REPÚBLICA DEL ECUADOR

UNIVERSIDAD ESTATAL DE MILAGRO

VICERRECTORADO DE INVESTIGACIÓN Y POSGRADO

PROYECTO DE INVESTIGACIÓN Y DESARROLLO PREVIO A LA OBTENCIÓN DEL TÍTULO DE:

MAGÍSTER EN GERENCIA DE TECNOLOGÍAS DE LA INFORMACIÓN

TEMA:

Análisis de herramientas de IA Generativa para el desarrollo de aplicaciones que usan Inteligencia Artificial

Autores:

Byron Vladimir Mayorga Albán
Leopoldo Javier Álava Vinueza

Director:

Ms. Mariuxi Vinueza Morales

Milagro, Ecuador, 2024

ACEPTACIÓN DEL TUTOR

En calidad de Tutora de Proyecto de Investigación, nombrada por el Comité Académico del Programa de Maestría en Gerencia Educativa.

CERTIFICO

Que he analizado el Proyecto de Investigación con el tema **ANÁLISIS DE HERRAMIENTAS DE IA GENERATIVA PARA EL DESARROLLO DE APLICACIONES QUE USAN INTELIGENCIA ARTIFICIAL**, elaborado por **BYRON VLADIMIR MAYORGA ALBÁN** y **LEOPOLDO JAVIER ÁLAVA VINUEZA**, el mismo que reúne las condiciones y requisitos previos para ser defendido ante el tribunal examinador, para optar por el título de **MAGÍSTER EN GERENCIA DE TECNOLOGÍAS DE LA INFORMACIÓN**.

Milagro, 1 de octubre de 2024.

Firmado electrónicamente por:
MARIUXI GEOVANNA
VINUEZA MORALES

Ms. Mariuxi Vinueza Morales
C. I.: 0917189664

Declaración de autoría de la investigación

Los autores de esta investigación declaran ante el Comité Académico del Programa de **Maestría en Gerencia de Tecnologías de la Información** de la **Universidad Estatal de Milagro**, que el trabajo presentado de nuestra propia autoría no contiene material escrito por otra persona, salvo el que está referenciado debidamente en el texto; parte del presente documento o en su totalidad no ha sido aceptado para el otorgamiento de cualquier otro título de una institución nacional o extranjera.

Milagro, 15 de octubre de 2024.

Firmado electrónicamente por:
BYRON VLADIMIR MAYORGA ALBAN

Byron Vladimir Mayorga Albán
C. I.: 0913854386

Firmado electrónicamente por:
LEOPOLDO JAVIER ALAVA VINUEZA

Leopoldo Javier Álava Vinueza
C. I.: 0912211208

VICERRECTORADO DE INVESTIGACIÓN Y POSGRADO
FACULTAD DE POSGRADO
CERTIFICACIÓN DE LA DEFENSA

El TRIBUNAL CALIFICADOR previo a la obtención del título de **MAGÍSTER EN GERENCIA DE TECNOLOGIAS DE LA INFORMACION**, presentado por **LIC. MAYORGA ALBAN BYRON VLADIMIR**, otorga al presente proyecto de investigación denominado "ANÁLISIS DE RECURSOS Y HERRAMIENTAS PARA EL DESARROLLO DE APLICACIONES QUE USAN INTELIGENCIA ARTIFICIAL", las siguientes calificaciones:

TRABAJO DE TITULACION	48.40
DEFENSA ORAL	35.33
PROMEDIO	**83.73**
EQUIVALENTE	**Bueno**

Firmado electrónicamente por:
INGRID NINOSHKA
RUIZ RUIZ

Ing. RUIZ RUIZ INGRID NINOSHKA
PRESIDENTE/A DEL TRIBUNAL

Firmado electrónicamente por:
TIBISAY MILENE
LAMUS DE RODRIGUEZ

Ph.D LAMUS DE RODRÍGUEZ TIBISAY MILENE
VOCAL

Firmado electrónicamente por:
PETITA ISABEL
SALAVARRIA MELO

Mag Edu SALAVARRIA MELO PETITA ISABEL
SECRETARIO/A DEL TRIBUNAL

VICERRECTORADO DE INVESTIGACIÓN Y POSGRADO
FACULTAD DE POSGRADO
CERTIFICACIÓN DE LA DEFENSA

El TRIBUNAL CALIFICADOR previo a la obtención del título de **MAGÍSTER EN GERENCIA DE TECNOLOGIAS DE LA INFORMACION**, presentado por **LIC. ALAVA VINUEZA LEOPOLDO JAVIER**, otorga al presente proyecto de investigación denominado "ANÁLISIS DE RECURSOS Y HERRAMIENTAS PARA EL DESARROLLO DE APLICACIONES QUE USAN INTELIGENCIA ARTIFICIAL", las siguientes calificaciones:

TRABAJO DE TITULACION	48.40
DEFENSA ORAL	35.00
PROMEDIO	**83.40**
EQUIVALENTE	**Bueno**

Firmado electrónicamente por:
INGRID NINOSHKA
RUIZ RUIZ

Ing. RUIZ RUIZ INGRID NINOSHKA
PRESIDENTE/A DEL TRIBUNAL

Firmado electrónicamente por:
TIBISAY MILENE
LAMUS DE RODRIGUEZ

Ph.D LAMUS DE RODRÍGUEZ TIBISAY MILENE
VOCAL

Firmado electrónicamente por:
PETITA ISABEL
SALAVARRIA MELO

Mag Edu SALAVARRIA MELO PETITA ISABEL
SECRETARIO/A DEL TRIBUNAL

DEDICATORIA

A Silvia… por todo.

Byron.

AGRADECIMIENTOS

A mis padres, a mis hijos.

Byron.

A mi mamá, que con su incondicional apoyo ha sido mi guía en este viaje.

Leopoldo.

Cesión de Derechos de Autor

Sr. Dr.
Jorge Fabricio Guevara Viejó
Rector de la Universidad Estatal de Milagro

Presente.

Mediante el presente documento, libre y voluntariamente procedemos a hacer entrega de la Cesión de Derecho del Autor del Trabajo realizado como requisito previo para la obtención de nuestro Título de Cuarto Nivel, cuyo tema fue **ANÁLISIS DE HERRAMIENTAS DE IA GENERATIVA PARA EL DESARROLLO DE APLICACIONES QUE USAN INTELIGENCIA ARTIFICIAL** y que corresponde al Vicerrectorado de Investigación y Posgrado.

Milagro, 15 de octubre de 2024.

Firmado electrónicamente por:
BYRON VLADIMIR MAYORGA ALBAN

Byron Vladimir Mayorga Albán
C. I.: 0913854386

Firmado electrónicamente por:
LEOPOLDO JAVIER ALAVA VINUEZA

Leopoldo Javier Álava Vinueza
C. I.: 0912211208

Lista de Figuras

Lista de Tablas

Índice / Sumario

Resumen

La investigación analiza las herramientas de IA generativa, específicamente modelos de lenguaje grande (LLMs) y sus APIs, destacando su importancia en el desarrollo de aplicaciones empresariales. Se evaluaron métricas de calidad, costos, rendimiento y latencia, destacando la necesidad de una selección cuidadosa para optimizar procesos y reducir costos. El estudio utilizó revisiones bibliográficas y benchmarks para proporcionar una guía detallada y práctica sobre la selección y uso de estas herramientas, subrayando su impacto en la eficiencia operativa y competitividad empresarial.

Los resultados muestran que OpenAI lidera en calidad, seguido de Anthropic y Google, mientras que OctoAI y Replicate son opciones más rentables. Fireworks AI y OpenAI destacan en rendimiento y latencia, siendo ideales para aplicaciones en tiempo real. Las recomendaciones se centran en elegir APIs que equilibren costos y calidad, adaptándose a las necesidades de PYMES, grandes corporaciones, startups tecnológicos e instituciones académicas. La investigación concluye que una evaluación exhaustiva y la selección adecuada de herramientas de IA son importantes para maximizar la eficiencia y efectividad en el desarrollo de aplicaciones de IA.

Palabras clave:

Herramientas de IA generativa, modelos de lenguaje grande, LLMs, APIs de LLMs, calidad del modelo, costo, rendimiento, latencia, eficiencia operativa, competitividad empresarial, OpenAI, Anthropic, Google, OctoAI, Replicate, Fireworks AI, aplicaciones empresariales, optimización de procesos, benchmarks, revisión bibliográfica.

Abstract

The research analyzes generative AI tools, specifically large language models (LLMs) and their APIs, highlighting their importance in developing business applications. Quality metrics, costs, performance, and latency were evaluated, emphasizing the need for careful selection to optimize processes and reduce costs. The study utilized literature reviews and benchmarks to provide a detailed and practical guide on selecting and using these tools, underscoring their impact on operational efficiency and business competitiveness.

The results show that OpenAI leads in quality, followed by Anthropic and Google, while OctoAI and Replicate are more cost-effective options. Fireworks AI and OpenAI excel in performance and latency, making them ideal for real-time applications. Recommendations focus on choosing APIs that balance costs and quality, tailored to the needs of SMEs, large corporations, tech startups, and academic institutions. The research concludes that thorough evaluation and appropriate selection of AI tools are important for maximizing efficiency and effectiveness in AI application development.

Keywords

Generative AI tools, large language models, LLMs, LLM APIs, model quality, cost, performance, latency, operational efficiency, business competitiveness, OpenAI, Anthropic, Google, OctoAI, Replicate, Fireworks AI, business applications, process optimization, benchmarks, literature review.

Introducción

Antecedentes

Una de las tecnologías que más se está desarrollando y aumentando su influencia en millones de personas es la Inteligencia Artificial (IA). Esta tecnología está impactando en el desarrollo de aplicaciones, permitiendo la automatización de procesos y mejorando la toma de decisiones.

Este avance ha sido posible gracias a una serie de herramientas y recursos que han facilitado la integración de IA en diversos sectores. Según un informe de McKinsey (Singla et al., 2024), el 67% de las empresas invertirán más en el IA, reflejando su creciente adopción. Herramientas como frameworks de desarrollo y modelos de lenguaje grande (LLM), APIs, entre otros, son importantes para este desarrollo, optimizando sectores como la salud, la gestión empresarial y la atención al cliente.

Estudios, como el de (Brown et al., 2020a), muestran cómo los modelos de lenguaje han mejorado significativamente, subrayando la relevancia y efectividad de estas herramientas. La evaluación de herramientas de inteligencia artificial es importante debido a la gran variedad de opciones disponibles, que incluyen tanto herramientas de código abierto como propietarias. Cada herramienta tiene sus propias ventajas y limitaciones, por lo que es importante realizar comparaciones detalladas para determinar cuál es la más efectiva y adecuada para necesidades específicas. Esta evaluación cuidadosa asegura que se elijan herramientas que no solo cumplan con los requisitos técnicos, sino que también se alineen con los objetivos del proyecto y los recursos disponibles

La importancia de investigar este tema radica en la creciente dependencia de las organizaciones en aplicaciones inteligentes para mejorar sus operaciones y servicios. La correcta elección de herramientas no solo optimiza el desarrollo y la implementación de aplicaciones de IA, sino que también influye directamente en la

competitividad y eficiencia de las organizaciones. Según un estudio de Gartner (Gartner, 2024), las empresas incrementarán el uso de APIs de IA en un 80%

Además, con la rápida evolución de la tecnología, la demanda de desarrolladores capaces de implementar soluciones de IA eficaces está en constante aumento. Esta investigación no solo beneficiará a las empresas que buscan optimizar sus procesos, sino también a los desarrolladores y gerentes de proyectos que necesitan orientación sobre las mejores prácticas y herramientas disponibles.

Sobre esta investigación

El objetivo general de esta investigación es evaluar la eficacia y eficiencia de herramientas de IA Generativa para optimizar el desarrollo de aplicaciones en empresas, mejorando la eficiencia y efectividad de los proyectos tecnológicos que usan IA.

De los tipos de IA se estudiará la IA Generativa y también se estudiará la IA Conversacional y NLP (Procesamiento de Lenguaje Natural) ya que se suelen usar con frecuencia en la IA Generativa.

Esta investigación propone una evaluación detallada y práctica de las herramientas de IA disponibles, proporcionando una guía para la recomendación de herramientas de desarrollo óptimas. Los resultados de esta investigación ayudarán a las empresas a tomar decisiones informadas sobre qué herramientas utilizar, optimizando sus procesos de desarrollo y reduciendo costos.

La originalidad de esta investigación reside en su enfoque práctico y actual en la recomendación de herramientas de desarrollo de IA Generativa, basado en un análisis comparativo exhaustivo de las herramientas disponibles.

Esta investigación proporcionará una visión integral que abarca múltiples herramientas, destacando sus fortalezas y debilidades en contextos reales de desarrollo empresarial.

Este estudio utilizará un enfoque descriptivo para evaluar las herramientas de IA Generativa. Las fuentes de datos incluirán literatura académica, documentación técnica, benchmarks.

La recolección de datos se realizará mediante una revisión bibliográfica técnica. Estas metodologías permitirán identificar patrones, temas y categorías relevantes, proporcionando una evaluación estructurada de cada herramienta en términos de rendimiento, facilidad de uso, costo y aplicabilidad.

1 Capítulo I: El problema de la investigación

1.1 Planteamiento del problema

Actualmente el uso de IA Generativa es en la automatización y mejora de los procesos representan una gran oportunidad de mejorar la eficiencia de las empresas, pero al mismo tiempo crea desafíos para seleccionar la que más se ajuste a sus necesidades y entorno ya que existen diferentes opciones como herramientas de código abierto y propietario, así como la evaluación de los modelos de IA y el uso de herramientas a través de APIs.

Hay estudios que indican que la IA se usa en múltiples sectores y estudios que indican que el 57% de las empresas han implantado IA (McKinsey Analytics (2021), lo que enfatiza la necesidad de poder realizar una evaluación efectiva de las herramientas a usar.

En el momento de elegir herramientas las organizaciones enfrentan dificultades en seleccionar las que se ajusten a sus necesidades y ofrezcan un balance entre rendimiento/costo/facilidad de uso.

1.2 Delimitación del problema

Esta investigación abarcará la evaluación de herramientas de IA Generativa que las empresas usan en los últimos 5 años. Dentro de las IA generativas el análisis se centrará en:

- LLM (Large Language Model, o Modelo de Lenguaje Largo en español), son modelos entrenados con grandes cantidades de texto para comprender y generar lenguaje natural de forma coherente.
- API (Application Programming Interface, o Interfaz de Programación de Aplicaciones en español), son un conjunto de reglas y protocolos que permite que las aplicaciones usen funciones o datos de otros servicios (en este caso específico de servicios de LLMs).

1.3 Formulación del problema

La gran cantidad de herramientas de IA Generativa disponibles, en nuestro caso de estudio específicamente serían las LLMs y APIs presentan una necesidad en conocer cual seria la mejor opción para optimizar eficiencia y efectividad en su uso para el desarrollo de software. Las características para considerar incluyen: costo, calidad, rendimiento y aplicación en diferentes contextos.

¿Cuales son los APIs y LLMs que se ajustarían mejor a las necesidades de los desarrollos de software?

Aspectos a evaluar:

Delimitado: Se define que se va a evaluar las LLMs y APIs de IA generativa.

Claro: La investigación se enfoca en analizar las características que determinar cómo las diversas herramientas de IA Generativa impactan la eficiencia del desarrollo de software, la efectividad de las aplicaciones y el costo-beneficio en el desarrollo de aplicaciones empresariales.

Evidente: La necesidad de esta investigación es evidente. Un informe de Deloitte (Ammanath et al., 2021)muestra que el 67% de las empresas que adoptan IA enfrentan dificultades significativas para seleccionar las herramientas adecuadas.

Concreto: Las empresas necesitan identificar y evaluar las herramientas de IA más efectivas para optimizar el desarrollo de aplicaciones.

Relevante: La investigación es de gran relevancia para las empresas tecnológicas, ya que la selección adecuada de herramientas de IA puede mejorar significativamente la eficiencia operativa y competitividad. (Gartner, 2022).

Original: El enfoque de esta investigación es novedoso, ya que combina un análisis exhaustivo de diversas herramientas de IA ofreciendo una perspectiva práctica y aplicable en el contexto empresarial.

1.4 Preguntas de investigación

Para realizar la investigación sobre los subproblemas de este tema se han planteado las siguientes preguntas:

- **Subproblema 1:** Se requiere conocer cuáles son los mejores **LLMs** para usarse en el desarrollo y operación de aplicaciones con IA
    - ¿Cuáles son las LLMs más avanzadas para el desarrollo de aplicaciones con IA?
    - ¿Cuáles son ventajas y desventajas presenta para cada LLM seleccionada?
- **Subproblema 2:** Se requiere conocer cuáles son los mejores APIs de IA para usarse en el desarrollo y operación de aplicaciones con IA
    - ¿Cuáles son las APIs de IA más avanzadas para el desarrollo de aplicaciones con IA?
    - ¿Cuáles son ventajas y desventajas presenta para cada API seleccionada?

1.5 Determinación del tema

El estudio investigará cómo las variables independientes (características de las herramientas y de IA) influyen en las variables dependientes (eficiencia, efectividad de las aplicaciones y relación costo-beneficio), proporcionando una guía práctica para la selección de herramientas de desarrollo de IA que optimicen estos factores en el entorno empresarial actual.

1.6 Objetivo general

El objetivo general de esta investigación es evaluar y comparar herramientas y recursos de inteligencia artificial (IA), tales como modelos de lenguaje grande (LLM) y plataformas de acceso API, con el fin de determinar su impacto en la calidad, velocidad y la relación costo-beneficios en aplicaciones que utilizan IA en su operación. Esta evaluación se realizará analizando las características de rendimiento, facilidad de uso

y costos de estas herramientas, para proporcionar una guía práctica que permita optimizar el desarrollo de aplicaciones de IA Generativa en el entorno empresarial actual.

1.7 Objetivos específicos

Los objetivos específicos son:

1. Determinar cuáles son las características de los LLMs con respecto a calidad, velocidad y menos costo en la operación de aplicaciones que usan IA.
2. Determinar cuáles son las características de APIs con respecto a calidad, velocidad y menos costo en la operación de aplicaciones que usan IA.

1.8 Sistematización de Variables

1.8.1 Formulación de las Variables

Tabla 1: Declaración de las Variables

Problema	**Formulación**	**Objetivo General**	**Variables**
Se requiere conocer cuáles son las mejores herramientas para aplicaciones con IA Generativa	¿Cuáles son las herramientas de IA que ofrecen mejores características para aplicaciones IA?	Identificar las herramientas de IA que ofrecen más calidad, velocidad y relación costo-beneficio.	Independientes: • Herramientas de IA. Dependientes: • Calidad • Eficiencia • Relación Costo - beneficio
Subproblemas	**Sistematización**	**Objetivos Específicos**	**Variables**
Se requiere determinar cuáles LLMs ofrecen más calidad, eficiencia y relación costo beneficio en la operación de aplicaciones que usan IA	• ¿Cuáles son las LLMs más avanzadas? • ¿Qué ventajas y desventajas tiene cada LLM a evaluar?	Identificar los LLMs que inciden en la eficiencia y efectividad del desarrollo de software.	Independientes: • LLMs Dependientes: • Calidad • Eficiencia • Relación Costo - beneficio
Se requiere determinar cuáles APIs ofrecen más calidad, eficiencia y relación costo calidad en la operación de aplicaciones que usan IA	• ¿Cuáles son las APIs de IA con más funcionalidades? • ¿Qué ventajas y desventajas tiene cada API a evaluar?	Identificar los APIs de IA que inciden en la eficiencia y efectividad del desarrollo de software.	Independientes: • APIs de IA Dependientes: • Calidad • Eficiencia • Relación Costo - beneficio

1.8.2 Sistematización de Variables Independientes

Variable	Definición conceptual	Indicadores	Escala de medición	Escala
Herramientas de IA.	Tecnologías y plataformas diseñadas para desarrollar, entrenar, implementar y aplicar inteligencia artificial en diversas tareas y procesos, optimizando la eficiencia y la innovación.	Resultados de frameworks de evaluación.	Intervalo	Puntajes (0-100)
		Tiempos de respuesta.	Razón	Tiempos de respuesta (0-2000 ms).
		Tarifas	Razón	Precio x 1M tokens ($0-$100)
LLMs	Algoritmos de inteligencia artificial entrenados con grandes cantidades de datos textuales para generar y comprender texto coherente y contextual.	Resultados de frameworks de evaluación	Intervalo	Puntajes (0-100)
		Tiempos de respuesta	Razón	Tiempos de respuesta (0-2000 ms).
		Tarifas	Razón	Precio x 1M tokens ($0-$100)
		Tamaño	Razón	Billones de datos (0-1000 B)
APIs de IA	Interfaces de programación que permiten integrar capacidades de inteligencia artificial en aplicaciones, facilitando el acceso a servicios como reconocimiento de voz, análisis de imágenes y generación de texto	Resultados de frameworks de evaluación.	Intervalo	Puntajes (0-100)
		Tiempos de respuesta.	Razón	Tiempos de respuesta (0-2000 ms).
		Tarifas	Razón	Precio x 1M tokens ($0-$100)

1.8.3 Sistematización de Variables Dependientes

Variable	Definición conceptual	Indicadores	Escala de medición	Escala
Relación Costo – beneficio de Herramientas IA	realizar sus funciones previstas de manera óptima, maximizando el rendimiento y los resultados positivos mientras minimiza los recursos utilizados y los costos asociados.	Índices de calidad.	Intervalo	Puntajes (0-100)
		Latencia.	Razón	Latencia (0-2000 ms).
		Análisis Costo/beneficio	Razón	Escala cualitativa
Calidad, Eficiencia, Relación Costo – beneficio de LLMs	Capacidad de una herramienta de inteligencia artificial para proporcionar respuestas precisas y útiles a las consultas de los usuarios, optimizando el uso de recursos y minimizando los costos asociados.	Índices de calidad.	Intervalo	Puntajes (0-100)
		Latencia.	Razón	Latencia (0-2000 ms).
		Análisis Costo/beneficio	Razón	Escala cualitativa
Calidad, Eficiencia, Relación Costo – beneficio de APIs	Capacidad de una API de inteligencia artificial para proporcionar acceso eficiente y efectivo a modelos de lenguaje grande (LLMs), mientras minimiza los costos asociados.	Índices de calidad.	Intervalo	Puntajes (0-100)
		Latencia.	Razón	Latencia (0-2000 ms).
		Análisis Costo/beneficio	Razón	Escala cualitativa

1.9 Justificación

El uso de la IA generativa en el desarrollo de aplicaciones empresariales es un indicador de innovación y mejoras en la eficiencia de procesos, sin embargo, la gran cantidad de herramientas en el mercado presenta un desafío para las empresas que buscan obtener los mejores rendimientos para sus inversiones en estas tecnologías.

Este estudio se propone abordar esta problemática identificando y comparando las herramientas de IA y asi seleccionar según sus características la que favorece al desarrollo de aplicaciones, proporcionando una guía práctica para su selección y uso.

1.10 Alcance y limitaciones

Alcance

Este estudio evalúa herramientas y recursos de inteligencia artificial (IA) Generativa utilizados en el desarrollo de aplicaciones empresariales entre 2020 y 2024. Se enfoca en empresas tecnológicas y de servicios, analizando:

- Modelos de Lenguaje Grande (LLM): Impacto en la calidad, velocidad y costos.
- Plataformas de Acceso API: Proveedores, modelos disponibles, latencia y costos.

Limitaciones

El estudio enfrenta varias limitaciones:

- Disponibilidad de Datos: Información limitada sobre rendimiento y características de algunas herramientas.
- Innovación Tecnológica Rápida: Herramientas pueden evolucionar o ser reemplazadas durante el período de estudio.
- Impacto Económico Variable: Análisis costo-beneficio puede variar entre empresas según presupuesto y recursos.

2 CAPÍTULO II: Marco teórico referencial

2.1 Antecedentes históricos

En la década de los 50s fueron los inicios de la Inteligencia Artificial y desde entonces se han desarrollado ideas fundamentales y hubo aportaciones de personas influyentes que han permitido crecer a la Inteligencia Artificial a lo que hoy conocemos.

2.1.1 Inicios de la IA (1950-1970)

En 1956 John McCarthy durante la conferencia de Dartmouth definió el término “Inteligencia Artificial”. Esta conferencia es considerada el punto de partida formal para el estudio de la IA (McCarthy et al., 2006) En este evento se reunió a destacados investigadores para explorar el potencial de las "máquinas pensantes" y establecer las bases de lo que se convertiría en un campo clave de la investigación tecnológica

Durante este periodo, se desarrollaron los primeros programas capaces de resolver problemas algebraicos, jugar ajedrez y demostrar teoremas lógicos, avances que demostraron el creciente potencial de la Inteligencia Artificial. A la vez, la "máquina de Turing" y su prueba de Turing jugaron un papel importante en estos primeros pasos.

En su influyente artículo "Computing Machinery and Intelligence" de 1950, Alan Turing presentó la prueba de Turing, un criterio diseñado para evaluar si una máquina podía exhibir un comportamiento inteligente comparable al humano. En esta prueba, un juez interactuaba por texto tanto con una persona como con una máquina. Si el juez no lograba distinguir cuál de las respuestas provenía de la máquina, esta se consideraba capaz de "pensar" (Alan, 1950).

2.1.2 Primeros Modelos y Redes Neuronales (1970-1990)

Las primeras redes neuronales surgieron tomando como referencia el funcionamiento del cerebro humano. Un paso decisivo en este proceso fue el desarrollo del algoritmo de retropropagación o método de aprendizaje supervisado, que hizo posible entrenar estas redes de forma mucho más eficiente. Aunque en esa época las limitaciones tecnológicas, como la capacidad de procesamiento y el acceso a datos, eran

evidentes, estas investigaciones pusieron las bases para los avances que más tarde transformarían el campo de la inteligencia artificial.

2.1.3 NLP y Machine Learning (1990-2010)

A principios de los años 2000, la inteligencia artificial dio un salto significativo gracias a los avances en hardware y la capacidad de manejar grandes volúmenes de datos. En el campo del procesamiento del lenguaje natural (NLP), se introdujeron modelos estadísticos que mejoraron la habilidad de las máquinas para entender y generar lenguaje humano. Durante este periodo, también surgieron técnicas de aprendizaje profundo, una rama dentro del machine learning que utiliza redes neuronales con múltiples capas para detectar patrones complejos en grandes conjuntos de datos. Estos avances permitieron que las máquinas aprendieran a distintos niveles de abstracción, mejorando tareas como el reconocimiento de imágenes y el procesamiento de lenguaje natural (LeCun et al., 2015a).

El machine learning también comenzó a ganar más relevancia. Esta área se divide en dos enfoques principales: el aprendizaje supervisado, donde los modelos se entrenan con datos etiquetados para aprender a partir de ejemplos claros, y el aprendizaje no supervisado, en el que los modelos descubren patrones y estructuras en datos no etiquetados, aplicándose en áreas como la agrupación y la reducción de dimensionalidad.

Es este periodo también hubo avances significativos en la visión por computadora permitiendo interpretar y procesar información visual. Las primeras aplicaciones fueron el reconocimiento de caracteres y su aplicación en medicina permitiendo el análisis de imágenes médicas.

2.1.4 Modelos de Lenguaje Grande (2010-presente)

En la última década, la inteligencia artificial ha avanzado de manera notable, especialmente con la creación de modelos de lenguaje grandes (LLM). Entre los ejemplos más destacados se encuentran BERT (Devlin et al., 2019) y GPT-3 (Brown

et al., 2020b), los cuales han demostrado ser sorprendentemente capaces en tareas como la generación de texto y la traducción.

El deep learning ha sido un pilar clave en estos avances recientes, potenciando la capacidad de análisis y predicción. La llegada de las redes neuronales convolucionales (CNNs) revolucionó la visión por computadora, permitiendo el desarrollo de aplicaciones como el reconocimiento facial, la conducción autónoma y diagnósticos médicos más precisos (LeCun et al., 2015a).

Un momento destacado de esta era fue la aparición de ChatGPT-3 en 2020. Creado por OpenAI, este modelo hizo que se masificara el uso de la IA, gracias a su capacidad para generar texto coherente en lenguaje natural. Su impacto ha sido visible en sectores como el servicio al cliente, la educación y la creación de contenido (Brynjolfsson et al., 2023).

2.2 Antecedentes referenciales

La presente investigación se contextualiza en el ámbito empresarial, específicamente en la implementación de aplicaciones de inteligencia artificial (IA) Generativa en el desarrollo de software durante el periodo comprendido entre 2020 y 2024.

A continuación, se relaciona este proyecto con otras investigaciones relevantes sobre el tema, indicando los títulos y autores de dichos estudios, y se señalan las diferencias, cómo se podrían complementar.

1. “Generative AI for Software Practitioners” – IA generativa para profesionales de software

Autores: Ebert C., Louridas P.

Revista: IEEE Software (2023)

Este artículo examina cómo herramientas de inteligencia artificial generativa, como Bard, ChatGPT y CoPilot, están transformando la productividad en la ingeniería de software. Los autores investigan cómo estas herramientas pueden mejorar la

productividad, su integración en el desarrollo de software y los posibles riesgos que conllevan. También se incluyen estudios de caso y recomendaciones prácticas desde una perspectiva industrial.

Ambas investigaciones reconocen el impacto positivo de las herramientas de IA generativa en la productividad y eficiencia del desarrollo de software. Tanto el estudio de Ebert y Louridas como nuestra propia investigación resaltan cómo estas herramientas benefician la ingeniería de software.

Mientras que el artículo de Ebert y Louridas ofrece una guía práctica y estudios de caso específicos en un contexto industrial, nuestra investigación se enfoca en una evaluación comparativa más amplia de diversas herramientas de IA generativa (como LLMs y APIs). Nuestro objetivo es proporcionar una guía práctica para la selección de estas herramientas en proyectos empresariales, evaluando su impacto en la eficiencia, efectividad y relación costo-beneficio.(Ebert & Louridas, 2023)

2. “Applications of AI in Classical Software Engineering”- Aplicaciones de IA en la ingeniería de software clásica

Autores: Barenkamp M., Rebstadt J., Thomas O.

Revista: AI Perspectives (2020)

El articulo incluye entrevistas e investigaciones con desarrolladores de software quienes emplean o utilizaran herramientas de IA en sus trabajos. El objetivo es la evaluación del estado actual y analizar el uso futuro de la misma incluyendo los riesgos de su aplicación en las diferentes etapas del ciclo de desarrollo de software.

El análisis destaca los principales logros y ventajas de la IA en este campo:

- La automatización de tareas rutinarias, así como la depuración y documentación.
- El análisis de grandes cantidades de datos, esto permite la identificación de patrones.

- La evaluación de estos datos a través de redes neuronales, permitiendo así una comprensión más profunda y precisa.

Los estudios revisados coinciden en que la IA puede mejorar considerablemente el desarrollo de software, especialmente en la automatización de tareas repetitivas. Aunque el estudio de Barenkamp et al. tiene un enfoque más amplio sobre la ingeniería de software en general, nuestra investigación se centra específicamente en la IA Generativa y su impacto en el desarrollo de software dentro de un entorno empresarial. (Barenkamp et al., 2020)

3. "The Next Frontier in Software Development: AI-Augmented Software Development Processes" - La próxima frontera en el desarrollo de software: procesos de desarrollo de software mejorados con IA

Autor: Ozkaya I.

Revista: IEEE Software (2023)

Este artículo explora cómo los procesos de desarrollo de software han mejorado notablemente al aplicar la automatización de manera efectiva para superar desafíos, y sugiere que la comunidad de desarrolladores debería adoptar una mentalidad similar al incorporar herramientas de IA Generativa (Ozkaya, 2023).

Tanto Ozkaya como nuestra investigación se centran en el potencial de la inteligencia artificial para transformar el desarrollo de software, enfocándose en cómo estas herramientas pueden eliminar obstáculos y aumentar la eficiencia. Mientras que Ozkaya se concentra en optimizar los procesos de desarrollo con IA Generativa, nuestra investigación adopta una perspectiva más amplia. Evaluamos una variedad de herramientas de IA y analizamos su impacto en términos de eficiencia, efectividad y rentabilidad en proyectos empresariales. (Ozkaya, 2023)

4. "Artificial Intelligence-Based Tools in Software Development Processes: Application of ChatGPT" - Herramientas basadas en inteligencia artificial en procesos de desarrollo de software: aplicación de ChatGPT

Autores: Özpolat Z., Yildirim Ö., Karabatak M.

Revista: European Journal of Technic (2023)

Este estudio explora cómo las herramientas basadas en IA, específicamente ChatGPT, pueden mejorar los procesos tradicionales de desarrollo de software. Los autores realizaron aplicaciones prácticas en un proyecto de software basado en las respuestas proporcionadas por ChatGPT, evaluando su rendimiento en el contexto del desarrollo de software(Özpolat et al., 2023).

Ambas investigaciones (ésta y la nuestra) y reconocen el valor de herramientas específicas de IA (como ChatGPT) en la mejora de procesos de desarrollo de software, especialmente en términos de eficiencia y calidad.

El estudio de Özpolat et al. se centra en la aplicación específica de ChatGPT, y nuestra investigación evalúa una gama más amplia de herramientas de IA Generativa (LLMs y APIs) y su impacto general en la eficiencia y efectividad del desarrollo de software en empresas.

Mientras los 4 estudios mencionados proporcionan una base sólida sobre el uso de herramientas de IA en el desarrollo de software, nuestra investigación se distingue por su enfoque específico en IA Generativa y su impacto en un contexto empresarial, buscando proporcionar una guía práctica para la selección de herramientas que optimicen la eficiencia, efectividad y costo-beneficio en proyectos tecnológicos.(ÖZPOLAT et al., 2023)

2.3 Definición de Términos y Conceptos Básicos

2.3.1 **Introducción a la Inteligencia Artificial (IA)**

La inteligencia artificial (IA) se define como un campo de la informática que se enfoca en la creación de sistemas capaces de realizar tareas que normalmente requieren inteligencia humana, como el reconocimiento de voz, la toma de decisiones y la traducción de idiomas (Russell & Norvig, 2021) .La IA se ha convertido en una

tecnología importante en diversos sectores debido a su capacidad para analizar grandes volúmenes de datos, aprender de ellos y tomar decisiones informadas de manera autónoma.

En el contexto empresarial, la IA ha demostrado ser una herramienta poderosa para optimizar procesos, mejorar la eficiencia operativa y fomentar la innovación. Las empresas están incorporando tecnologías de inteligencia artificial para automatizar tareas repetitivas, lo que permite a los empleados dedicarse a actividades que aportan mayor valor (Davenport et al., 2018). Por ejemplo, los chatbots impulsados por IA se utilizan ampliamente en el servicio al cliente, proporcionando respuestas rápidas y precisas a las consultas de los usuarios, lo que mejora la experiencia del cliente y reduce los costos operativos (Huang & Rust, 2021).

Además, la IA está revolucionando la toma de decisiones empresariales mediante el análisis predictivo, que utiliza datos históricos para prever tendencias futuras y comportamientos del mercado. Esto permite a las empresas anticiparse a las necesidades de los clientes y ajustar sus estrategias en consecuencia, aumentando así su competitividad (Madhavi, 2019)

La capacidad de la IA para procesar y analizar datos en tiempo real también ayuda a las empresas a identificar oportunidades y riesgos con mayor rapidez, mejorando la agilidad organizacional).

2.3.2 Machine Learning (Aprendizaje Automático)

Permiten a las computadoras mediante algoritmos y técnica imitar a la inteligencia humana pueden aprender de los datos y de esta manera predecir o tomar acciones sin asistencia humana.

Utiliza métodos estadísticos para identificar patrones en los datos y asi poder construir modelos predictivos. Dentro del machine learning se puede emplear el aprendizaje automático supervisado, no supervisado y por refuerzo.

El aprendizaje supervisado se basa en datos etiquetados para entrenar modelos.

El no supervisado busca identificar estructuras en datos sin etiquetar.

El aprendizaje por refuerzo entrena modelos a través de un sistema de recompensas y penalizaciones.

El resultado de la aplicación del machine learning depende de la calidad y cantidad de datos disponibles, así como de la selección adecuada de algoritmos. (Mitchell, 1997)

2.3.3 Deep Learning (Aprendizaje Profundo)

El Deep Learning es una técnica avanzada que ha surgido del Machine Learning y se basa en redes neuronales artificiales con múltiples capas ocultas. Esta metodología permite a los modelos interpretar y aprender a partir de datos en diferentes niveles de abstracción, de manera similar a cómo lo hace un ser humano. Las redes neuronales profundas pueden tener decenas o incluso cientos de capas, cada una especializada en identificar características cada vez más complejas en los datos. Por ejemplo, en el reconocimiento de imágenes, las primeras capas pueden detectar bordes y formas básicas, mientras que las capas más profundas son capaces de reconocer objetos complejos o incluso escenas completas.

El Deep Learning ha revolucionado campos como la visión por computadora, el procesamiento del lenguaje natural y el reconocimiento de voz, superando el rendimiento humano en ciertas tareas específicas. Esta tecnología ha permitido logros notables y ha establecido nuevos estándares en precisión y eficiencia. Su éxito se debe en gran medida a la disponibilidad de grandes volúmenes de datos y al aumento de la potencia computacional, especialmente con el uso de GPUs.

Los modelos de Deep Learning han demostrado una capacidad excepcional para extraer patrones complejos de datos brutos, lo que los hace particularmente útiles en tareas donde las características relevantes no son fácilmente identificables por los humanos. Sin embargo, estos modelos también presentan desafíos, como la necesidad de grandes cantidades de datos de entrenamiento, la complejidad computacional y la dificultad de interpretar sus decisiones.(LeCun et al., 2015b)

2.3.4 Algoritmos de IA

Los algoritmos de inteligencia artificial son procedimientos informáticos diseñados para ejecutar tareas que habitualmente requieren la inteligencia de un ser humano. Entre estas tareas se incluyen el aprendizaje, la resolución de problemas y la toma de decisiones. Estos algoritmos abarcan una amplia gama de técnicas, desde métodos basados en reglas hasta enfoques de aprendizaje automático y razonamiento probabilístico. Los algoritmos de IA pueden clasificarse en varias categorías, incluyendo algoritmos de búsqueda y optimización, como A* y los algoritmos genéticos; algoritmos de aprendizaje supervisado, como árboles de decisión y máquinas de vectores de soporte; algoritmos de aprendizaje no supervisado, como K-means y análisis de componentes principales; y algoritmos de aprendizaje por refuerzo, como Q-learning.

Cada clase de algoritmo posee características y ventajas únicas que lo hacen más apto para resolver ciertos tipos de problemas. Por ejemplo, los algoritmos de búsqueda son útiles para encontrar soluciones óptimas en espacios de problema bien definidos, mientras que los algoritmos de aprendizaje automático son más adecuados para tareas que involucran patrones complejos en grandes conjuntos de datos. La elección del algoritmo depende de factores como la naturaleza del problema, la cantidad y calidad de los datos disponibles, y los recursos computacionales. Con el avance de la IA, estos algoritmos se vuelven cada vez más sofisticados, capaces de manejar tareas más complejas y adaptarse a entornos dinámicos. (Vásquez-Quispesivana et al., 2022)

2.3.5 Redes Neuronales

Las redes neuronales están inspiradas en el sistema nervioso humano y fueron diseñados para aprender y reconocer patrones. Se trata del tipo de aprendizaje profundo, estas redes están formadas por nodos interconectados, llamados neuronas artificiales, que se organizan en capas. Durante el proceso de entrenamiento, los pesos de las conexiones se ajustan, lo que permite a la red aprender a partir de los datos. Las redes neuronales pueden ser aplicadas en diversas tareas, incluyendo

clasificación, regresión y agrupación, gracias a su capacidad de adaptación y aprendizaje (En et al., n.d.).

2.3.6 Redes Neuronales Convolucionales (CNN)

Son un tipo especializado de redes neuronales diseñadas para trabajar con datos estructurados en forma de cuadrícula. A diferencia de las redes neuronales tradicionales, las CNN utilizan operaciones de convolución en al menos una de sus capas en lugar de las multiplicaciones de matrices generales. Esta diferencia permite que las CNN sean particularmente eficaces en tareas de visión por computadora, como clasificar imágenes, detectar objetos y segmentar imágenes. Gracias a estas capacidades, las CNN han mejorado notablemente el procesamiento y análisis visual (En et al., n.d.).

2.3.7 Redes Neuronales Recurrentes (RNN):

Las RNN son un tipo de red neuronal diseñada para trabajar con secuencias de datos. A diferencia de las redes feedforward, las RNN tienen conexiones que forman ciclos, permitiéndoles mantener información a lo largo del tiempo. Esto las convierte en herramientas especialmente valiosas para el manejo de datos secuenciales. Ejemplos prominentes de este tipo de datos incluyen el procesamiento del lenguaje natural y la traducción automática (Roch & Reybaud, n.d.).

2.3.8 Redes Neuronales Transformers

Los Transformers son una arquitectura de redes neuronales que tienen el objetivo de procesar secuencias de datos utilizando mecanismos de atención. A diferencia de las RNN, los Transformers no procesan los datos secuencialmente, sino que consideran toda la secuencia simultáneamente. Han demostrado ser excepcionalmente efectivos en tareas de procesamiento del lenguaje natural y, más recientemente, en visión por computadora. (Roch & Reybaud, n.d.)

2.3.9 Procesamiento de Lenguaje Natural (NLP)

El procesamiento del lenguaje natural (NLP), un campo de la inteligencia artificial se dedica a facilitar la interacción entre computadoras y el lenguaje humano. Combina la lingüística computacional con algoritmos de aprendizaje automático y aprendizaje profundo, lo que permite a las máquinas entender, interpretar, manipular y generar lenguaje humano. Las aplicaciones de NLP son diversas e incluyen la traducción automática, el análisis de sentimientos, la generación de texto y la creación de sistemas de diálogo. En resumen, la inteligencia artificial está transformando el panorama empresarial al ofrecer soluciones innovadoras y eficientes que impulsan el crecimiento y la competitividad. Su aplicación en áreas como la automatización, el análisis predictivo y la mejora de la experiencia del cliente destaca su relevancia y potencial en el entorno empresarial actual. (Sapiens & Alexander Gelbukh, 2010)

2.4 Tipología de IA Basada en Áreas de Aplicación

Podemos clasificar las IAs por su área de aplicación de la siguiente manera:

- IA Generativa
- Procesamiento del Lenguaje Natural (NLP)
- Visión por Computadora
- IA Predictiva
- IA Conversacional
- IA de Toma de Decisiones
- Robótica e IOT (Internet de las Cosas)
- IA Emocional

2.4.1 IA Generativa

La IA Generativa es un campo dentro de la inteligencia artificial que se enfoca en producir contenido nuevo y único utilizando datos preexistentes como base. Utiliza algoritmos y modelos avanzados, como redes neuronales profundas y modelos de lenguaje grande, para generar texto, imágenes, música y otros tipos de contenido. La

IA Generativa puede aprender patrones y estructuras a partir de grandes volúmenes de datos y luego aplicar ese conocimiento para producir resultados que imitan la creatividad humana. Ejemplos incluyen la generación de imágenes mediante modelos como DALL-E y la creación de textos con modelos como GPT (Gómez Monsalve, 2023)

Nota: Las herramientas que vamos a estudiar en esta investigación corresponden a este tipo de IA.

2.4.2 Procesamiento del Lenguaje Natural (NLP)

La PNL es un campo de la Inteligencia Artificial y la Lingüística dedicado a hacer que las computadoras comprendan y generen texto en lenguajes humanos. E sta área de la IA abarca tareas como la traducción automática, el análisis de sentimientos, el reconocimiento de voz y la generación de texto (Khurana et al., 2023).Se usan algoritmos de NLP mencionados en un subtítulo anterior.

2.4.3 Visión por Computadora

La Visión por Computadora se dedica a enseñar a las computadoras a interpretar y comprender el mundo visual. Utilizando técnicas de aprendizaje automático y redes neuronales convolucionales (CNN), obtener información significativa de imágenes, videos y otros datos visuales para realizar tareas como el reconocimiento de objetos, la detección de anomalías y el seguimiento de movimientos (IBM, 2020).

2.4.4 IA Predictiva

Necesita grandes cantidades de datos de alta calidad para poder generar buenos modelos. Estos modelos pueden incluir algoritmos como arboles de decisión, redes neuronales, regresiones que permiten analizar los datos para identificar patrones y así poder hacer predicciones y ayudar a la toma de decisiones (Sghir et al., 2023).

2.4.5 IA Conversacional

La IA conversacional permite la interacción natural con humanos a través de texto o voz, empleada en asistentes virtuales y sistemas de atención al cliente automatizados (Huang & Rust, 2021).

2.4.6 IA de Toma de Decisiones

Se refiere al uso de modelos de IA para analizar datos y proporcionar recomendaciones o automatizar decisiones empresariales. Este enfoque combina técnicas de aprendizaje automático y optimización para mejorar la eficiencia y efectividad de las decisiones en diversos contextos, desde la gestión de la cadena de suministro hasta la planificación estratégica y el mantenimiento predictivo(Jeyanthi et al., 2022).

2.4.7 Robótica e IOT (Internet de las Cosas)

Usa sensores inteligentes y dispositivos interconectados con la capacidad de respuesta de los robots, lo que le permite que los robots recolecten, analicen y actúen evaluando los datos en tiempo real lo que permite mejorar la eficiencia en las tareas tales como manufacturas, logística y agricultura. La IA incrementa estas capacidades al complementarlas con análisis predictivos y automático (Rai et al., 2021).

2.4.8 IA Emocional

La computación afectiva, también llamada así, es una rama de la inteligencia artificial que se dedica al reconocimiento, comprensión, simulación y respuesta a las emociones humanas. Esta tecnología utiliza algoritmos avanzados para analizar expresiones faciales, tono de voz y otros indicadores no verbales para inferir el estado emocional de una persona (Somers, 2019).

2.5 Herramientas de IA

En la siguiente tabla se observa la clasificación de herramientas que pertenecen a varios Tipos de IA. Las celdas destacadas representan ejemplos de herramientas que serán analizadas en este estudio, es decir, LLMs y APIs de IA Generativa.

Tabla 2: Tipos de IA Generativa vs. Herramientas de IA

Tipo de IA	Modelos de Lenguaje Grande (LLMs)	APIs	Frameworks	Librerías de ML	Visión por Computadora
IA Generativa	GPT-3 (OpenAI), DALL-E, Codex, etc.	OpenAI API, Hugging Face API, IBM Watson Studio, Jasper API, etc.	TensorFlow, PyTorch, Keras	Hugging Face Transformers, TensorFlow, PyTorch	DALL-E (OpenAI), MidJourney, RunwayML
Procesamiento del Lenguaje Natural (NLP)	BERT, GPT-3, RoBERTa, T5, XLNet	Google Cloud Natural Language, IBM Watson NLP, Microsoft Azure Text Analytics, AWS Comprehend	SpaCy, NLTK, AllenNLP, Stanford NLP	Hugging Face Transformers, SpaCy, NLTK	No aplica directamente
Visión por Computadora	No aplica directamente	Google Cloud Vision, Amazon Rekognition, Microsoft Azure Computer Vision	OpenCV, TensorFlow, PyTorch, Darknet	scikit-image, PyTorch Vision, OpenCV, Detectron2	MS Azure Computer Vision, Google Cloud Vision, AWS Rekognition
IA Predictiva	No aplica directamente	IBM Watson Studio, Google AI Platform, MS Azure ML, AWS SageMaker	TensorFlow, PyTorch, H2O.ai	scikit-learn, XGBoost, CatBoost, LightGBM	No aplica directamente
IA Conversacional	GPT-3, DialoGPT, Meena, BlenderBot	Dialogflow, MS Bot Framework, AWS Lex, Rasa API	Rasa, Botpress, Microsoft Bot Framework	Hugging Face Transformers, Rasa NLU	No aplica directamente
IA de Toma de Decisiones	No aplica directamente	Qlik Sense, Tableau con AI insights, TIBCO, FICO Decision Management Suite	KNIME, RapidMiner	scikit-learn, PyMC3, TensorFlow Decision Forests	No aplica directamente
Robótica e IoT	No aplica directamente	Google Cloud IoT, Microsoft Azure IoT Hub, Amazon AWS IoT, IBM Watson IoT	ROS (Robot Operating System), OpenRAVE	No aplica directamente	Intel RealSense, OpenCV, Google Cloud IoT, Microsoft Azure Kinect
IA Emocional	No aplica directamente	Affectiva, IBM Watson Tone Analyzer, Microsoft Azure Emotion API	TensorFlow, PyTorch, Keras	Hugging Face Transformers, PyTorch, Keras	No aplica directamente

2.6 IA Generativa, Aplicaciones y Herramientas

Es importante definir que el alcance de este trabajo de investigación abarca el estudio de herramientas de IA Generativa. Al ver ejemplos de IA Generativa también nos encontraremos que interactúan con la IA Conversacional y el Procesamiento del Lenguaje Natural (NLP) con frecuencia. Por lo que también se mencionarán herramientas que utilicen estos tipos de IA.

La IA generativa ha revolucionado el ámbito empresarial al ofrecer herramientas avanzadas para la creación de contenido y la automatización de procesos creativos.

2.6.1 Aplicaciones de la IA Generativa

La IA generativa se usa actualmente en diversas áreas de las empresas permitiendo la automatización y personalización de procesos. A continuación, algunas de las principales.

2.6.1.1 Creación de Contenido

Herramientas como GTP-4 de OpenAI permiten generar artículos, blogs, guiones y otros textos dinámicamente, mejorando la eficiencia y reduciendo costos (Bengesi et al., n.d.), estas tecnologías permiten crear imágenes a partir de texto un ejemplo de es DALL-E (Ramesh et al., 2021)

2.6.1.2 Personalización y Marketing

La IA permite automatizar y personalizar campañas publicitarias y contenidos promocionales. Modelos como BERT pueden analizar grandes volúmenes de datos de clientes y generar mensajes personalizados (Jacob et al., 2021; Raffel et al., 2020). Esto incluye la creación de correos electrónicos, anuncios y recomendaciones de productos personalizados basados en el comportamiento del usuario.

2.6.1.3 Desarrollo de Productos y Prototipos

El uso de la IA permite crear múltiples variantes de diseños de productos optimizados por diferentes criterios. Esto es muy útil en industrias como la automotriz, la

manufactura y el diseño de productos donde el tiempo de creación de los prototipos es muy importante.

2.6.2 Beneficios y Desafíos

La aplicación de la IA genera ventajas competitivas, pero también retos técnicos y algunas veces éticos. A continuación, algunos ejemplos de implementaciones exitosas.

2.6.2.1 Ventajas Competitivas

La IA generativa ofrece varias ventajas que pueden proporcionar a las empresas una ventaja competitiva significativa. Una de las principales ventajas es la capacidad de automatizar la creación de contenido, lo que permite a las empresas producir textos, imágenes y videos a gran escala y con alta calidad (Feuerriegel et al., 2024). Además, la personalización avanzada impulsada por IA puede aumentar la relevancia de las interacciones con los clientes, mejorando la experiencia del usuario y la fidelidad del cliente (Devlin et al., 2019). La capacidad de la IA para analizar grandes volúmenes de datos y generar insights también puede acelerar la toma de decisiones y la innovación en productos.

2.6.2.2 Retos Técnicos y Éticos

A pesar de los beneficios, la IA generativa también enfrenta desafíos técnicos y éticos. Técnicamente, uno de los principales retos es la necesidad de grandes volúmenes de datos y poder computacional para entrenar y ejecutar modelos generativos, lo que puede ser costoso y complejo de gestionar (Bommasani et al., 2021). Además, los modelos de IA pueden generar contenido inapropiado o sesgado si no se entrenan y supervisan adecuadamente. Éticamente, la IA generativa plantea preocupaciones sobre la autoría y propiedad intelectual, así como el potencial para la creación de desinformación (Floridi et al., 2020) Las empresas deben navegar estos desafíos implementando medidas de transparencia, responsabilidad y supervisión ética.

2.6.2.3 Ejemplos de Implementación Exitosa

Existen varios casos de éxito en la implementación de IA generativa en el ámbito empresarial. Por ejemplo, Adobe ha utilizado IA generativa para potenciar herramientas de diseño, permitiendo a los usuarios crear contenido visual de alta calidad de manera más rápida y eficiente, OpenAI ha colaborado con diversas empresas para integrar GPT-3 en aplicaciones de atención al cliente, mejorando la eficiencia y la satisfacción del cliente (Brynjolfsson et al., 2023).

Este desarrollo abarca tanto los beneficios como los desafíos de la implementación de IA generativa en las empresas, y proporciona ejemplos concretos de cómo se ha aplicado con éxito en diferentes industrias.

2.6.3 Herramientas de IA Generativa

El campo de la Inteligencia Artificial Generativa ha experimentado un crecimiento exponencial en los últimos años. Este avance ha sido posible gracias a una amplia gama de herramientas sofisticadas que permiten a desarrolladores, investigadores y empresas aprovechar el poder de la IA para generar texto, imágenes, audio y otros tipos de contenido de manera autónoma y creativa (Ramesh et al., 2022).

Estas herramientas de IA Generativa abarcan desde modelos de lenguaje de gran escala (LLMs) y sus APIs, hasta frameworks especializados, plataformas de desarrollo integradas y hardware optimizado. Juntas, forman un ecosistema rico y diverso que está impulsando la innovación en múltiples sectores (Bommasani et al., 2021).

En esta sección, exploraremos las principales categorías de herramientas que pueden ser usadas en aplicaciones de IA Generativa, ofreciendo a las empresas y desarrolladores los recursos necesarios para crear aplicaciones de IA avanzadas y transformar procesos de negocio.

2.6.3.1 Modelos de Lenguaje Grande (LLMs)

Definición y Funcionamiento

Los Modelos de Lenguaje Grande (LLMs) son sistemas de inteligencia artificial diseñados para procesar y generar lenguaje natural a gran escala. Estos modelos se basan en arquitecturas de redes neuronales profundas y son entrenados con vastas cantidades de datos textuales(Minaee et al., 2024a). Los LLMs funcionan prediciendo la probabilidad de secuencias de palabras, lo que les permite generar texto coherente y realizar una variedad de tareas lingüísticas.

Evolución y Avances Recientes

Desde la introducción de GPT-3 en 2020, los LLMs han experimentado un rápido desarrollo. Modelos posteriores como GPT-4 y PaLM han demostrado capacidades aún más avanzadas en comprensión y generación de lenguaje (Sapiens & Alexander Gelbukh, 2010). Estos avances han ampliado significativamente el espectro de aplicaciones prácticas de los LLMs en diversos campos.

Aplicaciones en Negocios

Los LLMs están transformando numerosos aspectos de los negocios. Se utilizan para automatizar el servicio al cliente, generar contenido de marketing, analizar grandes volúmenes de datos textuales y asistir en la toma de decisiones (Bommasani et al., 2021). Empresas de diversos sectores están integrando LLMs en sus operaciones para mejorar la eficiencia y la innovación.

Desafíos y Consideraciones Éticas

A pesar de su potencial, los LLMs presentan desafíos significativos. Problemas como los sesgos en los datos de entrenamiento, la generación de información falsa y las preocupaciones sobre privacidad son áreas de investigación activa (Ortiz et al., 2024). Además, el uso ético y responsable de estos modelos es un tema de creciente importancia en la comunidad de IA.

Futuro de los LLMs

El futuro de los LLMs promete avances continuos en capacidades y eficiencia. Se espera que los próximos desarrollos se centren en mejorar la comprensión contextual,

aumentar la capacidad de razonamiento y reducir los requisitos computacionales (Wu et al., 2023). La integración de los LLMs con otras tecnologías de IA podría llevar a sistemas aún más sofisticados y versátiles.

2.6.3.2 APIs de IA Generativa

Específicamente nos referimos a APIs que sirven consultas a LLMs.

Definición y Propósito

Las APIs de IA Generativa para LLMs son interfaces de programación que permiten a desarrolladores y empresas acceder a las capacidades de los Modelos de Lenguaje Grande a través de servicios web. Estas APIs facilitan la integración de funcionalidades avanzadas de procesamiento y generación de lenguaje natural en aplicaciones y sistemas existentes, sin necesidad de desarrollar o alojar los modelos localmente (Bommasani et al., 2021).

Principales Proveedores y Servicios

Actualmente en el mercado existen varios proveedores que ofrecen APIs de LLMs, con diferentes características y precios. Open AI fue el pionero con su API GPT, tambien podemos mencionar a Google Cloud con su API de lenguaje natural y empresas como Anthropic y Cohere (Thoppilan et al., 2022).

Funcionalidades y Casos de Uso

Las APIs exponen las funcionalidades de los LLMs, estas incluyen generación de texto, traducciones, análisis de preguntas y sus respuestas. Se utilizan en diversas aplicaciones como chatbots, automatización de procesos, herramientas de escritura asistida, y análisis de grandes volúmenes de texto (Minaee et al., 2024b)

Integración y Desarrollo

La integración de estas APIs en aplicaciones existentes generalmente implica el uso de solicitudes HTTP y el manejo de respuestas en formato JSON. Los desarrolladores pueden personalizar los resultados ajustando parámetros como la temperatura de

generación o la longitud máxima de salida. Muchos proveedores ofrecen SDKs y bibliotecas para facilitar la integración en diferentes lenguajes de programación (Mei et al., 2024)

Desafíos y Consideraciones

El uso de APIs de LLMs presenta desafíos como la gestión de costos asociados al uso intensivo, la latencia en las respuestas para aplicaciones en tiempo real, y la necesidad de manejar apropiadamente la información sensible. Además, es importante considerar los aspectos éticos y de privacidad al procesar datos de usuarios a través de estos servicios (Finlayson et al., 2024)

Tendencias Futuras

Se esperaría que los modelos se puedan personalizar a las necesidades particulares específicas de las empresas, también una mejora en la eficiencia y reducción de costos. La integración modelos multimodales que permitan la combinación de texto con otros tipos de datos. (Patil et al., 2023)

Otras Herramientas

Los LLMs y los APIs son las herramientas que son analizadas en este estudio, pero adicionalmente hay otras que pueden ser usadas en aplicaciones de negocios. A continuación, algunas de ellas:

- **Frameworks de IA:** Estructuras de software que facilitan el desarrollo de aplicaciones de IA, como TensorFlow o PyTorch.
- **Librerías de IA:** Colecciones de funciones y herramientas predefinidas para tareas específicas de IA, como scikit-learn o Keras.
- **Plataformas de desarrollo de IA:** Entornos integrados que ofrecen herramientas y servicios para crear, entrenar y desplegar modelos de IA, como Google Cloud AI Platform o IBM Watson Studio.
- **Frameworks de aprendizaje profundo:** Herramientas especializadas para construir y entrenar redes neuronales profundas, como Caffe o Theano.

- **Herramientas de procesamiento de lenguaje natural (NLP):** Software diseñado para analizar y generar lenguaje humano, como spaCy o NLTK.
- **Motores de generación de imágenes:** Sistemas que crean imágenes a partir de texto o otros inputs, como DALL-E o Midjourney.
- **Herramientas de síntesis de voz:** Software que convierte texto en habla natural, como Amazon Polly o Google Text-to-Speech.
- **Plataformas de automatización de IA:** Sistemas que facilitan la integración y automatización de procesos basados en IA, como UiPath o Automation Anywhere.

3 CAPÍTULO III: Diseño metodológico

3.1 Tipo de Investigación

En este trabajo se aplican varios tipos de investigación dependiendo de los siguientes criterios:

Por su Objetivo Gnoseológico: Investigación Descriptiva

La investigación es descriptiva. Trata de recolectar y describir ciertas características de una lista de herramientas. En este caso de LLMs y APIs de IA Generativa.

Por su Finalidad: Investigación Básica

Es Investigación Básica porque trata de mostrar información base que podrá luego ser utilizada en otros estudios o en aplicaciones prácticas.

El objetivo es aumentar el conocimiento sobre un tema específico, no trata de resolver un problema concreto de manera práctica.

Por su Contexto: Investigación de Laboratorio

Se clasifica como investigación de laboratorio, dado que se basa en una revisión de literatura técnica, con especial atención a característica y benchmarks de rendimiento, sin realizar experimentos de campo.

Por el Control de Variables: Investigación No Experimental

Este estudio es no experimental, lo que significa que no se manipularán variables en condiciones controladas.

Por su Orientación Temporal: Investigación Transversal

Esta investigación se centra en herramientas que surgieron entre 2020 y 2024, priorizando la información más reciente sin un seguimiento a largo plazo.

3.1.1 Diseño de Investigación: Enfoque Cualitativo y Cuantitativo

El diseño del estudio adopta un enfoque mixto, combinando análisis **cualitativos y cuantitativos**:

3.2 Población y muestra

En esta **investigación descriptiva, basada en una revisión de literatura**, no se incluyen las secciones de población y muestra tradicionales, ya que no se recopilan datos primarios directamente de individuos o grupos.

El estudio se centra en el análisis de literatura técnica, como estudios previos y benchmarks.

3.3 Métodos y técnicas

Método Teórico: Analítico-sintético

Se utiliza para descomponer las herramientas de IA Generativa en sus características esenciales, evaluando cada una de ellas de forma individual y luego evaluando los resultados en un análisis completo.

Método Empírico: Estudio documental

Por otro lado, el estudio documental recopila y analiza datos secundarios provenientes de documentación técnica, especialmente benchamrks técnicos para extraer datos clave sobre el rendimiento y la efectividad de las herramientas.

3.4 Instrumentos

Se usarán una Matriz de Evaluación Técnica para la compilación de características y de indicadores de rendimiento de cada herramienta.

3.4.1 Matriz de evaluación de LLMs

A continuación, la **matriz de evaluación** que se usará para recopilar información de los LLMs:

Características	Modelo 1	Modelo 2
Creador		
Tamaño del Modelo (Billones de parámetros)		
Costo por Millón de Tokens (USD)		
Ventana de Contexto		
Licencia		

Benchmarks	Modelo 1	Modelo 2
Calidad		
Índice (Promedio normalizado)		
MMLU		
ChatBot Arena Index		
MT-Bench		
Humaneval		
Precios		
Precio de Entrada		
Precio de Salida		
Precio Combinado (3:1)		
Velocidad de Salida (Tokens/s)		
Mediana		
Percentil 5		
Percentil 25		
Percentil 75		
Percentil 95		
Latencia		
Mediana		

Percentil 5
Percentil 25
Percentil 75
Percentil 95

La información sobre rendimiento y características técnicas de los motores LLM y sus APIs ha sido obtenida mayormente de dos fuentes:

- Artificial Analysis (https://artificialanalysis.ai), un sitio web que provee herramientas para comparar LLMs llamando a sus APIs (Islam & Moushi, 2024)
- HELM, Holistic Evaluation of Language Models (https://crfm.stanford.edu/helm/), es una referencia para la comunidad, que se actualiza continuamente con nuevos escenarios, métricas y modelos (Bommasani et al., 2023). Es publicado por el Centro de Investigación sobre Modelos Fundamentales de la Universidad de Stanford (CRFM).

A continuación, se explican algunos de los datos usados:

- Calidad: El índice mide el rendimiento relativo promedio, normalizado según Chatbot Arena, MMLU y MT-Bench.
- Ventana de contexto: Define el número máximo de tokens combinados de entrada y salida, con un límite inferior variable para los tokens de salida según el modelo.
- Velocidad de salida: Representa la cantidad de tokens generados por segundo mientras el modelo está en funcionamiento, después de recibir el primer fragmento desde la API, clasificado en percentiles: P5, P25, P75 y P95.
- Latencia: Tiempo en segundos para recibir el primer token después de enviar la solicitud a la API.
- Precio: Costo por token, en USD por millón de tokens, combinando los precios de tokens de entrada y salida en una proporción de 3:1.

- Precio de salida: Costo en USD por millón de tokens generados por el modelo desde la API.
- Precio de entrada: Costo en USD por millón de tokens incluidos en la solicitud enviada a la API.
- Período de tiempo: Las métricas se basan en los últimos 14 días y se registran en tiempo real, con mediciones realizadas ocho veces al día para solicitudes individuales y dos veces al día para solicitudes paralelas.

3.4.2 Matriz de evaluación de APIs de IA

A continuación, la **matriz de evaluación** que se usará para recopilar información de los proveedores de APIs de IA:

Características	API 1	API 2
Api Provider		
Modelo		
Contexto		
Context Window		
License		
Compatible con Openai		
Api ID		
Calidad		
Index (promedio normalizado)		
Chatbot Arena		
MMLU		
MT Bench		
Humaneval		
Precio		
Input Price		
Output Price		
Combinado (3:1)		
Velocidad de Salida (Tokens/s)		
Mediana		
P5		

	P25
	P75
	P95
Latencia (Tokens/s)	
	Mediana
	P5
	P25
	P75
	P95

3.5 Alcance y Limitaciones del Enfoque Metodológico Elegido

En este estudio vamos a recopilar y analizar información extraída de benchmarks (rankings comparativos) de LLMs y APIs.

Se busca identificar las mejores alternativas para desarrollar aplicaciones empresariales que usen IA.

Los gráficos comparativos junto con los análisis cruzados permiten visualizar las características en donde destacan los diferentes productos evaluados, o sus puntos más débiles. También facilita la identificación de tendencias en el rendimiento y áreas de mejora.

Sin embargo, esta metodología tiene sus limitaciones.

- Los benchmarks tienen una gran utilidad al evaluar muchas herramientas, pero no siempre reflejan la complejidad de las aplicaciones reales.
- El avance acelerado en las herramientas que se ha visto en los últimos 2 años puede hacer que algunos resultados presentados queden obsoletos rápidamente.
- También estamos limitados por la disponibilidad de datos de benchmarks para ciertos modelos o APIs propietarias.

4 CAPÍTULO IV: Análisis e interpretación de resultados

De los tipos de IA existentes se eligió para este estudio la IA Generativa. De este tipo se seleccionó dos herramientas que consideramos vitales para la operación de aplicaciones empresariales que usan IA Generativa, los cuales que son:

- LLMs: Modelos de Lenguaje Grande
- APIs de LLMs: Servicios que proveen acceso a LLMs.

4.1 Selección de productos a evaluar

4.1.1 Proceso de selección de la muestra

Hay una cantidad considerable de productos que pertenecen a los tipos de herramientas a estudiar. Estudiarlos todos no es práctico por lo que se ha establecido un criterio para seleccionar 10 de cada tipo.

La selección de los productos a evaluar se realizó en base al **índice de calidad,** que es una métrica publicada por Artificial Analysis.

Las herramientas serán escogidas basándose en los siguientes criterios:

- LLMs: se seleccionarán los 10 mejores puntuados en el Índice de calidad de Artificial Analysis.
- APIs de IA: Se considerarán el Índice de calidad de Artificial Analysis ya mencionado. Para ello se considerará el mejor índice de calidad de un LLM para cada proveedor de APIs.LMs seleccionados

Los 10 mejores LLMs puntuados en el Índice de calidad de Artificial Analysis son:

Modelo	Creador	Licencia
GPT-4o	Openai	Propietario
Claude 3.5 Sonnet	Anthropic	Propietario
Gemini 1.5 PRO	Google	Propietario
GPT-4 Turbo	Openai	Propietario
Claude 3 Opus	Anthropic	Propietario

Reka Core	Reka	Propietario
GPT-4	Openai	Propietario
YI-Large	01.Ai	Propietario
Gemini 1.5 Flash	Google	Propietario
Llama 3 (70B)	Meta	Open Source

4.1.2 APIs de LLMs seleccionados

Los seleccionados son:

Proveedor	**Modelo ***
OpenAI	GPT-4o
ANTHROPIC	Claude 3 Opus
Gemini 1.5 Pro	Gemini 1.5 Pro
Microsoft Azure	GPT-4 Turbo
AWS	Claude 3 Opus
Reka	Reka Core
01.AI	Yi-Large
Fireworks AI	Yi-Large
Replicate	Llama 3, 70(B)
OctoAI	Llama 3, 70(B)

4.2 Presentación de los Datos

4.2.1 Métricas y Frameworks de Evaluación

Existen varios frameworks y métricas para evaluar los modelos de inteligencia artificial. A continuación, se describen algunos:

Context Windows: Se evalúa la capacidad de un modelo para comprender y generar texto a partir de una cantidad fija de texto previo (Peng et al., 2023).

MMLU (Massive Multitask Language Understanding): Mide la capacidad de un modelo de IA para realizar múltiples tareas de comprensión de lenguaje natural (Wang et al., 2024).

Chatbot Arena: Plataforma de evaluación que compara el desempeño de diferentes modelos de IA conversacionales. Se centra en la calidad y coherencia de las respuestas generadas por los chatbots en distintos contextos, ayudando a identificar las mejores prácticas y áreas de mejora. (Chiang et al., 2024)

HumanEval: Evaluación que utiliza revisores humanos para calificar la calidad de las respuestas generadas por modelos de IA (Gao et al., 2023).

Output Tokens: Métrica que mide la cantidad de tokens generados por un modelo de IA en un percentil. Los percentiles considerados son P5, P25, P75 y P90.

Los percentiles son medidas estadísticas que dividen un conjunto de datos en 100 partes iguales. Por ejemplo P5 es el valor por debajo del cual se encuentra el 5% de los datos, el P25 es el valor por debajo del cual se encuentra el 25% de los datos, etc.

Latency P5: Métrica que mide el tiempo de respuesta de un modelo de IA en un percentil. Los percentiles considerados son P5, P25, P75 y P90.

Precio: En la inteligencia artificial, principalmente en los modelos de procesamiento de lenguaje natural, el costo de utilizarlo se basa en tokens. Un token es una secuencia de caracteres que se trata como una unidad en el procesamiento de texto.

Generalmente el costo se define por millón de tokens y depende si son de entrada o salida, esta diferenciación se debe a la variación de los recursos y el procesamiento requerido. Existen varios factores que influyen en el costo en los diferentes modelos:

- **Complejidad del Modelo:** Esto afecta a la cantidad de recursos computacionales necesarios.
- **Velocidad de Procesamiento:** La rapidez con la generan respuestas. Con tarifas más altas para tiempos de respuesta más rápidos.
- **Escalabilidad y Volumen de Uso:** Los usuarios que procesan grandes volúmenes a menudo pueden obtener planes con tarifas más bajas por token.
- **Funciones Adicionales:** Algunos servicios adicionales pueden incluir: soporte multi-idiomas, multimodalidad (capacidad de procesar información diferente a texto como audio o video) entre otros.
- **Knowledge Cutoff :** Fecha de corte de los datos con que se entrenó un LLM.

4.2.2 Tabulación de Datos de LLMs.

Tabla 3: Métricas de LLMs

				Calidad					Precios		
Modelo	**Creador**	**Licencia**	**Context window**	**Atificial Analysis Index**	**Chatbot Arena**	**MMLU**	**MT Bench**	**HUMAN EVAL**	**Blended USD/1M tokens**	**Input USD/1M tokens**	**Output USD/1M tokens**
GPT-4o	Openai	Propietario	128k	100	1287	0.887		90.2	$7.50	$5.00	$15.00
Claude 3.5 Sonnet	Anthropic	Propietario	200k	98	1272	0.887		92	$6.00	$3.00	$15.00
Gemini 1.5 PRO	Google	Propietario	1m	95	1265	0.859		84.1	$5.25	$10.00	$10.50
GPT-4 Turbo	Openai	Propietario	128k	94	1256	0.864	9.32	85.4	$15.00	$15.00	$30.00
Claude 3 Opus	Anthropic	Propietario	200k	93	1249	0.868			$30.00	$3.00	$75.00
Reka Core	Reka	Propietario	128k	90		0.832		76.8	$6.00	$30.00	$15.00
GPT-4	Openai	Propietario	8k	84	1186	0.864		88.4	$37.50	$3.00	$60.00
YI-Large	01.Ai	Propietario	32k	84	1217				$3.00	$0.35	$3.00
Gemini 1.5 Flash	Google	Propietario	1m	84	1231	0.789		74.3	$0.53	$0.90	$1.05
Llama 3 (70B)	Meta	Open Source	8k	83	1207	0.82		81.7	$0.90	$0.90	$0.90

(Continúa…)

(Continúa de la tabla anterior…)

Modelo	Output tokens					Latency							
	Median token	P5	P25	P75	P95	Median chunk	P5	P25	P75	P95	Python Coding	Grade School Math	Math Problems
GPT-4o	87.1	57.4	72.6	106.7	129.7	0.47	0.32	0.41	0.68	0.95	0.90	-	0.77
Claude 3.5 Sonnet	78.7	62	71.6	86.3	91.8	0.97	0.79	0.86	0.16	1.69	0.92	0.96	0.71
Gemini 1.5 PRO	61.1	54.5	58.8	64	66.3	1.02	0.79	0.9	1.33	1.45			
GPT-4 Turbo	30	15.9	24.6	37.9	50.6	0.59	0.51	0.55	0.69	1.03			
Claude 3 Opus	24.8	19.3	22.4	27.8	31	1.92	1.57	1.72	2.12	2.93	0.85	0.95	0.60
Reka Core	14.4	7.9	11.8	14.9	15.6	1.16	1.06	1.11	1.22	1.64			
GPT-4	25.3	14	20.8	31.5	38.2	0.67	0.5	0.55	0.87	1.33	0.67	0.92	0.53
YI-Large	66.1	24.2	58.2	72.9	89.1	0.85	0.32	0.37	1.86	2.37			
Gemini 1.5 Flash	165	140.2	152.1	175.5	188.3	1.06	0.74	0.87	1.34	1.54			
Llama 3 (70B)	57.7	14.8	40.9	108.4	347.9	0.46	0.23	0.3	1.2	6.3	0.62	0.80	0.30

(Continúa de la tabla anterior…)

Modelo	Narrative QA	Natural Questions (open)	Natural Questions (closed)
GPT-4o	0.804	0.803	0.501
Claude 3.5 Sonnet	0.746	0.749	0.502
Gemini 1.5 PRO	0.783	0.748	0.378
GPT-4 Turbo	0.761	0.795	0.482
Claude 3 Opus	0.351	0.264	0.441
Reka Core			
GPT-4	0.768	0.790	0.457
YI-Large	0.373	0.586	0.428
Gemini 1.5 Flash	0.783	0.723	0.332
Llama 3 (70B)	0.798	0.743	0.475

4.2.3 Tabulación de Datos de APIs.

Tabla 4: Métricas de APIs

		Contexto			Calidad				
Proveedor	**Modelo ***	**Context window**	**Licencia**	**API ID**	**Artificial Analysis Index**	**Chatbot Arena**	**MMLU**	**MT Bench**	**Human Eval**
OpenAI	**GPT-4º**	128k	Propietario	gpt-4º	100	1287	0.887		90.2
ANTHROPIC	**Claude 3 Opus**	200k	Propietario	claude-3-5-sonnet-20240620	98	1272	0.887		92.0
Gemini 1.5 Pro	**Gemini 1.5 Pro**	1m	Propietario	Gemini-1.5-flash-lasted	95	1265	0.859		84.1
Microsoft Azure	**GPT-4 Turbo**	128k	Propietario	gpt-4/1106-Preview	94	1256	0.864	9.32	85.4
AWS	**Claude 3 Opus**	200k	Propietario	anthropic.claude-3-opus-20240229-v1:0	93	1249	0.868		
Reka	**Reka Core**	128k	Propietario	Reka-core	90		0.832		76.8
01.AI	**Yi-Large**	32k	Propietario	yi-large	84	1217			
Fireworks AI	**Yi-Large**	32k	Propietario	accounts/yi-01-ai/models/yi-large	84	1217			
Replicate	**Llama 3 70(MB)**	8k	Open	meta/meta-llama-3-70b-instruct	83	1207	0.82		81.7
OctoAI	**Llama 3 70(MB)**	8k	Open	meta-llama-3-70b-instruct	83	1207	0.82		81.7

* Se seleccionó el mejor modelo de cada API.

(Continúa…)

(Continúa de la tabla anterior…)

		Precio			Output tokens					Latency				
Proveedor	Modelo	Precio Blended USD/1M tokens	Precio Input USD/1M tokens	Precio Output USD/1M tokens	Output Median Token	P5	P25	P75	P95	Median First Chunk	P5	P25	P75	P95
Openai	GPT-4o	$7.50	$5.00	$15.00	87.1	57.4	106.7	106.7	129.4	0.45	0.32	0.39	0.65	0.88
Anthropic	Claude 3 Opus	$6.00	$3.00	$15.00	78.8	61.6	72.3	84.2	91.2	1.08	0.84	0.95	1.21	1.75
Gemini 1.5 Pro	Gemini 1.5 Pro	$5.25	$3.50	$10.50	61.1	54.5	64.0	64.0	66.3	0.29	0.79	0.9	1.33	1.54
Gpt-4 Turbo	Microsoft Azure	$15.00	$10.00	$30.00	33.5	13.8	25.1	43.1	0.5	0.57	0.51	0.54	0.62	0.89
Aws	Claude 3 Opus	$30.00	$15.00	$75.00	24.0	19.3	22.1	25.9	28.4	1.83	1.52	1.65	1.98	2.34
Reka	Reka Core	$6.00	$3.00	$15.00	14.4	9.0	12.3	14.9	15.4	1.14	1.03	1.09	1.2	1.43
01.Ai	Yi-Large	$3.00	$3.00	$3.00	66.4	24.4	26.0	74.0	88.4	1.97	1.45	1.59	2.25	2.49
Fireworks AI	Yi-Large	$3.00	$3.00	$3.00	73.4	64.2	70.0	79.2	92.4	0.34	0.26	0.31	0.38	0.42
Replicate	Llama 3 70(MB)	$1.18	$0.65	$2.75	79.6	63.4	74.5	82.5	86.1	1.56	1.19	1.26	6.32	6.89
Octoai	Llama 3 70(MB)	$0.90	$0.90	$0.90	62.8	56.5	59.2	65.3	70.3	0.29	0.19	0.27	0.3	0.45

4.3 Análisis de las Métricas

4.3.1 Análisis de Modelos LLMs

En este análisis, evaluaremos varias métricas de modelos de lenguaje grande (LLMs). Especialmente en la calidad, costos, rendimiento y latencia de cada modelo.

4.3.1.1 Análisis de Calidad

Tabla 5: LLMs, Análisis de Calidad

Métrica	Análisis
Index	Refleja la calidad general de los modelos. OpenAI lidera con 100, seguido de Anthropic con 98 y Google con 95.
Chatbot Arena	OpenAI también encabeza esta métrica con un puntaje de 1287, indicando su alta capacidad para mantener conversaciones naturales y coherentes. Anthropic (1272) y Google (1265) le siguen de cerca.
MMLU	Esta métrica mide la capacidad de los modelos para manejar tareas de lenguaje natural. Google (0.859) y Anthropic (0.887) muestran un buen desempeño, pero OpenAI (0.887) sigue a la delantera.
MT Bench	No hay suficientes datos de esta métrica para hacer un análisis.
HUMAN EVAL	OpenAI lidera con 90.2, luego Anthropic con 92 y Google con 84.1.

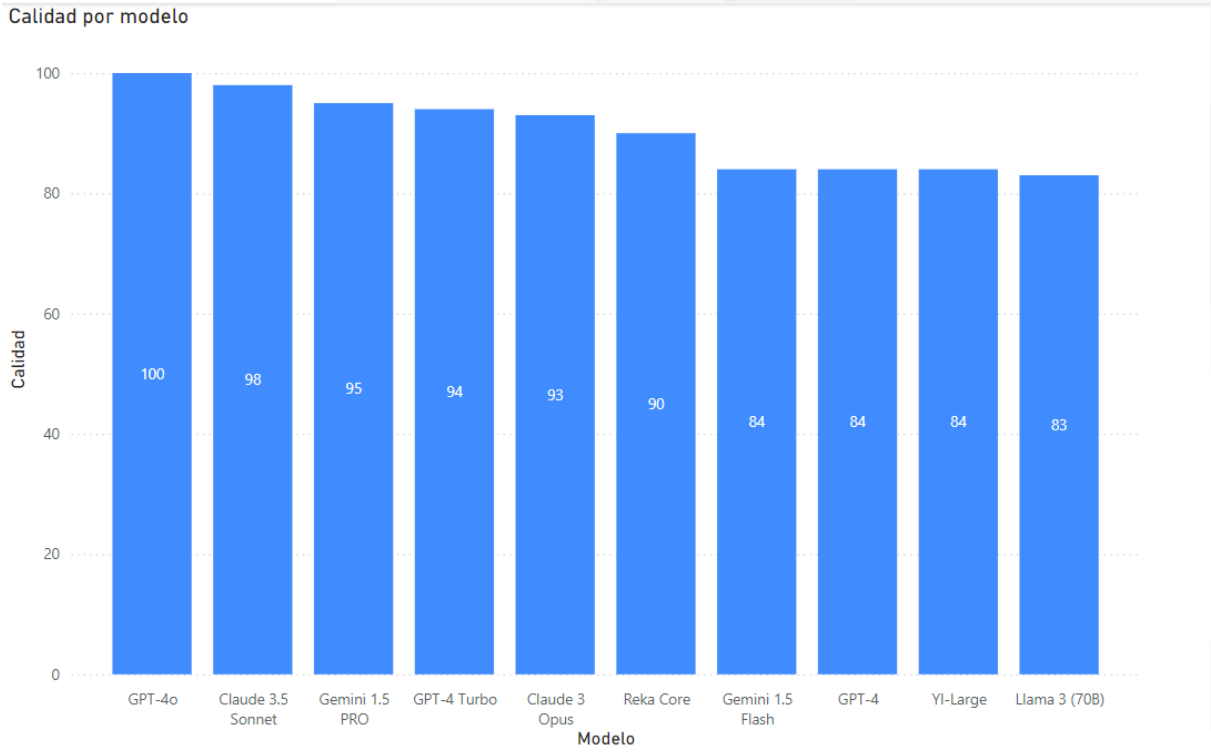


Ilustración 1: LLMs, Análisis de Calidad

4.3.1.2 Análisis de Costos

Tabla 6: LLMs, Análisis de Costos

Métrica	Análisis
USD/1M tokens (Input):	Google ofrece el menor costo ($0.53), seguido de Meta ($0.90) y Anthropic ($6.00). OpenAI es bastante caro en esta métrica ($15.00).
USD/1M tokens (Output):	Google lidera con ($0.90), OpenAI ($15.00) y Reka ($30.00) son bastante más altos.
USD/1M tokens (Total):	El precio "blended" combina los costos de entrada y salida en una proporción 3:1. Aquí, Google nuevamente presenta el costo más bajo ($1.05), seguido de Meta ($0.90) y Anthropic ($15.00).

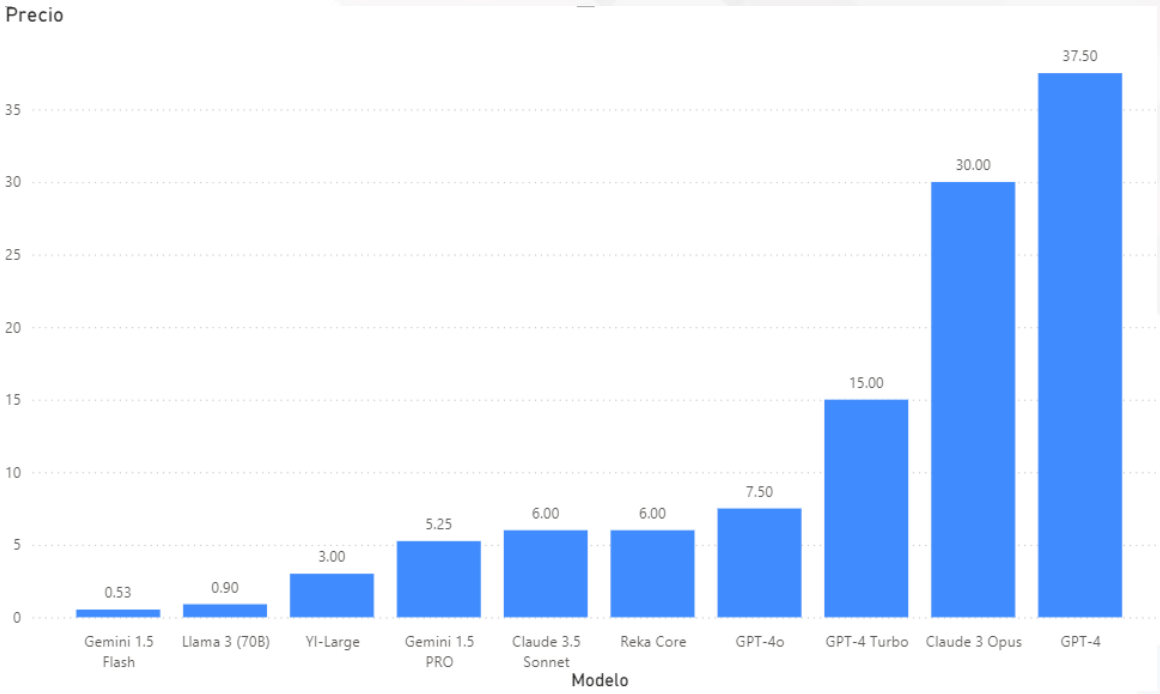


Ilustración 2: LLMs, Análisis de Costos

4.3.1.3 Rendimiento y Latencia

Tabla 7: LLMs, Análisis de Rendimiento y Latencia

Métrica	Análisis
Output Median token:	Reka Core es el más rápido (165 tokens/s), seguido de YI-Large (66.1 tokens/s) y GPT-4o (87.1 tokens/s).
P5, P25, P75, P95:	Estas métricas indican la distribución de la velocidad de generación. GPT-4o mantiene una buena consistencia, con percentiles relativamente altos en comparación con otros modelos como Claude 3 Opus, que muestra más variabilidad.
Latency Median chunk:	Llama 3 (70B) tiene la menor latencia (0.46s), seguido de GPT-4o (0.47s) y YI-Large (0.85s).
P5, P25, P75, P95:	Llama 3 (70B) también muestra buenos valores de latencia

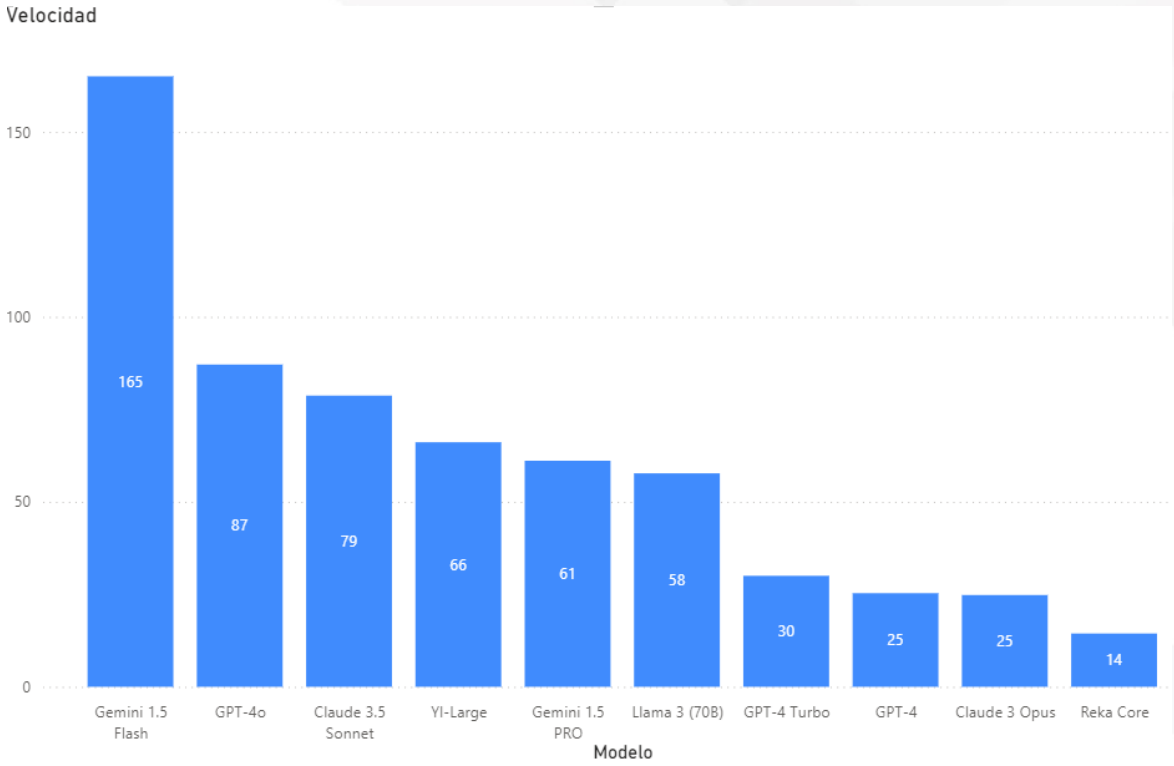


Ilustración 3: LLMs, Análisis de Velocidad

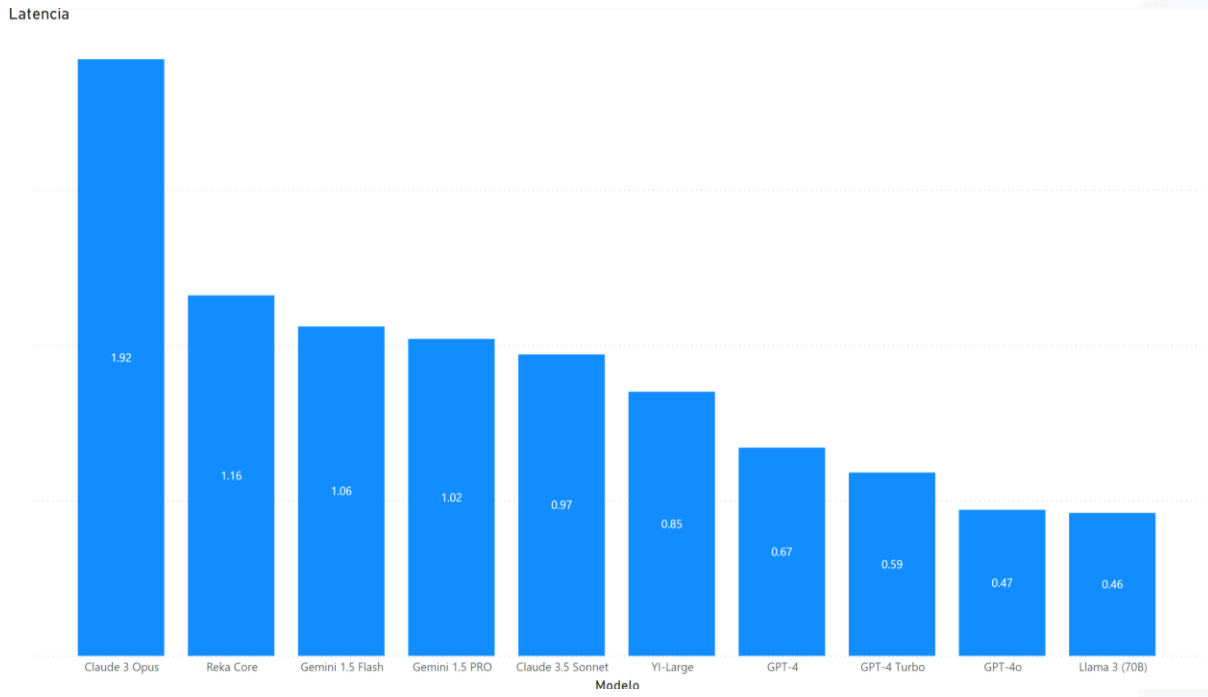


Ilustración 4: LLMs, Análisis Latencia

4.3.1.4 Relaciones entre Variables

LLMs: Calidad vs. Precio

Se observa que hay una relación notable entre los costos y las métricas de calidad. OpenAI. Aunque tiene costos más altos, también lidera en la mayoría de las métricas de calidad (Index, Chatbot Arena, MMLU, HUMAN EVAL).

Google presenta una buena relación costo-calidad. Esto sugiere que hay casos específicos que ofrecen una alta calidad a un costo más accesible.

OpenAI con el modelo GPT-4o es el valor más alto en precio/calidad. Una buena opción es el modelo de Google Gemini 1.5 PRO en donde la relación precio/calidad es baja y su nivel de calidad se mantiene entre los más altos.

Hay otras opciones como GPT4 turbo de OpenAI que, si bien la relación precio/calidad es baja, la calidad está en las últimas posiciones del ranking. Se puede también inferir que los nuevos modelos mejoran esta relación precio/calidad, como por ejemplo Chatgpt-4o y Claude 3.5 sonnet en comparación a Chatgpt turbo y Claude 3 opus respectivamente.

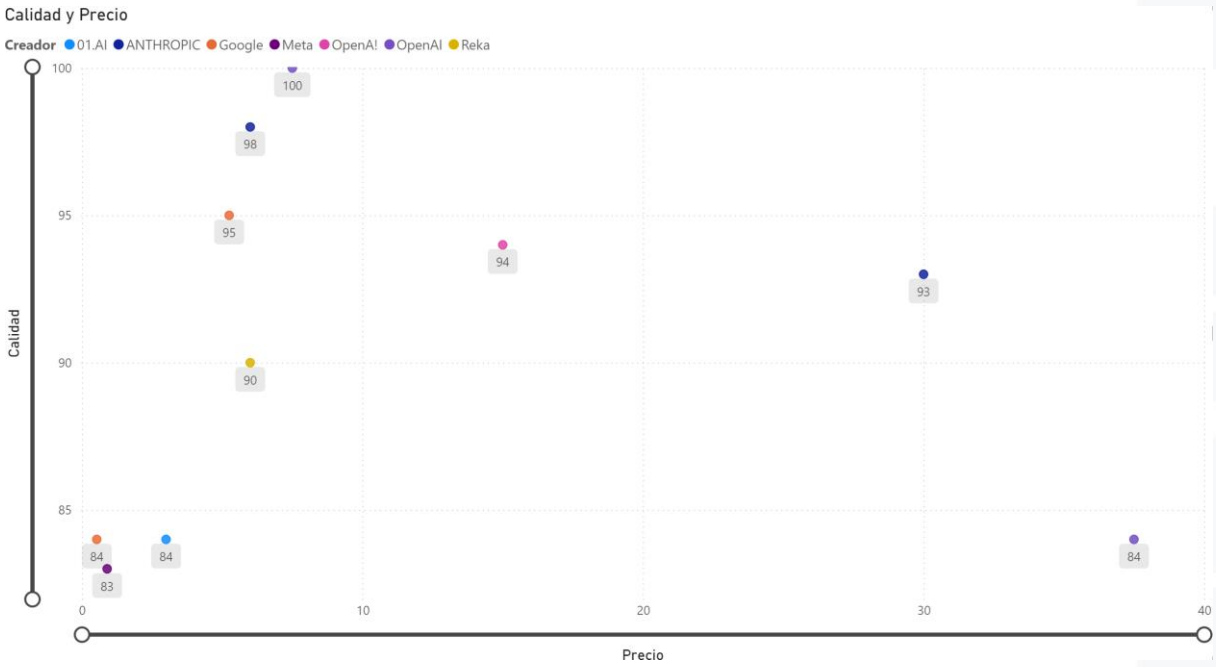


Ilustración 5: LLMs, Calidad vs. Precio

LLMs, Velocidad de Generación vs. Calidad

Reka Core que tiene alta velocidad de generación, no está entre las mejores métricas de calidad. Esto indica que una mayor velocidad de generación no siempre implica una mejor calidad.

OpenAI mantiene un equilibrio entre velocidad razonable y alta calidad.

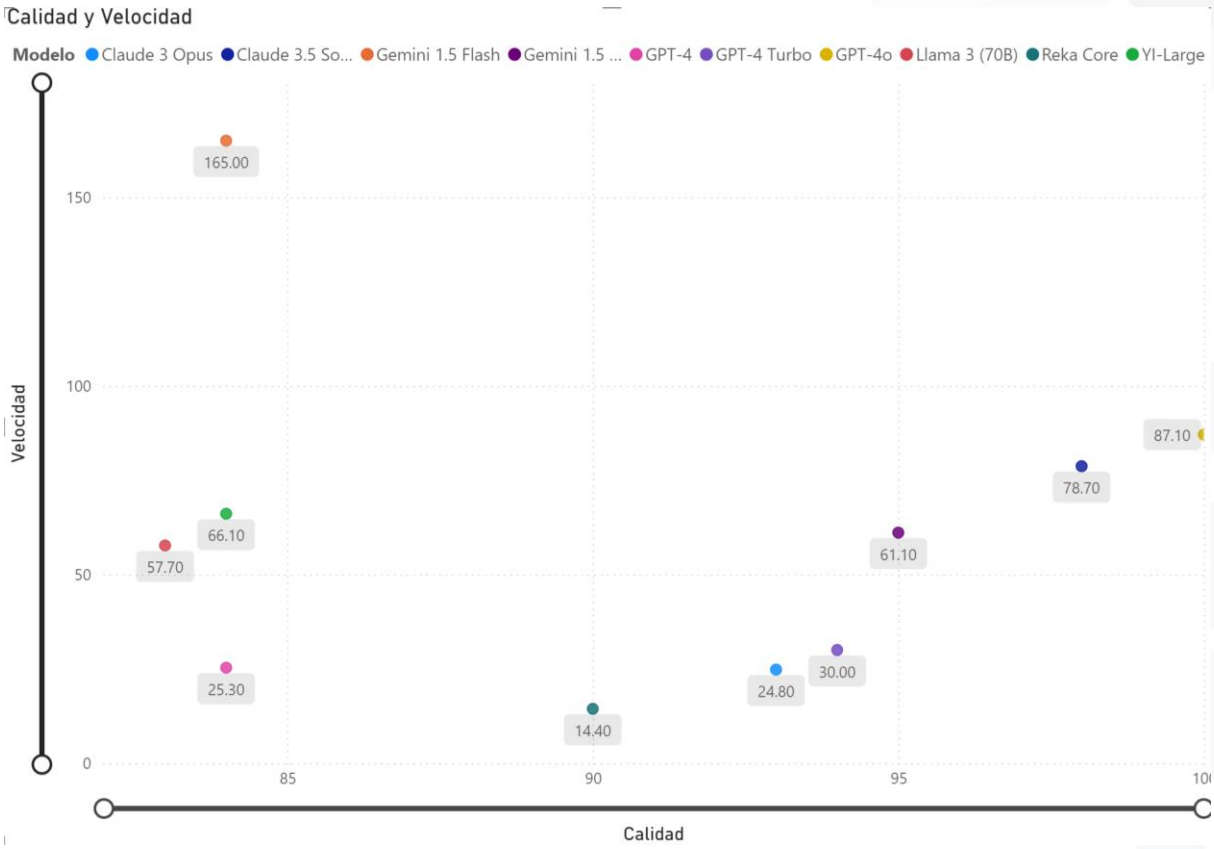


Ilustración 6: LLMs, Velocidad de Generación vs. Calidad

LLMs, Latencia vs. Calidad

GPT-4o y GPT turbo presentan las menores relaciones latencia/calidad. Si bien Llama 3 tiene la latencia más baja, su índice de calidad también lo es.

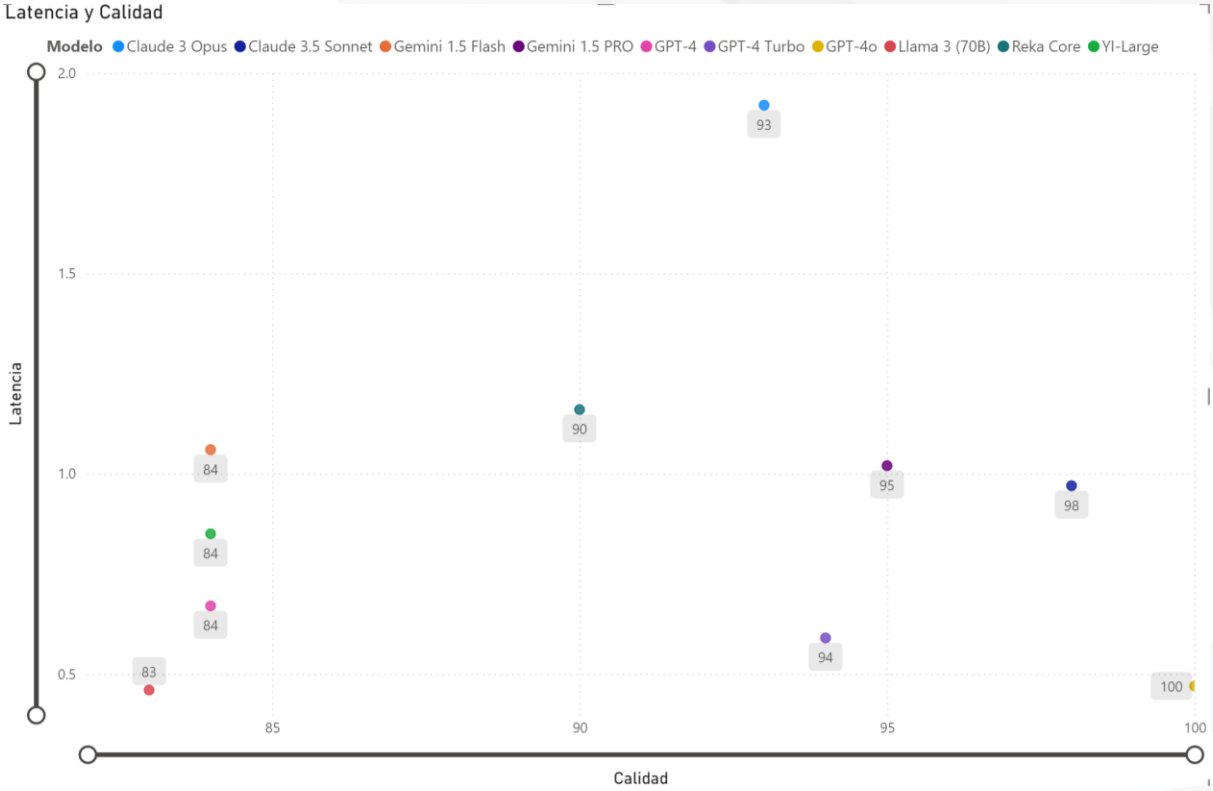


Ilustración 7: LLMs, Latencia vs. Calidad

LLMs, Precios vs Velocidad

Gemini 1.5 Flash destaca como la opción más rentable en términos de velocidad y costo. GPT-4o y Claude 3.5 Sonnet son buenas opciones si se busca un equilibrio entre alta velocidad y costos moderados. YI-Large y Llama 3 (70B) son opciones muy económicas con velocidades moderadas.

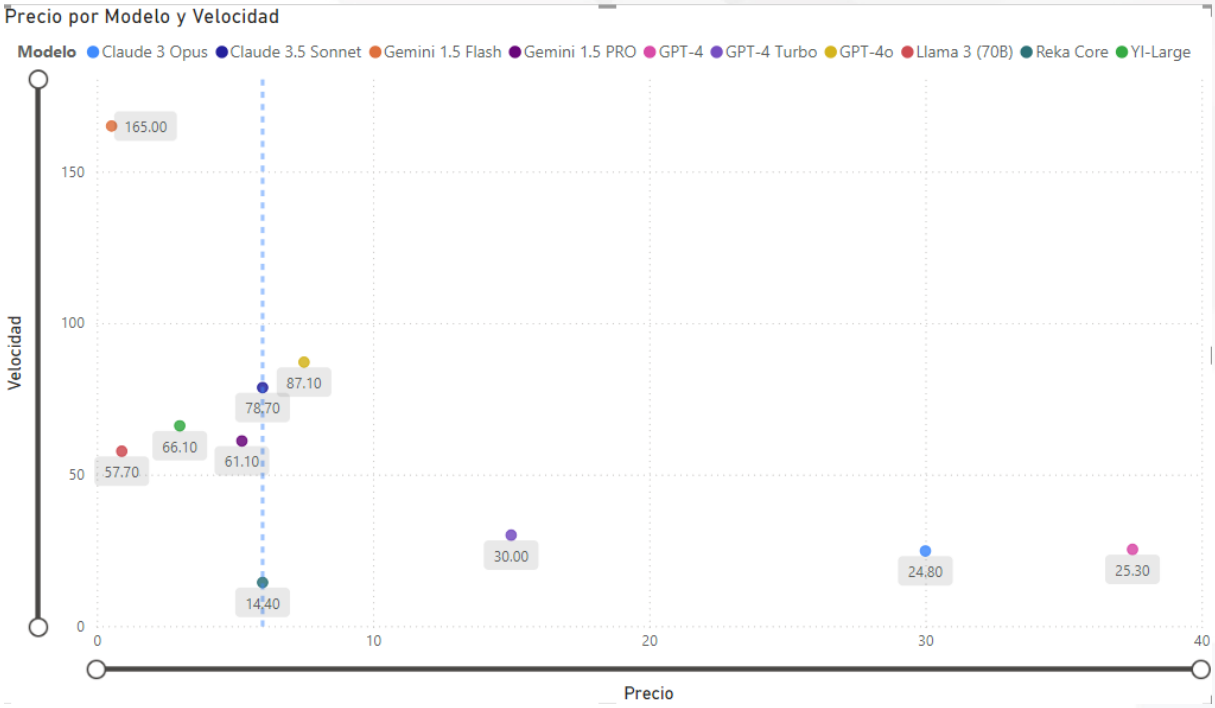


Ilustración 8: LLMs, Precios vs Velocidad

4.3.1.5 Patrones y Tendencias

Costos y Calidad:

Se nota que modelos más caros tienden a tener mejores métricas de calidad. Sin embargo, Google logra ofrecer buena calidad con costos relativamente bajos.

Consistencia en la Velocidad y Latencia:

GPT-4o y Llama 3 (70B) tienen consistencia en latencia y velocidad. Esto es útil en aplicaciones que requieren un rendimiento previsible.

4.3.2 Análisis de los APIs de LLMs

4.3.2.1 Análisis de Calidad

Tabla 8: APIs, Análisis de Calidad

Métrica	Análisis
Index	OpenAI lidera con 100. Le siguen Anthropic (98) y Gemini 1.5 Pro (95).
Chatbot Arena	OpenAI encabeza con 1287 puntos. Anthropic y Gemini 1.5 Pro también tienen puntuaciones altas (1272 y 1265).
MMLU	OpenAI y Anthropic tienen las mismas puntuaciones (0.887). Gemini 1.5 Pro les sigue con 0.859.
MT Bench	No hay suficientes datos
HUMAN EVAL	OpenAI, Anthropic y Gemini 1.5 Pro tienen altas puntuaciones (90.2, 92.0 y 84.1).

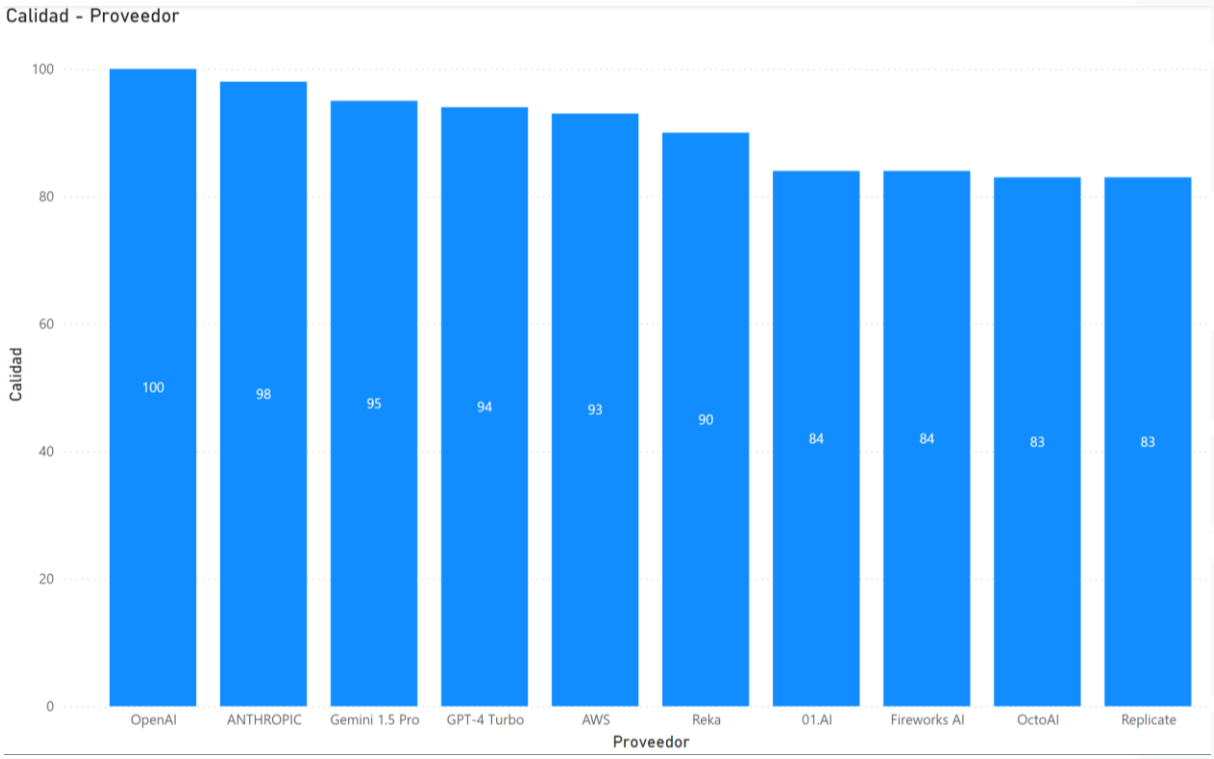


Ilustración 9: APIs, Análisis de Calidad

4.3.2.2 Análisis de Costos

Tabla 9: APIs, Análisis de Costos

Métrica	Análisis
USD/1M tokens (Input):	El costo de entrada más bajo es de Replicate ($0.65), y OctoAI ($0.90). OpenAI y AWS tienen costos de entrada bastante más altos ($5.00 y $15.00).
USD/1M tokens (Output):	Replicate y OctoAI nuevamente ofrecen costos de salida bajos ($2.75 y $0.90). Los costos de OpenAI y AWS son bastante altos ($15.00 y $75.00).
USD/1M tokens (Total):	OctoAI tiene el costo más bajo ($0.90), AWS tiene el costo total más alto ($75.00).

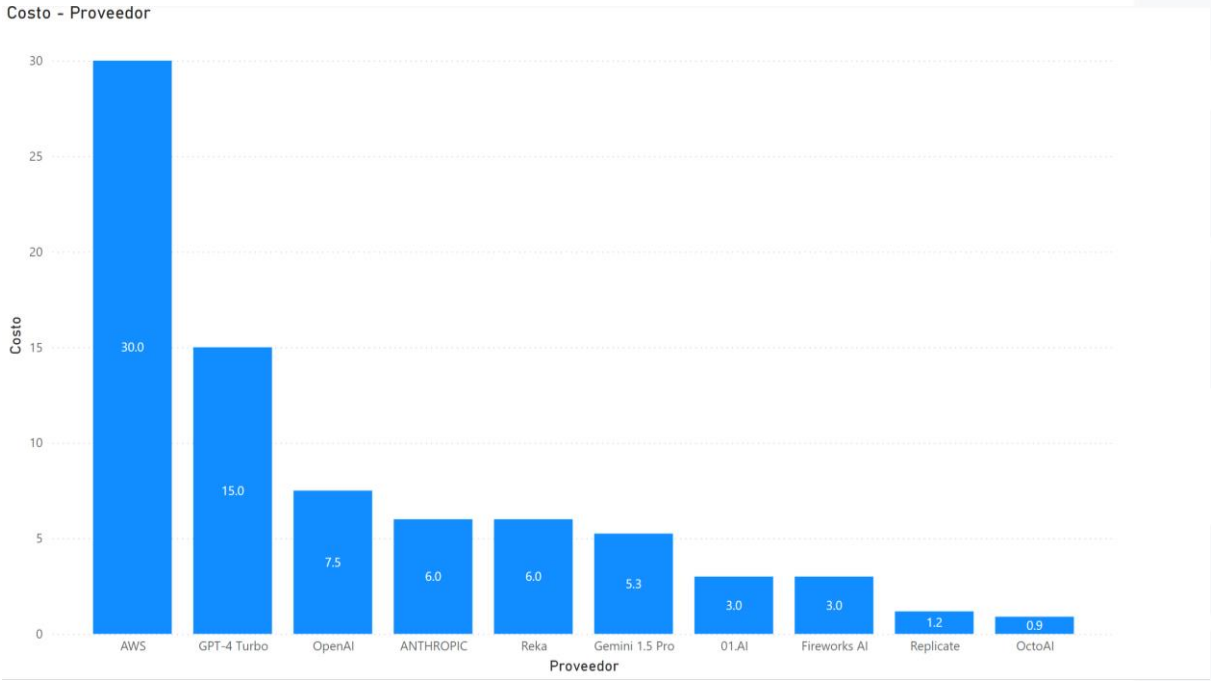


Ilustración 10: APIs, Análisis de Costos

4.3.2.3 Análisis de Rendimiento y Latencia

Tabla 10: APIs, Análisis de Rendimiento y Latencia

Métrica	Análisis
Output Median token:	OpenAI lidera con 87.1 tokens/s, seguido de Anthropic (78.8) y Fireworks AI (73.4).

P5, P25, P75, P95:	Las métricas percentiles muestran la variabilidad en la velocidad de generación. OpenAI y Fireworks AI presentan consistencia en sus velocidades.
Latency Median chunk:	Fireworks AI tiene la menor latencia (0.34s), luego OpenAI (0.45s). AWS tiene la latencia más alta (1.83s).
P5, P25, P75, P95:	Fireworks AI y OpenAI tienen latencias consistentes y bajas, AWS tiene mayor variabilidad y mayor latencia.

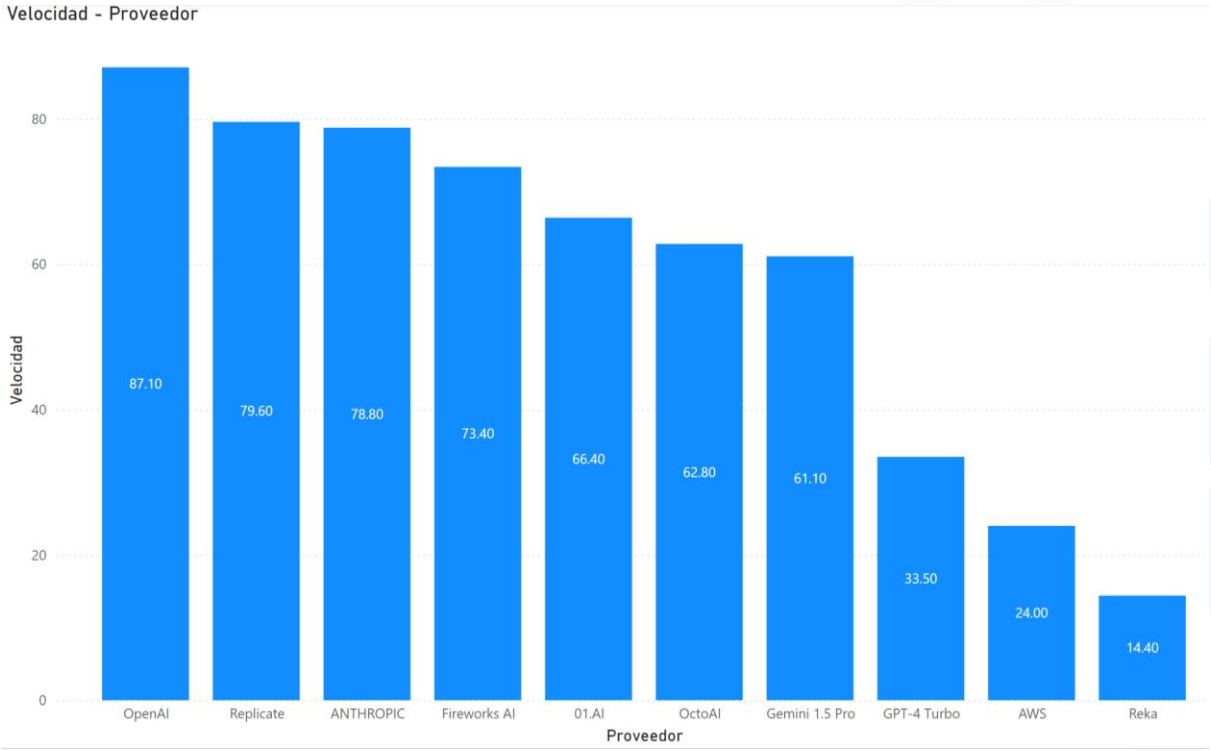


Ilustración 11: APIs, Análisis de Velocidad

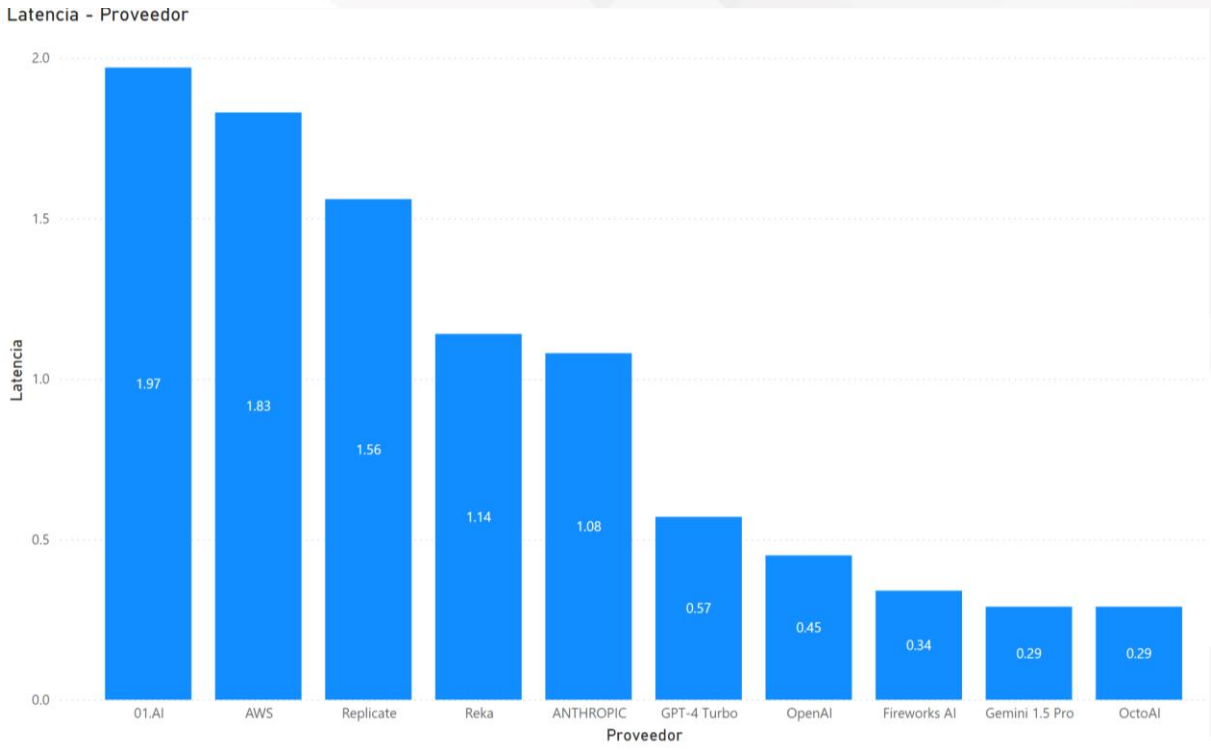


Ilustración 12: APIs, Análisis de Latencia

4.3.2.4 Relación entre Variables

Calidad vs. Costo

Comparando las métricas de calidad (Index, Chatbot Arena, MMLU, MT Bench, HUMAN EVAL) con los costos (USD/1M tokens input, output y blended), se nota que OpenAI lidera en las métricas de calidad aunque tiene costos más altos ($7.50 blended).

OctoAI, tiene el costo más bajo ($0.90 blended) y ofrece una calidad decente, pero bastante menor que OpenAI. Esto sugiere una fuerte correlación entre costo y calidad.

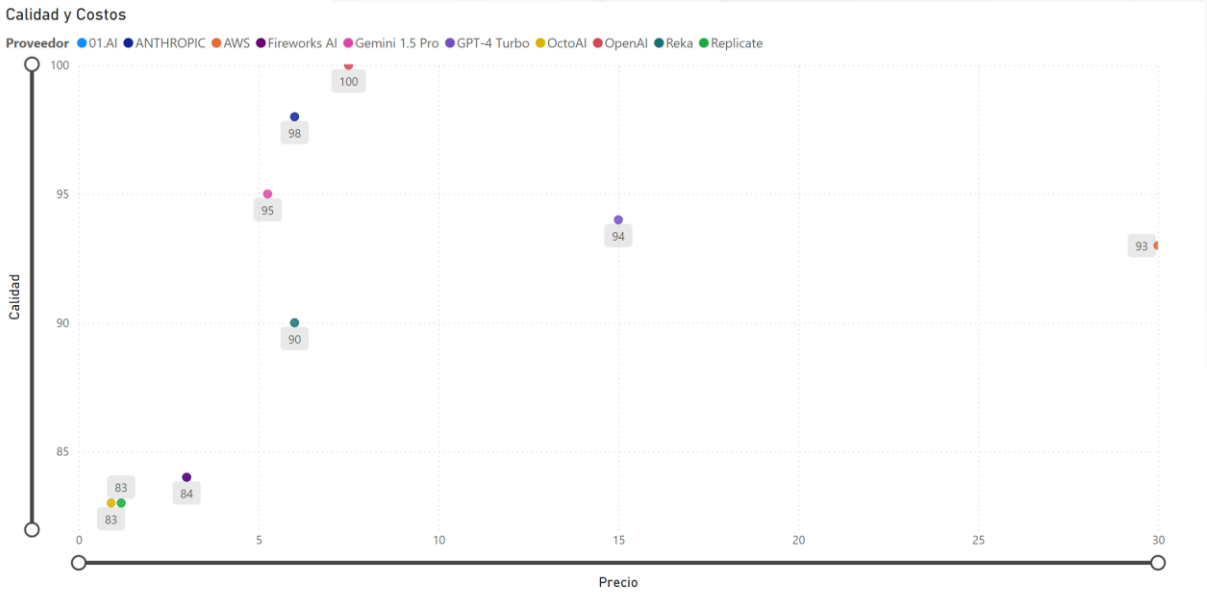


Ilustración 13: APIs, Calidad vs Costo

Rendimiento vs. Costo

La velocidad de generación de tokens tiene variaciones importantes según el costo. OpenAI, con alta velocidad de 87.1 tokens/s, justifica su alto costo mientras que Fireworks AI y Replicate, son más económicos, pero tienen buenas velocidades (73.4 y 79.6 tokens/s).

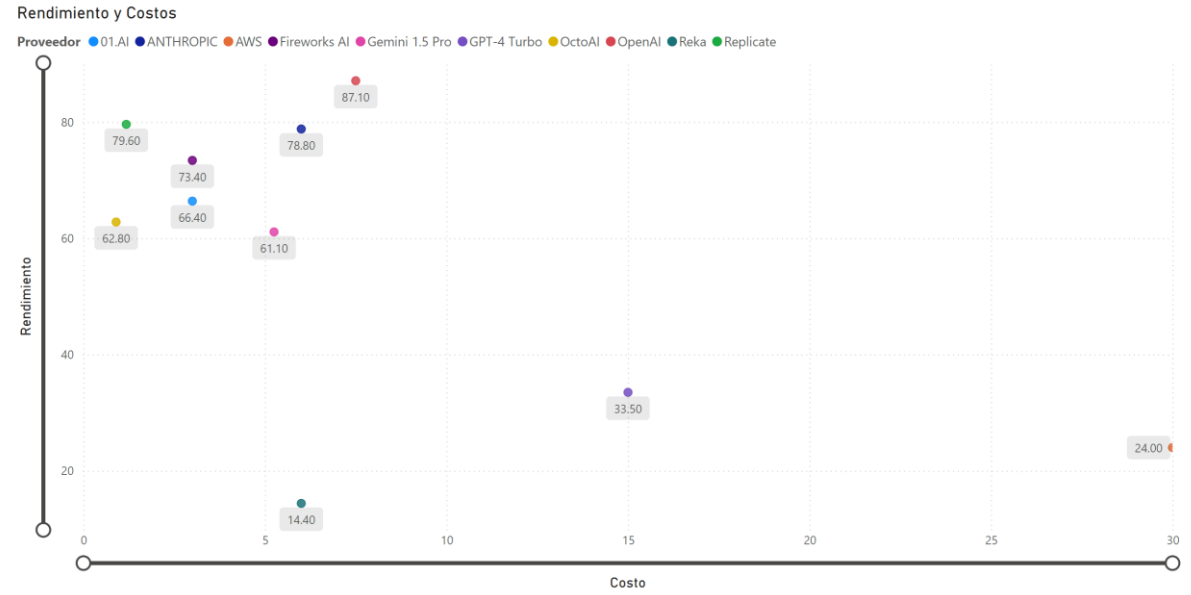


Ilustración 14: APIs: Rendimiento vs Costo

Latencia vs. Costo

Fireworks AI y OpenAI tienen las latencias más bajas (0.34s y 0.45s, respectivamente) y costos relativamente altos, lo que indica una correlación negativa: a menor latencia, mayor costo. AWS presenta la mayor latencia (1.83s) y también uno de los costos más altos, lo que rompe esta tendencia en algunos casos.

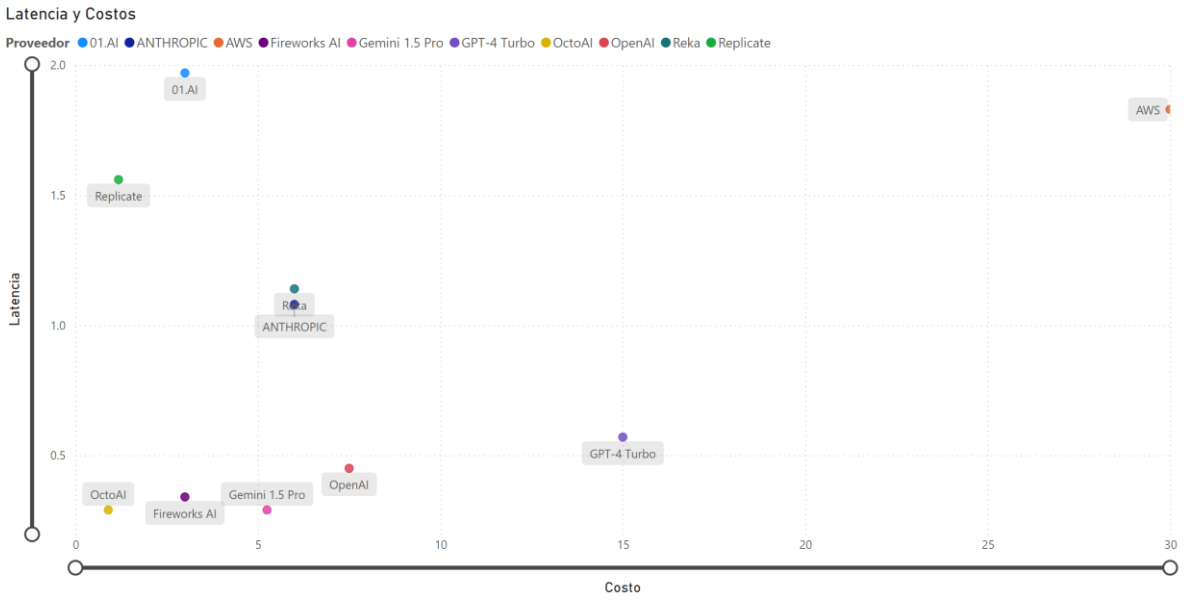


Ilustración 15: APIs, Latencia vs. Costo

4.3.2.5 Patrones y Tendencias

Costo y Calidad

Los modelos de OpenAI y Anthropic muestran que altos costos están consistentemente asociados con alta calidad. Modelos más económicos como OctoAI y Replicate ofrecen un rendimiento razonable para su costo, aunque no destaquen en calidad. Esto los hace opciones efectivas en costo.

Consistencia en Rendimiento

Fireworks AI y OpenAI son consistentes en velocidad y latencia. Esto los hace convenientes para aplicaciones en tiempo real.

AWS no es tan consistente a pesar de ser más caro.

4.4 Otras fuentes de datos

Adicionalmente al estudio realizado hemos encontrado fuentes parecidas de datos que por ciertas razones decidimos no utilizar:

- Chatbot Arena estudia productos sólo de la categoría chatbots, lo que de igual manera hubiera limitado mucho el alcance del estudio.
- Hugging Face está centrado solamente en productos Open Source. Considerando que parte importante del desarrollo de IA la están realizando productos propietarios consideramos que sería inconveniente limitarnos a sólo productos de código abierto.

De todas maneras, hacemos un breve análisis de estas fuentes para complementar nuestra investigación.

4.4.1 Tabulación de datos de LLMs (Chatbot Arena)

Ésta es la tabulación de las métricas que ofrece el benchmark de Chatbot Arena.

Tabla 11 Datos de Chatbot Arena

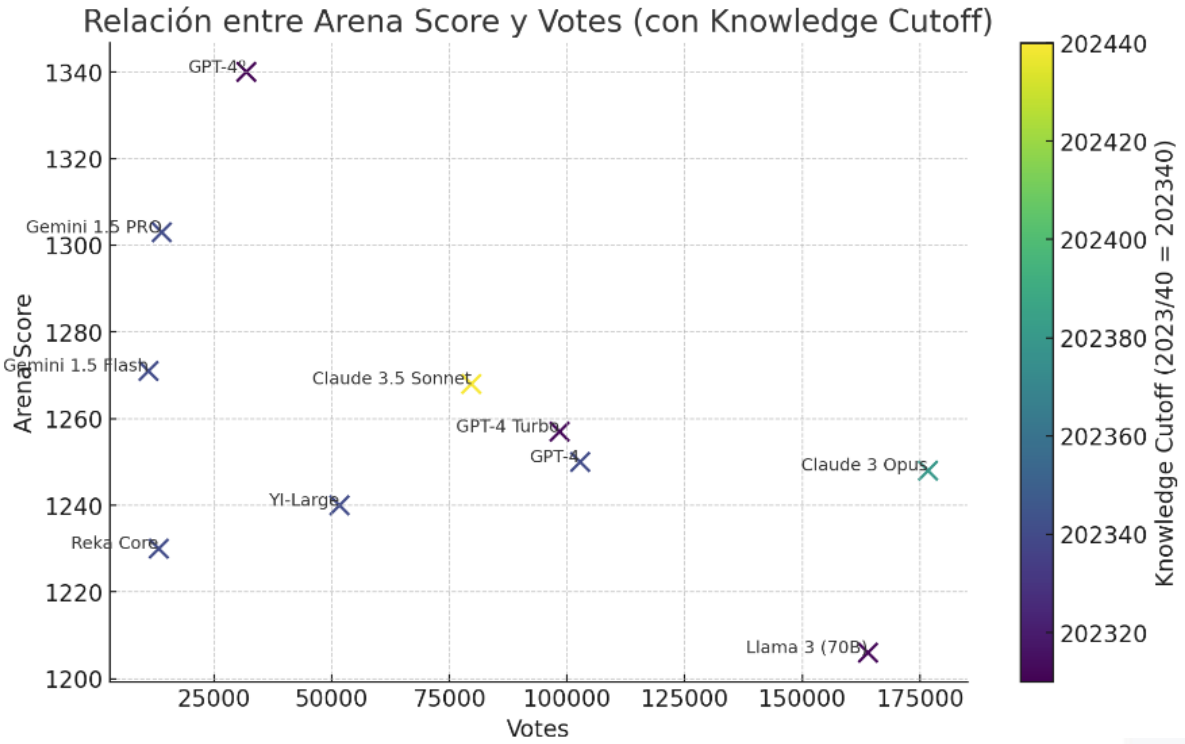
Modelo	Creador	Licencia	Arena Score	Votes	Knowledge Cutoff
GPT-4º	Openai	Propietario	1340	31927	2023/10
Claude 3.5 Sonnet	Anthropic	Propietario	1268	79710	2024/40
Gemini 1.5 PRO	Google	Propietario	1303	13957	
GPT-4 Turbo	Openai	Propietario	1257	98456	2023/12
Claude 3 Opus	Anthropic	Propietario	1248	176723	2023/80
Reka Core	Reka	Propietario	1230	13330	
GPT-4	Openai	Propietario	1250	102833	2023/40
YI-Large	01.Ai	Propietario	1240	51724	
Gemini 1.5 Flash	Google	Propietario	1271	11155	
Llama 3 (70B)	Meta	Open Source	1206	163966	2023/12

Knowledge cutoff es la fecha máxima de los datos con los cuales cada motor fue entrenado.

A continuación, presentamos un breve análisis de este benchmark:

Respecto a la métrica Arena Score, GTP-4o es el que mejor performance en chatbots presenta. Claude 3 Opus es el que más votos registra, lo cual es consistente con su antigüedad.

Ilustración 16 Relacion de datos de Chatbot Arena



Si bien los modelos con mayor cantidad de votos pueden indicar mayor popularidad o confiabilidad, no siempre estos modelos tienen el mejor puntaje general.

4.4.2 Tabulación de datos de LLMs (Hugging Face – open source)

Tabla 12 Datos de Hugging Face

Modelo	Propietario	Average	MATH	MMLU
CalmeRys-78B-Orpo-v0.1	dfurman	50.78	37.92	66.88
calme-2.4-rys-78b	MaziyarPanahi	50.26	37.69	66.69

Rombos-LLM-V2.5-Qwen-72b	rombodawg	45.39	47.58	54.83
RYS Improvement	Dnhkng	44.75	38.97	49.20
calme-2.1-rys-78b	MaziyarPanahi	44.14	36.40	49.38

El mejor rendimiento de los modelos open lo tiene CalmeRys-78B-Orpo-v0.1, podemos observar que los modelos tienen un rendimiento general sin grandes diferencias, destacándose Rombos-LLM-V2.5-Qwen-72b en matemáticas y CalmeRys-78B-Orpo-v0.1 en la capacidad de realizar múltiples tareas de comprensión de lenguaje natural

5 CAPÍTULO V: Conclusiones y Recomendaciones

5.1 Conclusiones

En este capítulo se presentan las conclusiones y recomendaciones del análisis de los Modelos de Lenguaje Grande (LLMs) y APIs estudiados.

5.1.1 Sobre los LLMs

Calidad:

OpenAI tiene los mejores indicadores en la mayoría de las métricas. Esto hace que ofrezca el mejor rendimiento general.

Google y Anthropic también ofrecen buena. Luego aparecen Llama 3 y YI-Large como una alternativa interesante.

Análisis de Costos:

Google y Meta ofrecen las opciones más convenientes. OpenAI justifica su alto precio con una calidad superior.

Los costos combinados (blended) muestran que modelos de menor costo pueden alcanzar un rendimiento parecido al de modelos más caros. Este es el caso de Google.

Rendimiento y Latencia:

Reka Core y Gemini 1.5 Flash ofrecen una buena velocidad, pero su calidad no se acerca a las de OpenAI y Anthropic.

Llama 3 ofrece la mejor latencia, lo que lo hace ideal para aplicaciones en tiempo real. Se observa una correlación entre latencia y calidad.

Chatbot:

GPT-4º y Claude 3.5 son los mejores modelos de pago para chatbot. El modelo open mejor posicionado es Llama 3

5.1.2 Sobre los APIs

Calidad

Los datos de calidad muestran que OpenAI lidera en varias métricas (Index, Chatbot Arena, HUMAN EVAL), seguido de cerca por Anthropic y Gemini 1.5 Pro.

Las APIs de Fireworks AI y Replicate ofrecen una buena relación precio calidad. Son buenas opciones para presupuestos más limitados.

Análisis de Costos

OctoAI y Replicate tienen los precios blended más bajos: $0.90 y $1.18 USD/1M tokens.

OpenAI y AWS tienen los precios más altos: $7.50 y $30.00 USD/1M tokens.

Rendimiento y Latencia

OpenAI y Fireworks AI tienen las mejores velocidades (87.1 y 73.4 tokens/s) y bajas latencias (0.45s y 0.34s). Estos modelos son ideales para aplicaciones en tiempo real.

AWS tiene una latencia bastante mayor (1.83s). Esto sorprende ya que tiene un alto costo

5.2 Recomendaciones

5.2.1 Guía de Selección de IA Generativa

Pequeñas y Medianas Empresas (PYMES)

Para las PYMES se recomienda LLMs que ofrezcan una buena relación costo-calidad. Los modelos de Google y Meta destacan como opciones accesibles y de calidad.

En cuanto a las APIs, OctoAI y Replicate proporcionan un rendimiento sólido a precios convenientes, ideales para presupuestos más ajustados.

Grandes Empresas

Las grandes empresas, pueden optar por LLMs de alto rendimiento como los de OpenAI y Anthropic, que ofrecen calidad superior en generación de texto y capacidad de respuesta, aunque a un costo más elevado. APIs de OpenAI y Anthropic son recomendadas para aplicaciones complejas que exijan precisión y rendimiento de alto nivel.

Startups Tecnológicos

Los startups tecnológicos tienen necesidad de resultados de alta calidad a precios razonables. Por ello OpenAI y Google son muy buenas alternativas.

Las APIs de OpenAI y Fireworks AI, les puede ofrecer la flexibilidad necesaria para su crecimiento.

Instituciones Académicas y de Investigación

Para las instituciones académicas y de investigación las herramientas de código abierto ofrecen muchas ventajas, tanto en costos como en apertura para la experimentación. Llama 3 (70B) de Meta tiene esas ventajas or lo que es una gran alternativa para este segmento.

Entre las APIs, Replicate y OcotAI son alternativas de código abierto que facilitan la investigación y no comprometen su presupuesto.

5.2.2 Recomendaciones para la Implementación

A continuación, se presentan algunas recomendaciones clave para la etapa de implementación de una aplicación que utiliza IA:

- Es importante realizar una planificación detallada antes de emprender un proyecto usando modelos de IA. Hay que cuidar que la integración con sistemas actuales no provoque interrupciones en las operaciones.
- Es recomendable empezar con un proyecto piloto de alcance limitado. Esto permitirá probar la estrategia e ir mejorando en cada iteración.

- La planificación de la infraestructura tiene gran importancia. Se debe garantizar que se cuente con capacidad suficiente para poder escalar según aumente la demanda.
- Se recomienda planificar una etapa de capacitación a los usuarios para prepararlos en el manejo de las funcionalidades de la aplicación a implementar.

Queremos profundizar en dos tipos de organizaciones que pueden aprovechar de gran manera el desarrollo de soluciones basadas en IA: emprendedores e instituciones educativas…

5.2.2.1 Recomendaciones para Emprendedores

Para que un emprendedor pueda llevar a cabo exitosamente un proyecto con IA, hay que considerar sus necesidades y limitaciones específicas. Para este tipo de usuarios tenemos las siguientes recomendaciones para las diferentes etapas del proceso:

Análisis inicial: Se debe considerar el propósito de la solución de la IA. Hay que definir si se usará para automatizar actividades, crear contenido o personalizar las experiencias de los clientes. Esto será un factor importante a la hora de seleccionar las herramientas de IA.

Selección de Herramientas: Al evaluar modelos de lenguaje (LLMs) y APIs podemos considerar modelos con equilibrio entre calidad y costos, como APIs de acceso abierto o LLMs de precio accesible. Google, Meta, OctoAI y Replicate son algunas herramientas que convienen a este segmento.

Prototipado Rápido: Se recomienda desarrollar un proyecto piloto (o prototipo) y así comprobar que la solución funcione correctamente. Al usar ambientes de prueba se evita que errores temporales durante el desarrollo afecten otros procesos.

Escalabilidad: Un correcto análisis de la demanda de recursos nos acercará a dimensionar apropiadamente la inversión. También es importante considerar un futuro crecimiento de la demanda. La infraestructura en la nube es una buena solución en donde se gasta sólo por el uso de esos recursos.

Monitoreo: Un constante monitoreo de los indicadores de la solución nos permitirá responder a eventualidades y considerar incrementos (o decrementos, de ser el caso) de recursos según la variación de la demanda. Hay que desarrollar un plan de mantenimiento para actualizar componentes de software (de código abierto o propietario) y corregir defectos no detectados en el desarrollo e implementación.

Retro alimentación: Es útil dar la posibilidad a usuarios y clientes de comunicar observaciones, reclamos o sugerencias. De esta manera se puede obtener información útil para corregir defectos y desarrollar mejoras.

Siguiendo estas recomendaciones los emprendedores podrán hacer una inversión eficiente que priorice la relación costo / beneficio y facilite la adaptabilidad para cambios futuros en demandas y necesidades.

5.2.2.2 Recomendaciones para Instituciones Educativas

Para que una institución educativa implemente exitosamente soluciones basadas en IA, es importante considerar sus objetivos pedagógicos y su entorno y características particulares. Presentamos recomendaciones prácticas para el sector educativo:

Análisis Inicial: El primer paso es definir el propósito de las soluciones IA a crear. Hay que evaluar si se empleará para personalizar el aprendizaje, automatizar procesos administrativos o crear contenidos interactivos. Así podemos alinear estas necesidades con la selección de herramientas IA.

Selección de Herramientas: Para las organizaciones educativas se recomienda evaluar herramientas que prioricen la calidad y la accesibilidad. Las APIs de IA y modelos de lenguaje (LLMs) de código abierto pueden ser una buena opción para instituciones con presupuestos limitados.

Para este segmento recomendamos Llama 3 de Meta por ser open source y gratuito. También la posibilidad de acceder a código fuente de los modelos LLM permitirá a los estudiantes experimentar con la modificación y creación de variantes de modelos. Si se prefiere servicios provistos por terceros, Google ofrece alternativas económicas y de buena calidad.

Planificación para la Escalabilidad: Considerar el posible crecimiento de las necesidades de los usuarios o de la complejidad de tareas es necesario en las organizaciones educativas. Implementar el uso de una infraestructura escalable, preferentemente en la nube (o implementaciones locales que simula estas funcionalidades) permite que crezca la capacidad sin una inversión inicial elevada.

Monitoreo de Desempeño y Calidad: Implementar un sistema de monitoreo continuo para evaluar la precisión y efectividad de la IA en la generación de contenidos y apoyo a la enseñanza. Es una buena norma de la industria, definir alertas y reportes regularmente para detectar y corregir errores en las soluciones.

Feedback de Usuarios: Permitir a docentes y estudiantes proporcionar feedback sobre la IA instalada ayuda a identificar áreas de mejora. Encuestas o foros de discusión pueden ser valiosos para recolectar sus opiniones, ajustando la herramienta para mejorar su adaptabilidad y alineación con los objetivos pedagógicos.

Capacitación Continua para el Personal: Recomendamos definir una etapa de capacitación al personal que use estas herramientas y así aproveche sus funcionalidades al máximo. Conviene una capacitación continua a los usuarios y desarrolladores considerando que estas herramientas presentan un continuo cambio.

Con estas recomendaciones, las organizaciones educativas podrán obtener beneficios de las aplicaciones que usan IA. Se usarán eficientemente recursos y se adaptarán a las demandas y necesidades cambiantes de estudiantes y docentes.

5.2.3 Casos de uso de IA Generativa

La IA Generativa ya se está usando en diferentes industrias. Los siguientes son algunos casos de uso de aplicaciones desarrolladas con herramientas de IA:

Creación de Contenidos

Muchas empresas están ya produciendo textos, imágenes, artículos y anuncios publicitarios usando motores de generación basados en LLMs. Estas herramientas ayudan a producir estos contenidos a gran velocidad y con alta personalización para diferentes audiencias

En instituciones educativas se puede usar la IA Generativa para elaborar contenidos, materiales, planes de estudio, etc. Para ayudar con las necesidades educativas,

Asistentes Virtuales y Chatbots Inteligentes

Sectores que atienden a clientes finales como bancos o minoristas están creando herramientas que permiten a los usuarios conversar con chatbots especializados en determinados campos.

No solo se puede dar información y ayuda a usuarios, también se puede guiar en procesos y trámites a través de la automatización de procesos y de su integración con otros sistemas transaccionales.

Desarrollo de Prototipos y Diseño de Productos

Existen herramientas que ayudan a crear productos desde su diseño hasta los procesos de fabricación. La IA en arquitectura y construcción, por ejemplo, ayudan a crear planos, diseños interiores cumpliendo condiciones geográficas, regulatorias o de preferencia de clientes.

Análisis de Datos

Una manera importante de aprovechar la gran cantidad de datos que algunas organizaciones poseen es a través de herramientas de Inteligencia de Negocios asistida por IA. En el campo financiero se usan para elegir mejores estrategias y productos de inversión.

También se usan en procesos automatizado de generación de informes ayudando a investigación y a la toma de decisiones.

Personalización de Comercio Electrónico

Uno de los tipos de empresas que aprovecha mucho la gran cantidad de información que tienen son los retails online. El análisis por medio de la IA permite generar recomendaciones personalizadas de productos. Por ejemplo, Amazon y Alibaba se basan en el comportamiento de los usuarios y sus ventas previas para preparar seleccionar que productos nuevos mostrarles.

También lo usan para generar de manera automática descripciones y reseñas de productos.

Atención Médica

La IA generativa en medicina ayuda a elaborar diagnósticos y tratamientos personalizados. Al analizar millones de historias clínicas logran encontrar patrones escondidos a simple vista. Muchos de estos diagnósticos aprovechan la capacidad de análisis y generación visual de algunas herramientas.

6 Referencias bibliográficas

Alan, M. (1950). Turing. Computing machinery and intelligence. *Mind*, *59*(236).

Ammanath, B., Mittal, N., Saif, I., & Anderson, S. (2021). *Becoming an AI-fueled organization Deloitte's State of AI in the Enterprise, 4th Edition A report from the Deloitte AI Institute and the Deloitte Center for Integrated Research*. www.deloitte.com/us/cir.

Barenkamp, M., Rebstadt, J., & Thomas, O. (2020). Applications of AI in classical software engineering. *AI Perspectives*, *2*(1). https://doi.org/10.1186/s42467-020-00005-4

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., Arx, S. Von, Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Chelsea, L. F., Trevor, F., Lauren, G., Karan, G., Noah, G., Lisa, X., Xuechen, L., Tengyu, L., Ali, M., … Liang, P. (2021). On the Opportunities and Risks of Foundation Models arXiv : 2108 . 07258v2 [cs . LG] 18 Aug 2021. *ArXiv*, *2108.07258*.

Bommasani, R., Liang, P., & Lee, T. (2023). Holistic Evaluation of Language Models. *Annals of the New York Academy of Sciences*, *1525*(1). https://doi.org/10.1111/nyas.15007

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. (2020a). *Language Models are Few-Shot Learners*. http://arxiv.org/abs/2005.14165

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. (2020b). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, *2020-December*.

Brynjolfsson, E., Li, D., Raymond, L. R., Acemoglu, D., Autor, D., Axelrod, A., Dillon, E., Enam, Z., Garicano, L., Frankel, A., Manning, S., Mullainathan, S., Pierson, E., Stern, S., Rambachan, A., Reenen, J. Van, Sadun, R., Shaw, K., & Stanton, C. (2023). *GENERATIVE AI AT WORK*. http://www.nber.org/papers/w31161

Chiang, W.-L., Zheng, L., Sheng, Y., Angelopoulos, A. N., Li, T., Li, D., Zhang, H., Zhu, B., Jordan, M., Gonzalez, J. E., & Stoica, I. (2024). *Chatbot Arena: An Open Platform for Evaluating LLMs by Human Preference*. http://arxiv.org/abs/2403.04132

Davenport, T. H., Ronanki, R., Wheaton, J., & Nguyen, A. (2018). FEATURE ARTIFICIAL INTELLIGENCE FOR THE REAL WORLD 108 HARVARD BUSINESS REVIEW. *Harvard Business Review*.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, *1*.

Ebert, C., & Louridas, P. (2023). Generative AI for Software Practitioners. *IEEE Software*, *40*(4), 30–38. https://doi.org/10.1109/MS.2023.3265877

En, G., Matemática, I., De, F., Universidad, M., & De Madrid, C. (n.d.). *REDES NEURONALES CONVOLUCIONALES Y APLICACIONES*.

Feuerriegel, S., Hartmann, J., Janiesch, C., & Zschech, P. (2024). Generative AI. *Business and Information Systems Engineering*, *66*(1), 111–126. https://doi.org/10.1007/s12599-023-00834-7

Finlayson, M., Ren, X., & Swayamdipta, S. (2024). *Logits of API-Protected LLMs Leak Proprietary Information*. http://arxiv.org/abs/2403.09539

Floridi, L., Cowls, J., King, T. C., & Taddeo, M. (2020). How to Design AI for Social Good: Seven Essential Factors. *Science and Engineering Ethics*, *26*(3). https://doi.org/10.1007/s11948-020-00213-5

Gao, M., Ruan, J., Sun, R., Yin, X., Yang, S., & Wan, X. (2023). *Human-like Summarization Evaluation with ChatGPT*. http://arxiv.org/abs/2304.02554

Gartner. (2022). Top Strategic Technology Trends for 2022- Gartner. *Gartner*, *December 2020*.

Gartner. (2024). *Top Strategic Technology Trends 2024*.

Gómez Monsalve, W. D. (2023). Inteligencia artificial generativa e Inteligencia colectiva crowdsourcing para desarrollar Ebooks periodísticos. *Miguel Hernández Communication Journal*, *14*, 471–481. https://doi.org/10.21134/mhjournal.v14i.1997

Huang, M. H., & Rust, R. T. (2021). A strategic framework for artificial intelligence in marketing. *Journal of the Academy of Marketing Science*, *49*(1). https://doi.org/10.1007/s11747-020-00749-9

IBM. (2020). What is Computer Vision? | IBM. In *IBM Topics - Computer Vision*.

Islam, R., & Moushi, O. M. (2024). *GPT-4o: The Cutting-Edge Advancement in Multimodal LLM*.

Jacob, D., Ming-Wei, C., Kenton, L., & Kristina, T. (2021). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. (arXiv:1810.04805v2 [cs.CL] UPDATED). *ArXiv Computer Science*.

Jeyanthi, P. M., Polay, D. H., & Choudhury, T. (2022). The Rise of Decision Intelligence: AI That Optimizes Decision-Making. In *EAI/Springer Innovations in Communication and Computing*. https://doi.org/10.1007/978-3-030-82763-2_7

Khurana, D., Koli, A., Khatter, K., & Singh, S. (2023). Natural language processing: state of the art, current trends and challenges. *Multimedia Tools and Applications*, *82*(3). https://doi.org/10.1007/s11042-022-13428-4

LeCun, Y., Bengio, Y., & Hinton, G. (2015a). Deep learning. Nature. *Nature*, *521*(7553).

LeCun, Y., Bengio, Y., & Hinton, G. (2015b). Deep learning. Nature. *Nature*, *521*(7553).

Madhavi, M. (2019). *Artificial Intelligence in Business Decision Making*. https://blog.zoominfo.com/statistics-about-artificial-intelligence/

McCarthy, J., Minsky, M. L., Rochester, N., & Shannon, C. E. (2006). A proposal for the Dartmouth summer research project on artificial intelligence. *AI Magazine*, *27*(4).

McKinsey Analytics. (2021). The State of AI in 2021. *McKinsey Digital*, *December*.

Mei, K., Li, Z., Xu, S., Ye, R., Ge, Y., & Zhang, Y. (2024). *AIOS: LLM Agent Operating System*. https://github.com/agiresearch/AIOS.

Minaee, S., Mikolov, T., Nikzad, N., Chenaghlu, M., Socher, R., Amatriain, X., & Gao, J. (2024a). *Large Language Models: A Survey*. http://arxiv.org/abs/2402.06196

Minaee, S., Mikolov, T., Nikzad, N., Chenaghlu, M., Socher, R., Amatriain, X., & Gao, J. (2024b). *Large Language Models: A Survey*. http://arxiv.org/abs/2402.06196

Mitchell, T. M. (1997). Machine learning. WCB. *Mac Graw Hill*.

Ortiz, J. M., Juan, Z., Dias, M., Avenburg, A., & Imanol Gonzalez Quiroga, J. (2024). *Sesgos algorítmicos y representación social en los modelos de lenguaje generativo (LLM)*.

Ozkaya, I. (2023). The Next Frontier in Software Development: AI-Augmented Software Development Processes. In *IEEE Software* (Vol. 40, Issue 4, pp. 4–9). IEEE Computer Society. https://doi.org/10.1109/MS.2023.3278056

ÖZPOLAT, Z., YILDIRIM, Ö., & KARABATAK, M. (2023). Artificial Intelligence-Based Tools in Software Development Processes: Application of ChatGPT. *European Journal of Technic*. https://doi.org/10.36222/ejt.1330631

Özpolat, Z., Yildrim, Ö., & Karabatak, M. (2023). Artificial Intelligence-Based Tools in Software Development Processes: Application of ChatGPT. *European Journal of Technic*. https://doi.org/10.36222/ejt.1330631

Patil, S. G., Zhang, T., Wang, X., & Gonzalez, J. E. (2023). *Gorilla: Large Language Model Connected with Massive APIs*. http://arxiv.org/abs/2305.15334

Peng, B., Quesnelle, J., Fan, H., & Shippole, E. (2023). *YaRN: Efficient Context Window Extension of Large Language Models*. http://arxiv.org/abs/2309.00071

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, *21*.

Rai, A., Sharma, D., Rai, S., Singh, A., & Singh, K. K. (2021). IoT-Aided Robotics Development and Applications with AI. In *Advances in Science, Technology and Innovation*. https://doi.org/10.1007/978-3-030-66222-6_1

Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., & Sutskever, I. (2021). Zero-Shot Text-to-Image Generation. *Proceedings of Machine Learning Research*, *139*.

Roch, M., & Reybaud, L. (n.d.). *Arquitectura Transformers: descripción y aplicaciones*.

Russell, S., & Norvig, P. (2021). *Artificial Intelligence, Global Edition A Modern Approach*. Pearson Deutschland. https://elibrary.pearson.de/book/99.150005/9781292401171

Sapiens, K., & Alexander Gelbukh, por. (2010). *Artículos de divulgación Procesamiento de Lenguaje Natural y sus Aplicaciones*. www.google.com.mx/language

Sghir, N., Adadi, A., & Lahmer, M. (2023). Recent advances in Predictive Learning Analytics: A decade systematic review (2012–2022). *Education and Information Technologies*, *28*(7). https://doi.org/10.1007/s10639-022-11536-0

Singla, A., Sukharevsky, A., Yee, L., Chui, M., & Hall, B. (2024). *The state of AI in early 2024: Gen AI adoption spikes and starts to generate value*.

Somers, M. (2019). Emotion AI, explained. In *MIT Sloan*.

Vásquez-Quispesivana, W., Inga, M., & Betalleluz-Pallardel, I. (2022). Artificial intelligence in aquaculture: basis, applications, and future perspectives. In *Scientia Agropecuaria* (Vol. 13, Issue 1, pp. 79–96). Universidad Nacional de Trujillo. https://doi.org/10.17268/SCI.AGROPECU.2022.008

Wang, Y., Ma, X., Zhang, G., Ni, Y., Chandra, A., Guo, S., Ren, W., Arulraj, A., He, X., Jiang, Z., Li, T., Ku, M., Wang, K., Zhuang, A., Fan, R., Yue, X., & Chen, W. (2024). *MMLU-Pro: A More Robust and Challenging Multi-Task Language Understanding Benchmark*. http://arxiv.org/abs/2406.01574

Wu, J., Yang, S., Zhan, R., Yuan, Y., Wong, D. F., & Chao, L. S. (2023). *A Survey on LLM-Generated Text Detection: Necessity, Methods, and Future Directions*. http://arxiv.org/abs/2310.14724

UNEMI

UNIVERSIDAD ESTATAL DE MILAGRO

¡Evolución académica!

@UNEMIEcuador