ESCUELA POLITÉCNICA NACIONAL
FACULTAD DE INGENIERÍA EN SISTEMAS
MÉTODOS NUMÉRICOS

Nombre: Camila Caicedo

Curso: GR1CC

Fecha de entrega: 05/11/2025

Actividad extracurricular 4: Costos relacionados a los modelos de lenguaje pendiente

	ChatGPT	Grok	Gemini	Copilot	Titan
Inferencia / entrenamiento	**Entrenamiento:** usa alta precisión y grandes lotes. **Inferencia:** usa baja precisión y es continua.	**Entrenamiento:** con datos sintéticos y paralelismo extremo. **Inferencia:** distribuida con baja latencia.	**Entrenamiento:** multimodal intensivo. **Inferencia:** optimizada para apps móviles y búsqueda.	**Entrenamiento:** masivo. **Inferencia:** optimizada para productividad (código, texto).	**Entrenamiento:** con alto paralelismo. **Inferencia:** optimizada para bajo costo y alta velocidad.
Modelo de GPU utilizado	NVIDIA H100, A100	NVIDIA H100	Google TPU v4/v5e	Azure NVIDIA A100/H100 clusters	AWS Trainium
Costo del hardware	Aproximadamente 300 – 600 millones USD	Aproximadamente 200–500 millones USD	Aproximadamente 10–20 millones USD	Aproximadamente 900 millones – 3 mil millones USD	Aproximadamente 30–100 millones USD
Tiempo de entrenamiento	Varios meses	Varios meses	Fine-tuning continuo	Semanas a meses	Meses
Consumo energético (watts)	**Entrenamiento:** Aproximadamente 7–15 MW (30–90 días) → cientos MWh **Inferencia:** ~0.01–0.5 Wh por consulta	**Entrenamiento:** 20–25 MW (~10–25 GWh). **Inferencia:** MW escala global	**Entrenamiento:** varios MW (decenas GWh) **Inferencia:** MW total a escala global	**Entrenamiento:** MWh por corrida **Inferencia:** Wh por consulta; uso total en MW	**Entrenamiento:** MW por pod (MWh totales) **Inferencia:** MW en despliegue global

- **DIFERENCIA ENTRE ENTRENAMIENTO E INFERENCIA**

  - **Inferencia:** proceso mediante el cual un modelo de IA entrenado genera nuevos resultados razonando y haciendo predicciones sobre nuevos datos, clasificando las entradas y aplicando el conocimiento aprendido en tiempo real.

  - **Entrenamiento:** es la primera fase de un modelo de IA, puede implicar un proceso de ensayo y error, o un proceso de mostrar al modelo ejemplos de las entradas y salidas deseadas, o ambos.

  - La inferencia no puede darse sin entrenamiento.

Fuentes de consulta:

- https://www.cloudflare.com/learning/ai/inference-vs-training/