

Entraînement vs inférence



1. Entraînement

Le modèle apprend à générer du texte.

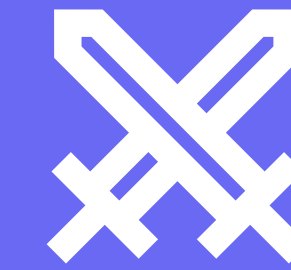
Beaucoup d'énergie consommé pendant plusieurs semaines, mais fait qu'une seule fois par modèle.



2. Utilisation (inférence)

Le modèle génère du texte à partir d'un prompt.

Peu d'énergie consommée par demande, mais si beaucoup de demande, la consommation cumulée devient considérable.



Entraînement vs inférence



1. Entraînement

Le modèle apprend à générer du texte.

Beaucoup d'énergie consommé pendant plusieurs semaines, mais fait qu'une seule fois par modèle.

Exemple de consommation d'entraînement

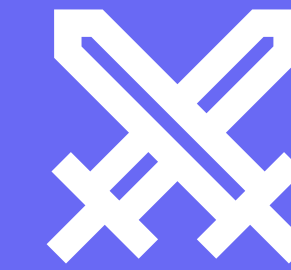


Meta/Llama 4

= 21 000 GWh
consommés pour
l'entraînement

= 10 000 ménages
pendant un an

source : Données de Meta sur LLama 4



Entraînement vs inférence



2. Utilisation (inférence)

Le modèle génère du texte à partir d'un prompt.

Peu d'énergie consommée par demande, mais si beaucoup de demande, la consommation cumulée devient considérable.

Exemple de consommation d'inférence :



OpenAI/**GPT 4o**

Méthodologie : Ecologits

= 14.9 Wh
consommés pour
écrire un mail
(170 jetons)

= 1 smartphone
rechargé



Entraînement vs inférence



2. Utilisation (inférence)

Le modèle génère du texte à partir d'un prompt.

Peu d'énergie consommée par demande, mais si beaucoup de demande, la consommation cumulée devient considérable.

Exemple de consommation d'inférence :



OpenAI/**GPT 4o**

Méthodologie : Ecologits

= 14.9 Wh
consommés pour
écrire un mail
(170 jetons)

= 1 smartphone
rechargé

Si 1% de la population mondiale le faisait tous les jours pendant un an, ce serait équivalent à ✈️ 1,502 vols Paris ↔ NYC en termes d'émissions.



Entraînement vs inférence



2. Utilisation (inférence)

Le modèle génère du texte à partir d'un prompt.

Peu d'énergie consommée par demande, mais si beaucoup de demande, la consommation cumulée devient considérable.



Question posée sur X (Twitter)

« Combien d'argent OpenAI [les créateurs de ChatGPT] perd à cause des gens qui disent "s'il te plaît" et "merci" à leurs modèles ? »

Réponse de Sam Altman, PDG d'OpenAI

« Des dizaines de millions de dollars [...] »