Escuela Politécnica Nacional

Métodos Numéricos

Nombre:Lenin Amangandi

Costos relacionados a los modelos de lenguaje

1. ¿Qué es inferencia y entrenamiento?

El **entrenamiento** es cuando el modelo aprende. Se le dan millones de textos y ejemplos para que ajuste sus parámetros. Este proceso se hace una sola vez o muy pocas veces, y requiere miles de GPUs y mucha energía.

La **inferencia** ocurre después, cuando el modelo ya está listo y solo se usa para responder preguntas o generar texto. Aquí no aprende, solo aplica lo que ya sabe.

Palabra Clave

- **Entrenamiento** = Aprender.
- **Inferencia** = Usar lo aprendido.

2. Modelo de GPU utilizado

- **GPT-4 (OpenAI):** GPUs NVIDIA A100 y H100.
- **Claude 4 (Anthropic):** Estimación similar a A100 o H100.
- **Gemini 2.5 Pro (Google):** TPUs v4 o v5.
- **GPT-3 (OpenAI):** GPUs NVIDIA V100 y A100.
- **LLaMA 3 (Meta):** Se cree que GPUs A100 o equivalentes.

3. Costo del hardware (GPU × número de GPUs)

- **GPT-4 (OpenAI):** Se estima que se usaron unas 25 000 GPUs A100, con un costo aproximado de US$ 700 millones solo en hardware.
- **Claude 4 (Anthropic):** No hay cifras oficiales, pero se calcula que el gasto estuvo en decenas de millones de dólares.
- **Gemini 2.5 Pro (Google):** El desarrollo de la versión más grande habría costado alrededor de US$ 191 millones.
- **GPT-3 (OpenAI):** Se estima que el costo del hardware fue entre US$ 4 y 12 millones.
- **LLaMA 3 (Meta):** No hay datos confirmados, aunque se cree que el gasto estuvo en el rango de millones de dólares, dependiendo del tamaño del modelo.

4. Tiempo de entrenamiento

- **GPT-4:** varios meses (≈90 días).
- **Claude 4:** meses (estimado).
- **Gemini:** varios meses.
- **GPT-3:** 1–2 meses.
- **LLaMA 3:** semanas o meses, según el tamaño.

5. Consumo energético (Watts)

- **Entrenamiento:** Una GPU A100 consume 400–700 W/h. Miles de GPUs trabajando semanas corresponde a millones de kWh.
- **Inferencia:** Consume menos que el entrenamiento, pero sigue siendo alto con uso masivo.

Ejemplos estimados:

- **GPT-4:** consumo total muy alto, equivalente al de una ciudad pequeña durante el entrenamiento.
- **Gemini y Claude:** similares, aunque más eficientes con chips nuevos.
- **GPT-3 y LLaMA:** menor consumo que GPT-4, por ser modelos más pequeños.