Actividad_extracurricular_04

November 4, 2025

1 Escuela Politécnica Nacional

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

1.1.1 Nombre: Luis Alexander Lema Delgado

1.1.2 Fecha: 02/11/2025

1.1.3 Curso: GR1CC

2 Investigación sobre Modelos de Lenguaje Comerciales

En esta parte se investigaron algunas características sobre distintos modelos de lenguaje comerciales actuales, como ChatGPT, Claude, Gemini, Llama y Mistral.
El objetivo es comparar qué tipo de hardware utilizan, cuánto cuestan, el tiempo que tardan en entrenarse y su consumo de energía.

2.1 ¿Qué es inferencia y entrenamiento?

Primero es importante entender la diferencia entre entrenamiento e inferencia. **Entrenamiento:**
Es la etapa en la que el modelo aprende a partir de una gran cantidad de texto. Durante este proceso se ajustan los parámetros internos para que el sistema pueda reconocer patrones y generar respuestas coherentes.
Esta fase es la más costosa y requiere miles de GPUs trabajando durante semanas o meses.

Inferencia:
Ocurre cuando el modelo ya entrenado genera respuestas a partir de una entrada. Por ejemplo, cuando se le hace una pregunta a ChatGPT, el modelo realiza una inferencia.
Aquí ya no aprende, solo utiliza lo que sabe para dar una salida.

En pocas palabras:
El entrenamiento es el proceso de aprendizaje, mientras que la inferencia es la aplicación de lo aprendido.

2.2 Aspectos a comparar

Para cada modelo se consideraron los siguientes puntos:

1. Tipo de GPU utilizada.

2. Costo total del hardware (precio de una GPU por el número de unidades).

3. Tiempo aproximado de entrenamiento.

4. Consumo energético tanto en el entrenamiento como en la inferencia.

2.3 Tabla resumen de modelos de lenguaje

Los valores son aproximados y se basan en estimaciones públicas. No todas las empresas revelan cifras exactas.

Modelo de Lenguaje	GPU utilizada	Costo total aprox. (USD)	Tiempo de Entrenamiento	Consumo Energético	Comentario
ChatGPT (GPT-4)	NVIDIA H100 ($8,000 c/u) × 10,000	$80 millones	2 a 3 meses	Entrenamiento: ~1.2 GWhInferencia: ~0.5 kWh / 1k tokens	Modelo muy grande y costoso, pero de alto rendimiento.
Claude 3 (Anthropic)	NVIDIA A100	$40 millones	2 meses	900 MWh / 0.3 kWh	Optimizado para manejar contextos largos.
Gemini 1.5 (Google)	TPU v5e (Google)	$50 millones	2 a 3 meses	1 GWh / 0.4 kWh	Usa hardware propio de Google.
Mistral 7B	NVIDIA A100 80GB × 512	$4 millones	1 mes	100 MWh / 0.1 kWh	Modelo más pequeño y de código abierto.
Llama 3 (Meta)	NVIDIA H100 × 2,000	$16 millones	1.5 meses	300 MWh / 0.2 kWh	Modelo abierto y relativamente eficiente.

2.4 Conclusión

Al comparar los modelos, se puede ver que el entrenamiento es la parte más exigente en recursos. Requiere gran cantidad de hardware, altos costos y mucho tiempo. Por eso, solo empresas grandes como OpenAI o Google pueden entrenar modelos de este nivel.

En cambio, la inferencia es mucho más ligera, lo que permite que los usuarios comunes podamos usar estos sistemas en línea sin requerir tanto poder de cómputo.
También se nota que los modelos más nuevos buscan ser más eficientes y menos costosos energéticamente, lo cual es importante para reducir el impacto ambiental y hacerlos más accesibles.

Resumen general: - El entrenamiento es la fase donde el modelo aprende, y la inferencia es cuando usa lo aprendido.
- Las GPUs más usadas son las NVIDIA A100 y H100, y las TPUs en el caso de Google.
- Entrenar un modelo grande puede costar decenas de millones de dólares.
- Los nuevos modelos intentan mejorar la eficiencia energética y el acceso abierto.