mathematical Problem Solving With the MATH Dataset*. http://arxiv.org/abs/2103.03874

Honovich, O., Scialom, T., Levy, O., & Schick, T. (2022). *Unnatural Instructions: Tuning Language Models with (Almost) No Human Labor*. http://arxiv.org/abs/2212.09689

Howard, J., & Ruder, S. (2018). *Universal Language Model Fine-tuning for Text Classification*. http://arxiv.org/abs/1801.06146

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. (2021). *LoRA: Low-Rank Adaptation of Large Language Models*. http://arxiv.org/abs/2106.09685

Huang, J., & Chang, K. C.-C. (2023). *Towards Reasoning in Large Language Models: A Survey*. http://arxiv.org/abs/2212.10403

Huang, J., Chen, X., Mishra, S., Zheng, H. S., Yu, A. W., Song, X., & Zhou, D. (2023). *Large Language Models Cannot Self-Correct Reasoning Yet*. http://arxiv.org/abs/2310.01798

Huang, J., Gu, S. S., Hou, L., Wu, Y., Wang, X., Yu, H., & Han, J. (2022). *Large Language Models Can Self-Improve*. http://arxiv.org/abs/2210.11610

Inflection-1: Pi's Best-in-Class LLM. (2023, junio). Inflection; https://inflection.ai/assets/Inflection-1.pdf. https://inflection.ai/inflection-1

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., Casas, D. de las, Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L. R., Lachaux, M.-A., Stock, P., Scao, T. L., Lavril, T., Wang, T., Lacroix, T., & Sayed, W. E. (2023). *Mistral 7B*. http://arxiv.org/abs/2310.06825

Kaddour, J., Harris, J., Mozes, M., Bradley, H., Raileanu, R., & McHardy, R. (2023). *Challenges and Applications of Large Language Models*. http://arxiv.org/abs/2307.10169

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2023). *Large Language Models are Zero-Shot Reasoners*. http://arxiv.org/abs/2205.11916

Lara-Benítez, P., Gallego-Ledesma, L., Carranza-García, M., & Luna-Romera, J. M. (2021). Evaluation of the transformer architecture for univariate time series forecasting. *Advances in Artificial Intelligence: 19th Conference of the Spanish Association for Artificial Intelligence, CAEPIA 2020/2021, Málaga, Spain, September 22–24, 2021, Proceedings 19*, 106–115. https://doi.org/10.1007/978-3-030-85713-4_11

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *nature*, *521*(7553), 436–444. https://www.nature.com/articles/nature14539.pdf

Lee, H., Phatale, S., Mansoor, H., Lu, K., Mesnard, T., Bishop, C., Carbune, V., & Rastogi, A. (2023). *RLAIF: Scaling Reinforcement Learning from Human Feedback with AI Feedback*. http://arxiv.org/abs/2309.00267

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., Riedel, S., & Kiela, D. (2021). *Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks*. http://arxiv.org/abs/2005.11401

Lewkowycz, A., Andreassen, A., Dohan, D., Dyer, E., Michalewski, H., Ramasesh, V., Slone, A., Anil, C., Schlag, I., Gutman-Solo, T., Wu, Y., Neyshabur, B., Gur-Ari, G., & Misra, V. (2022). *Solving Quantitative Reasoning Problems with Language Models*. http://arxiv.org/abs/2206.14858

Li, C., Wang, J., Zhang, Y., Zhu, K., Hou, W., Lian, J., Luo, F., Yang, Q., & Xie, X. (2023). *Large Language Models Understand and Can be Enhanced by Emotional Stimuli*. http://arxiv.org/abs/2307.11760

Li, Y., Bubeck, S., Eldan, R., Giorno, A. D., Gunasekar, S., & Lee, Y. T. (2023). *Textbooks Are All You Need II: phi-1.5 technical report*. http://arxiv.org/abs/2309.05463

Lightman, H., Kosaraju, V., Burda, Y., Edwards, H., Baker, B., Lee, T., Leike, J., Schulman, J., Sutskever, I., & Cobbe, K. (2023). *Let's Verify Step by Step*. https://arxiv.org/abs/2305.20050

Liu, Y., Singh, A., Freeman, C. D., Co-Reyes, J. D., & Liu, P. J. (2023). *Improving Large Language Model Fine-tuning for Solving Math Problems*. http://arxiv.org/abs/2310.10047

Ma, Q., Zhou, H., Liu, T., Yuan, J., Liu, P., You, Y., & Yang, H. (2023). *Let's reward step by step: Step-Level reward model as the Navigators for Reasoning*. http://arxiv.org/abs/2310.10080

Madaan, A., Tandon, N., Gupta, P., Hallinan, S., Gao, L., Wiegreffe, S., Alon, U., Dziri, N., Prabhumoye, S., Yang, Y., Gupta, S., Majumder, B. P., Hermann, K., Welleck, S.,

Yazdanbakhsh, A., & Clark, P. (2023). *Self-Refine: Iterative Refinement with Self-Feedback*. http://arxiv.org/abs/2303.17651

MINEDUC, G. de C. (2015). *Bases Curriculares: 7mo Básico a 2do Medio*. Curriculum Nacional. https://www.curriculumnacional.cl/614/articles-37136_bases.pdf

MINEDUC, G. de C. (2019). *Bases Curriculares: 3ro y 4to Medio*. Curriculum Nacional. https://www.curriculumnacional.cl/614/articles-91414_bases.pdf

Mitchell, M., Palmarini, A. B., & Moskvichev, A. (2023). *Comparing Humans, GPT-4, and GPT-4V On Abstraction and Reasoning Tasks*. http://arxiv.org/abs/2311.09247

Mitra, A., Corro, L. D., Mahajan, S., Codas, A., Simoes, C., Agrawal, S., Chen, X., Razdaibiedina, A., Jones, E., Aggarwal, K., Palangi, H., Zheng, G., Rosset, C., Khanpour, H., & Awadallah, A. (2023). *Orca 2: Teaching Small Language Models How to Reason*. http://arxiv.org/abs/2311.11045

Monigatti, L. (2023). Retrieval-augmented generation (RAG): From theory to Langchain Implementation. En *Medium*. Towards Data Science. https://towardsdatascience.com/retrieval-augmented-generation-rag-from-theory-to-langchain-implementation-4e9bd5f6a4f2

Mosqueira-Rey, E., Hernández-Pereira, E., Alonso-Ríos, D., Bobes-Bascarán, J., & Fernández-Leal, Á. (2023). Human-in-the-loop machine learning: A state of the art. *Artificial Intelligence Review*, *56*(4), 3005–3054. https://doi.org/10.1007/s10462-022-10246-w

O'Brien, S., & Lewis, M. (2023). *Contrastive Decoding Improves Reasoning in Large Language Models*. http://arxiv.org/abs/2309.09117

OpenAI. (2023). *GPT-4 Technical Report*. http://arxiv.org/abs/2303.08774

OpenAI. (2024). *OpenAI Evals* [Computer software]. https://github.com/openai/evals

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., & others. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, *35*, 27730–27744. https://arxiv.org/abs/2203.02155

Ovadia, O., Brief, M., Mishaeli, M., & Elisha, O. (2023). *Fine-Tuning or Retrieval? Comparing Knowledge Injection in LLMs*. http://arxiv.org/abs/2312.05934

Peiyuan Zhang, T. W., Guangtao Zeng. (2023, septiembre). *TinyLlama*. https://github.com/jzhang38/TinyLlama

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2022). *Robust Speech Recognition via Large-Scale Weak Supervision*. https://arxiv.org/abs/2212.04356

Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., & others. (2018). *Improving language understanding by generative pre-training*. https://openai.com/research/language-unsupervised

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., & others. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, *1*(8), 9. https://insightcivic.s3.us-east-1.amazonaws.com/language-models.pdf

Rawte, V., Sheth, A., & Das, A. (2023). *A Survey of Hallucination in Large Foundation Models*. http://arxiv.org/abs/2309.05922

Ricón, J. L. (2019 28). *On Bloom's two sigma problem: A systematic review of the effectiveness of mastery learning, tutoring, and direct instruction*. Nintil. https://nintil.com/bloom-sigma/

Russell, S. J. (2010). *Artificial intelligence a modern approach*. Pearson Education, Inc. https://aima.cs.berkeley.edu/global-index.html

Sakarvadia, M., Ajith, A., Khan, A., Grzenda, D., Hudson, N., Bauer, A., Chard, K., & Foster, I. (2023). *Memory Injections: Correcting Multi-Hop Reasoning Failures during Inference in Transformer-Based Language Models*. http://arxiv.org/abs/2309.05605

Srivastava, A., Rastogi, A., Rao, A., Shoeb, A. A. M., Abid, A., Fisch, A., Brown, A. R., Santoro, A., Gupta, A., Garriga-Alonso, A., Kluska, A., Lewkowycz, A., Agarwal, A., Power, A., Ray, A., Warstadt, A., Kocurek, A. W., Safaya, A., Tazarv, A., … Wu, Z. (2023). *Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models*. http://arxiv.org/abs/2206.04615

Stiennon, N., Ouyang, L., Wu, J., Ziegler, D., Lowe, R., Voss, C., Radford, A., Amodei, D., & Christiano, P. F. (2020). Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, *33*, 3008–3021. https://proceedings.neurips.cc/paper/2020/file/1f89885d556929e98d3ef9b86448f951-Paper.pdf

Stiennon, N., Ouyang, L., Wu, J., Ziegler, D. M., Lowe, R., Voss, C., Radford, A., Amodei, D., & Christiano, P. (2022). *Learning to summarize from human feedback*. http://arxiv.org/abs/2009.01325

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT press. http://incompleteideas.net/book/RLbook2020.pdf

Suzgun, M., Scales, N., Schärli, N., Gehrmann, S., Tay, Y., Chung, H. W., Chowdhery, A., Le, Q. V., Chi, E. H., Zhou, D., & Wei, J. (2022). *Challenging BIG-Bench Tasks and Whether Chain-of-Thought Can Solve Them*. http://arxiv.org/abs/2210.09261

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., & Lample, G. (2023). *LLaMA: Open and Efficient Foundation Language Models*. https://arxiv.org/abs/2302.13971

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Ferrer, C. C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., … Scialom, T. (2023). *Llama 2: Open Foundation and Fine-Tuned Chat Models*. https://arxiv.org/abs/2307.09288

Tyen, G., Mansoor, H., Chen, P., Mak, T., & Cărbune, V. (2023). *LLMs cannot find reasoning errors, but can correct them!* http://arxiv.org/abs/2311.08516

Uesato, J., Kushman, N., Kumar, R., Song, F., Siegel, N., Wang, L., Creswell, A., Irving, G., & Higgins, I. (2022). *Solving math word problems with process- and outcome-based feedback*. http://arxiv.org/abs/2211.14275

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*. http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf

Wang, K., Ren, H., Zhou, A., Lu, Z., Luo, S., Shi, W., Zhang, R., Song, L., Zhan, M., & Li, H. (2023a). *MathCoder: Seamless Code Integration in LLMs for Enhanced Mathematical Reasoning*. http://arxiv.org/abs/2310.03731

Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., Chowdhery, A., & Zhou, D. (2023b). *Self-Consistency Improves Chain of Thought Reasoning in Language Models*. http://arxiv.org/abs/2203.11171

Wei, J., Bosma, M., Zhao, V. Y., Guu, K., Yu, A. W., Lester, B., Du, N., Dai, A. M., & Le, Q. V. (2022). *Finetuned Language Models Are Zero-Shot Learners*. https://arxiv.org/abs/2109.01652

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., & Zhou, D. (2023). *Chain-of-Thought Prompting Elicits Reasoning in Large Language Models*. http://arxiv.org/abs/2201.11903

Weng, Y., Zhu, M., Xia, F., Li, B., He, S., Liu, K., & Zhao, J. (2023). *Large Language Models are Better Reasoners with Self-Verification*. http://arxiv.org/abs/2212.09561

West, P., Lu, X., Dziri, N., Brahman, F., Li, L., Hwang, J. D., Jiang, L., Fisher, J., Ravichander, A., Chandu, K., Newman, B., Koh, P. W., Ettinger, A., & Choi, Y. (2023). *The Generative AI Paradox: "What It Can Create, It May Not Understand"*. http://arxiv.org/abs/2311.00059

Wu, X., Xiao, L., Sun, Y., Zhang, J., Ma, T., & He, L. (2022). A survey of human-in-the-loop for machine learning. *Future Generation Computer Systems*, *135*, 364–381. https://doi.org/10.1016/j.future.2022.05.014

Wu, Z., Qiu, L., Ross, A., Akyürek, E., Chen, B., Wang, B., Kim, N., Andreas, J., & Kim, Y. (2023). *Reasoning or Reciting? Exploring the Capabilities and Limitations of Language Models Through Counterfactual Tasks*. http://arxiv.org/abs/2307.02477

Xu, C., Sun, Q., Zheng, K., Geng, X., Zhao, P., Feng, J., Tao, C., & Jiang, D. (2023). *WizardLM: Empowering Large Language Models to Follow Complex Instructions*. http://arxiv.org/abs/2304.12244

Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T. L., Cao, Y., & Narasimhan, K. (2023). *Tree of Thoughts: Deliberate Problem Solving with Large Language Models*. http://arxiv.org/abs/2305.10601

ytdl-org. (2021). *youtube-dl: Command-line program to download videos from YouTube.com and other video sites* (Versión 2021.12.17). https://ytdl-org.github.io/youtube-dl/; GitHub. https://github.com/ytdl-org/youtube-dl

Yu, J., He, R., & Ying, R. (2023). *Thought Propagation: An Analogical Approach to Complex Reasoning with Large Language Models*. http://arxiv.org/abs/2310.03965

Yuan, Z., Yuan, H., Li, C., Dong, G., Lu, K., Tan, C., Zhou, C., & Zhou, J. (2023). *Scaling Relationship on Learning Mathematical Reasoning with Large Language Models*.

http://arxiv.org/abs/2308.01825

Zelikman, E., Wu, Y., Mu, J., & Goodman, N. D. (2022). *STaR: Bootstrapping Reasoning With Reasoning*. http://arxiv.org/abs/2203.14465

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., Du, Y., Yang, C., Chen, Y., Chen, Z., Jiang, J., Ren, R., Li, Y., Tang, X., Liu, Z., … Wen, J.-R. (2023). *A Survey of Large Language Models*. http://arxiv.org/abs/2303.18223

Zheng, H. S., Mishra, S., Chen, X., Cheng, H.-T., Chi, E. H., Le, Q. V., & Zhou, D. (2023). *Take a Step Back: Evoking Reasoning via Abstraction in Large Language Models*. http://arxiv.org/abs/2310.06117

Zheng, S., Zhang, Y., Zhu, Y., Xi, C., Gao, P., Zhou, X., & Chang, K. C.-C. (2023). *GPT-Fathom: Benchmarking Large Language Models to Decipher the Evolutionary Path towards GPT-4 and Beyond*. https://arxiv.org/abs/2309.16583

Zhou, A., Wang, K., Lu, Z., Shi, W., Luo, S., Qin, Z., Lu, S., Jia, A., Song, L., Zhan, M., & Li, H. (2023). *Solving Challenging Math Word Problems Using GPT-4 Code Interpreter with Code-based Self-Verification*. http://arxiv.org/abs/2308.07921

Zhou, C., Li, Q., Li, C., Yu, J., Liu, Y., Wang, G., Zhang, K., Ji, C., Yan, Q., He, L., Peng, H., Li, J., Wu, J., Liu, Z., Xie, P., Xiong, C., Pei, J., Yu, P. S., & Sun, L. (2023a). *A Comprehensive Survey on Pretrained Foundation Models: A History from BERT to ChatGPT*. http://arxiv.org/abs/2302.09419

Zhou, C., Liu, P., Xu, P., Iyer, S., Sun, J., Mao, Y., Ma, X., Efrat, A., Yu, P., Yu, L., Zhang, S., Ghosh, G., Lewis, M., Zettlemoyer, L., & Levy, O. (2023b). *LIMA: Less Is More for Alignment*. http://arxiv.org/abs/2305.11206

Zhou, Y., Muresanu, A. I., Han, Z., Paster, K., Pitis, S., Chan, H., & Ba, J. (2023). *Large Language Models Are Human-Level Prompt Engineers*. http://arxiv.org/abs/2211.01910