Métodos Números

Isaac González

18/10/2025

[Actividad extracurricular 04] Costos relacionados a los modelos de lenguaje

Indicaciones

Investigue sobre las siguientes características para al menos **5 modelos de lenguaje comerciales** (i.e. ChatGPT, Claude, Gemini, etc.) y cree una tabla resumen.

Aspectos a analizar:

- ¿Qué es inferencia y entrenamiento? ¿Cuál es la diferencia?
- Modelo de GPU utilizado/s
- Costo del hardware (costo GPU × número de GPUs) en inferencia y entrenamiento
- Tiempo de entrenamiento
- Consumo energético (watts) en inferencia y entrenamiento

Comparativa de GPUs, potencia y costos de entrenamiento (2025)

Modelo	GPUs utilizadas	Potencia típica (W)	Costo por GPU (USD)	Costo total estimado (USD)
GPT-5 (OpenAI)	≈ 200 000 H100/H200	700 W (SXM) / 350 W (PCIe)	2 – 7 USD / h	≈ 500 – 700 M USD
Grok-4 (xAI)	≈ 246 M H100 horas	700 W	1.9 – 2.2 USD / h	≈ 490 M USD
Llama 3 / 3.1 (Meta)	≈ 16 000 H100	700 W	2 – 5 USD / h	≈ 50 – 100 M USD
Claude 3.5 / 3.7 Sonnet (Anthropic)	No divulgado	700 W	—	≈ 20 – 30 M USD
DeepSeek V3 / V3.2 (China)	≈ 2 048 H800	500 – 700 W	—	< 6 M USD

> *Datos aproximados basados en informes técnicos y estimaciones 2024 – 2025.*

Diferencia entre **entrenamiento** e **inferencia**

- **Entrenamiento:** proceso donde el modelo aprende ajustando sus parámetros a partir de grandes volúmenes de datos. Requiere un uso intensivo de GPU, energía y tiempo.
- **Inferencia:** Fase en la que el modelo ya entrenado genera respuestas o predicciones. Es menos costosa y más rápida, pero depende del tamaño del modelo y su despliegue.
