ESCUELA POLITÉCNICA NACIONAL

FACULTAD DE INGENIERÍA DE SISTEMAS

INGENIERÍA *[carrera]*

PERÍODO ACADÉMICO: 2025-B

ASIGNATURA: ICCD412 Métodos Numéricos GRUPO: GR1CC

TIPO DE INSTRUMENTO: Actividad Extracurricular N°4

FECHA DE ENTREGA LÍMITE: *[29/10/2025]*

ALUMNO: Kevin Eduardo Garcia Rodríguez

TEMA

- *Costos relacionados a los modelos de lenguaje*

OBJETIVOS

- *Conocer los costos creación y mantenimiento de las inteligencias artificiales.*

DESARROLLO

La diferencia fundamental entre ambos conceptos es:

- **Entrenamiento:** Es un proceso de aprendizaje y construcción que consume muchos recursos, se hace una sola vez (o periódicamente para actualizar el modelo) y es muy lento (meses).
- **Inferencia:** Es un proceso de aplicación y ejecución que es rápido (segundos), consume comparativamente pocos recursos *por consulta* y se hace constantemente.

Característica	GPT-4 (OpenAI)	Claude 3 Opus (Anthropic)	Gemini 1.5 Pro/Ultra (Google)	Llama 3 (Meta) (Modelos 70B y 400B+)	Mixtral 8x7B (Mistral AI)
Entrenamiento: GPU/Hardware	Estimado: ~25,000 GPUs NVIDIA A100 (para la versión inicial) o H100.	Estimado: Decenas de miles de GPUs NVIDIA H100 (en clusters de AWS y Google Cloud).	Google TPUs (v4 y v5p). Google usa su propio hardware (Tensor Processing Units), no GPUs de NVIDIA.	NVIDIA H100. Meta reportó públicamente usar 2 clústeres de 24,576 H100s cada uno.	Estimado: Cientos o miles de GPUs NVIDIA H100.
Entrenamiento: Costo Hardware	Especulativo: >$100 millones (solo en costo de cómputo, sin incluir I+D).	Especulativo: Similar o superior a GPT-4. Cientos de millones de dólares.	N/A (Costo Interno). Google fabrica sus TPUs, por lo que el costo es de fabricación, no de compra.	>$1.5 mil millones. (Estimando 24,576 H100s a ~$30k-40k c/u, solo para *un* clúster).	Especulativo: Decenas de millones (mucho más barato que los modelos más grandes).
Entrenamiento: Tiempo	Meses. (Estimaciones varían de 3 a 6 meses para la versión inicial).	Meses. (Se especula un tiempo similar o superior a GPT-4).	Meses. (Tiempo de desarrollo y afinado).	7.7M GPU-horas (para el 400B+); 1.7M GPU-horas (para el 70B). (M = Millones).	Semanas a pocos meses. (Mucho más rápido debido a su arquitectura MoE más pequeña).

Inferencia: GPU/Hardware	GPUs de centro de datos (A100, H100, H200) en la infraestructura de Microsoft Azure.	GPUs H100 y otros aceleradores en AWS y Google Cloud.	Google TPUs (v4, v5e). Conocidos por ser muy eficientes en inferencia.	GPUs H100, A100, L40S (desplegado por muchos proveedores).	GPUs H100, A100.
Inferencia: Consumo Energético	Estimado: ~0.01-0.02 kWh (10-20 Wh) por consulta. (Varía mucho según la longitud).	Estimado: Similar a GPT-4.	Estimado: Muy eficiente. Se cree que los TPUs tienen un menor consumo (Watts) por operación que las GPUs.	Estimado: Comparable a otros modelos de su tamaño.	Estimado: Muy eficiente. Al ser MoE (Mixture of Experts), solo usa una fracción (2 de 8) de sus parámetros por token, reduciendo significativamente el consumo.
Entrenamiento: Consumo Energético (Total)	Estimado: Múltiples GWh (GigaWatts-hora). (Estimaciones para GPT-3 fueron ~1.3 GWh; GPT-4 es mucho mayor).	Estimado: Múltiples GWh.	Estimado: Múltiples GWh.	Estimado: 5.4 - 7.7 GWh (para el 400B+, basado en 7.7M GPU-horas y el TDP del H100).	Estimado: Significativamente menor que los modelos monolíticos gigantes (GPT-4, Llama 400B+).

CONCLUSIONES

- Primero, la diferencia entre entrenamiento e inferencia es simplemente bestial. El entrenamiento es el "modo dios": es donde se queman literalmente *cientos de millones* (o hasta miles de millones, viendo lo que gasta Meta) de dólares. Requiere tener un estadio lleno de las GPUs más caras del planeta (las H100) funcionando a tope durante meses, chupando una cantidad de energía absurda (hablamos de Gigawatts-hora).
- La inferencia, en cambio, es el "modo usuario". Es lo que usamos todos los días, y aunque necesita hardware potente, es muchísimo más barato y rápido por consulta.

- Segundo, esto es una carrera de gigantes. Lo que más me voló la cabeza es el secretismo. OpenAI, Google y Anthropic guardan sus "recetas" (cuántas GPUs, cuánto tiempo) como si fuera un secreto de estado. Es una caja negra. Meta, con Llama 3, al menos es más transparente y básicamente presume que usó clústeres de 24,000 H100s, una salvajada de plata. Google va por libre con sus TPUs, así que ni siquiera podemos comparar costos directamente.
- Básicamente, el resumen es que crear un "cerebro" como GPT-4 o Claude 3 es obscenamente caro, un lujo que solo 3 o 4 empresas en el mundo pueden pagar. Y luego tienes a gente como Mistral, que son como los listos de la clase: usan "trucos" de arquitectura (como el MoE) para intentar conseguir resultados parecidos sin tener que gastar tanto. Al final, todo se reduce a quién tiene más poder de cómputo y quién está dispuesto a pagar la factura de luz más alta.

REFERENCIAS

[1] D. Carrasco, "Historia de OpenAI: así nació la famosa compañía que creó ChatGPT y DALL-E," Marketing4ecommerce.net, 23 de nov. de 2023. [En línea]. Disponible en: https://marketing4ecommerce.net/historia-de-openai-asi-nacio-la-famosa-compania-que-creo-chatgpt-y-dall-e. [Accedido: 26 de oct. de 2025].

[2] Jenni AI, "Guía Maestra de ChatGPT," Jenni.ai. [En línea]. Disponible en: https://jenni.ai/es/chat-gpt/creator. [Accedido: 26 de oct. de 2025].

[3] Marketing Zone Icesi, "La historia OpenAI, la empresa detrás de ChatGPT y DALL-E," Marketing Zone Icesi, 19 de feb. de 2024. [En línea]. Disponible en: https://www.icesi.edu.co/marketingzone/la-historia-openai-la-empresa-detras-de-chatgpt-y-dall-e. [Accedido: 26 de oct. de 2025].