ESCUELA POLITÉCNICA NACIONAL

FACULTAD DE INGENIERÍA EN SISTEMAS

METODOS NUMERICOS ICCD412

JUAN FRANCISCO PINTO ANDRANGO

Actividad Extracurricular 04 Costos relacionados a los modelos de Lenguaje

GR1CC

FECHA DE ENTREGA 05 DE NOVIEMBRE DEL 2025

Indicaciones

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

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

Que es inferencia y entrenamiento, cual es la diferencia

la inferencia es un proceso con el cual podemos obtener conclusiones, realizar decuciones he interpretaciones con la informacion disponible como datos, hechos y evidencias. la inferencia es una habilidad importante para la toma de decisiones, resolucion de conflictos y deducion de situaciones completas. la inferencia es una etapa en la que el modelo aplica los conocionmentos para entregar una respuesta o generar una prediccion.

el entrenamiento son actividades de tipo sistematico y planificado para generar o mejorar capacidades especificas, tienede a ser una etapa en la que un modelo aprende en base a grandes cantidades de datos esta etapa es intensa y costosa.

las inferencias se diferencian del entrenamiento por la aplicacion y uso, la respuesta o predicion, un dessarrollo en poco tiempo. En el entrenamiento se aprende y construye un modelo, trabaja con grandes volumenes de datos, tiende a ser costoso y siempre precede a la implementacion.

Modelos de lenguaje	GPU utilizados	Costo del Hadware	Tiempo de Entrenamiento	Consumo Energetico Entrenamiento	Consumo Energetico Inferencia
GTP-4 (OpenAI)	NVIDIA H100 O A1000	500 $ millones USD	de 60 a 90 Dias	25.000 - 50.000 MWh	Alto consumo por token de $20/1M
DeepSeek-V3 (DeepSeek)	NVIDIA H800 (2.048 unidades)	5-20 $ millones USD	60 dias	10.7000 MWh	Bajo consumo por token de $1/1 M
Claude 3 Opus (Anthropic)	NVIDIA H100 o A100	200 $ millones USD	45 - 60 dias	15.000 - 30.000 MWh	Alto consumo por token de $ 75/1 M

| Gemini Ultra (Google)|Google TPUv4/v5 |N/A(Hardware propietario) |60 - 120 dias | 30.000 - 60.000 MWh | Optimizado por TPU |Llama 3 70B (Meta/Open)| NVIDIA H100(16.000 unidades)| 10 - 20 $millonesUSD|20dias|1.200MWh|Moderapuedeserauto-alojado||Mixtral8x22B(MinstralAI/Open)|NVIDIAH100|5-10$ millones USD | 14 -21 dias | 800 MWh | Bajo (optimizado para MoE) |

link del repositorio de Git-hub

https://github.com/JuanfranPinto/Metodos-Numericos-