Aplicaciones LLM

Coste de las LLM Apps

Coste

- Inference cost.
- In-context learning cost.
- Training cost.

Coste: inference cost

- Pagas por el número de tokens procesados.
- Este coste puede dispararse muy rápidamente.
  - Resumir una página en chatGPT utiliza 700 tokens y cuesta 0.015 dólares.
  - Resumir 1500 páginas utiliza 1M tokens y cuesta 20 dólares.

Coste: in-context learning cost

- La forma más habitual de in-context learning es construir una LLM App con la técnica RAG:
  - Utilizar un modelo fundacional para generar lenguaje.
  - Añadir una base de datos privada.

- Es una alternativa más barata que pre-training o fine-tuning.
- No tiene training costs.
- Es el enfoque más recomendable para la mayoría de proyectos. Construir un LLM desde cero y fine-tuning son alternativas muy caras.

Coste: fine-tuning

- Fine-tuning: entrenar un modelo fundacional con tu propia base de datos. Tiene un coste prohibitivo en consumo y GPUs para la gran mayoría de las empresas.

Coste: training an LLM model from scratch

- Training from scratch: entrenar un modelo LLM desde cero requiere cientos de miles de horas de computación. El coste aumenta exponencialmente con el tamaño del modelo y de la base de datos de entrenamiento. Tiene un coste prohibitivo en consumo y GPUs para la gran mayoría de las empresas.