



L'empreinte environnementale de l'IA générative et des agents de développement

ANF DevLog IA4Dev · Grenoble · octobre 2026

Tristan Coignon · Inria - Université de Lille

Qui suis-je ?

Tristan Coignon

Maître de conférences
Université de Lille

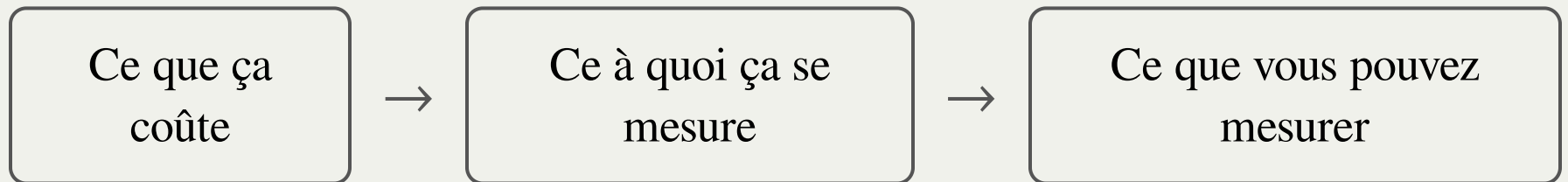
Mes recherches portent sur les impacts
environnementaux et sociétaux des IA
génératives

tristan.coignon.fr



Une semaine de développeur remplacée par un run
d'agent de 5 minutes :
meilleur pour l'environnement ?

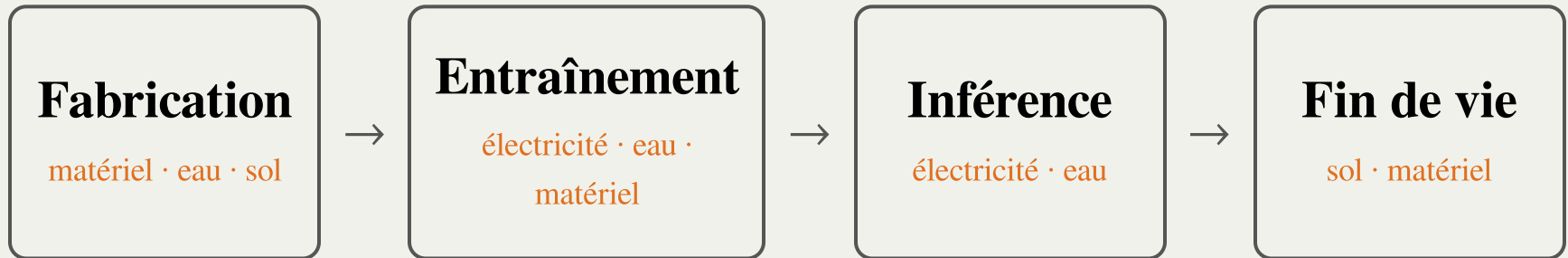
Objectif : utiliser l'IA générative en connaissance de cause



2

Où se situe physiquement
l'empreinte d'un modèle ?

L'empreinte de l'IA s'étale sur quatre ressources : électricité, eau, matériel, sol



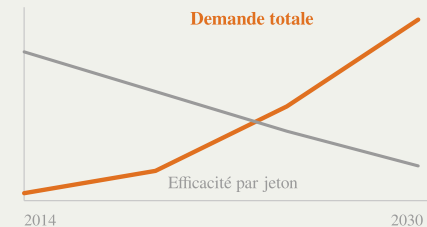
La fin de vie est l'étape que la plupart des ACV omettent.

La consommation électrique des datacenters croît plus vite que les gains d'efficacité

165 → 420 TWh

datacenters, 2014 → 2024

1 500 TWh projetés en 2030



L'eau se consume en trois endroits

Refroidissement
datacenter

Refroidissement
des centrales

Eau ultra-pure
pour puces

4,2-6,6 milliards de m³

d'eau pour l'IA d'ici 2027, soit 4-6 Danemark, la moitié du Royaume-Uni

Les datacenters évaporent 1-9 L par kWh d'énergie serveur

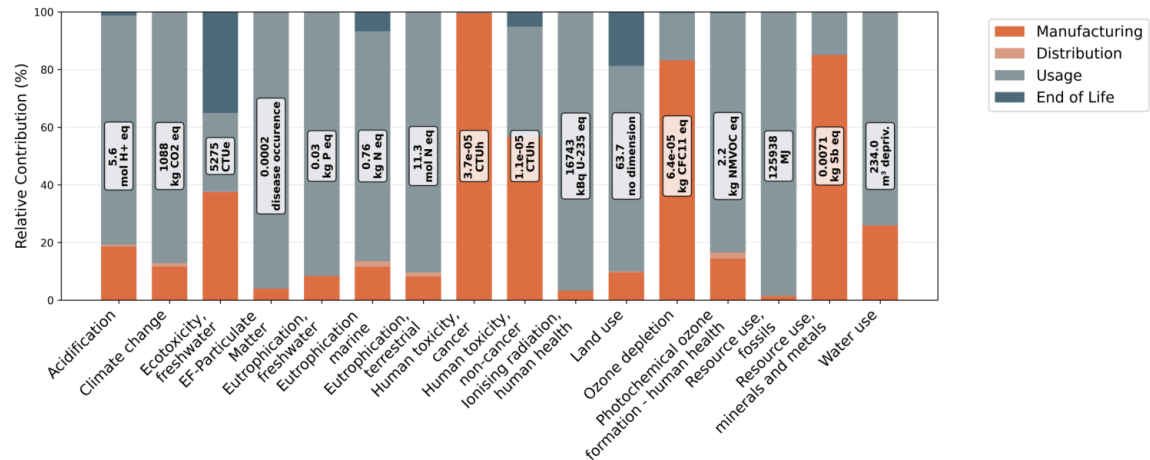


1 datacenter sur 4 est déjà en zone de déficit d'eau au moins un mois par an,
en concurrence avec l'eau potable et l'irrigation.

La fabrication : toxicité et épuisement minéral

99 %

de l'impact toxicité
humaine (cancer)
de l'A100 SXM
vient de la
fabrication



Contribution des étapes du cycle de vie par catégorie d'impact, GPU A100 SXM

La fabrication commence dans des mines



Creuseurs de la mine artisanale de cobalt de Shabara, près de Kolwezi (RDC) : 76 % du cobalt mondial y est extrait

Photo : Junior Kannah / AFP, 12 octobre 2022, mine artisanale de Shabara près de Kolwezi (via La Croix)

U.S. Geological Survey, *Mineral Commodity Summaries*, janvier 2025

Amnesty International & Afrewatch 2016, « *This Is What We Die For* » : *Human Rights Abuses in the DRC Power the Global Trade in Cobalt*

Quatre ressources,
un multiplicateur : l'usage

3

Combien coûte
réellement chaque usage ?

Requête Google vs requête LLM : une comparaison populaire, et pourquoi elle est contestée

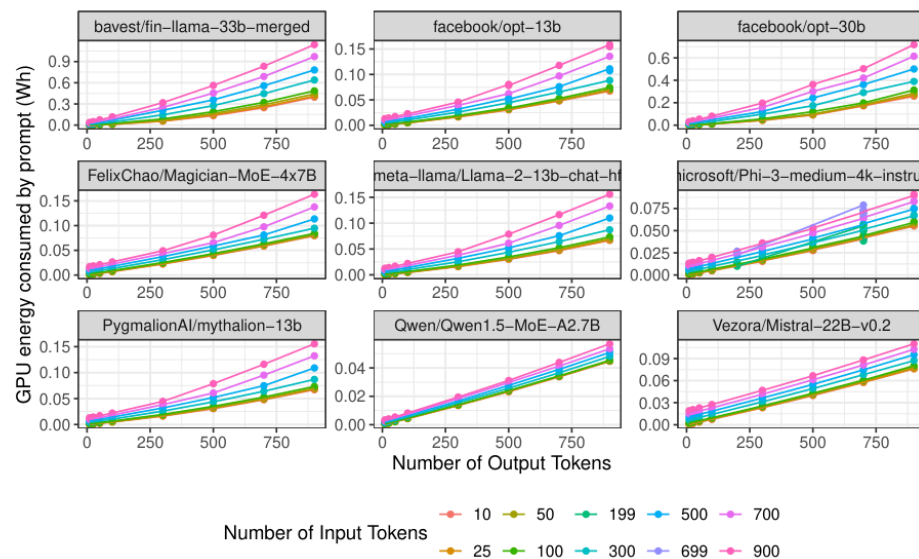
estimations 2009-2023

0,3 Wh par recherche Google (auto-estimation de 2009, jamais re-mesurée) vs **2,9 Wh** par requête ChatGPT : $\times 10$

mesures 2025

0,24 Wh par prompt Gemini (Google, production) · **~0,3 Wh** par requête GPT-4o (Epoch AI)

La longueur du prompt est secondaire ; la longueur de la réponse pilote le coût



L'énergie croît quasi linéairement avec les jetons de sortie, et bien plus qu'avec les jetons d'entrée : contexte long bon marché, générations longues non.

Une image coûte $\approx 60\times$ une requête texte, une vidéo courte $\approx 30\times$ une image

l'effort du prompt est le même pour les trois : ce qu'il génère fixe le prix · longueurs de barres en échelle logarithmique, non à l'échelle



✧ la génération de vidéo : coût $\approx$ quadratique en résolution et en durée

Un run d'agent multiplie le coût de la requête par le nombre de tours

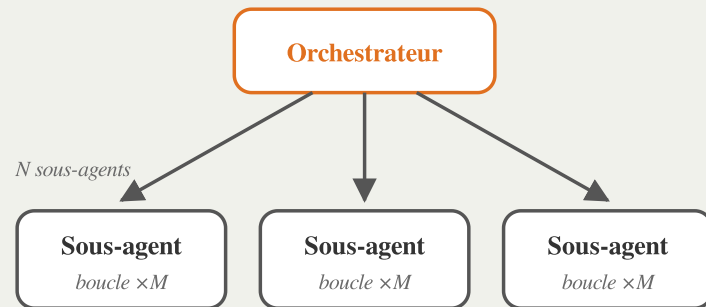
13×

l'énergie par requête : requêtes 15× plus longues



répéter : chaque tour relit le contexte et régénère

Les sessions multi-agents se ramifient : N sous-agents $\times M$ tours, et le framework coûte jusqu'à $9,4\times$



$9,4\times$

AutoCodeRover a consommé $9,4\times$
l'énergie d'OpenHands sur les
mêmes tâches

La multiplication se compose : l'architecture est le premier facteur énergétique, avant la taille des modèles.

En une requête : 2,7 % des *issues* résolues. En boucle d'agent : 18 %

	Qualité issues SWE-bench Lite résolues	Coût par tâche
Requête unique	2,7 % (recherche puis réponse)	0,13 \$
Agent en boucle	18,0 % (même modèle)	1,67 \$
Multi-agents	gains croissant avec la difficulté	se compose : $N \times M$ tours

La qualité suit l'architecture, mais les rendements décroissent avec le nombre d'appels.

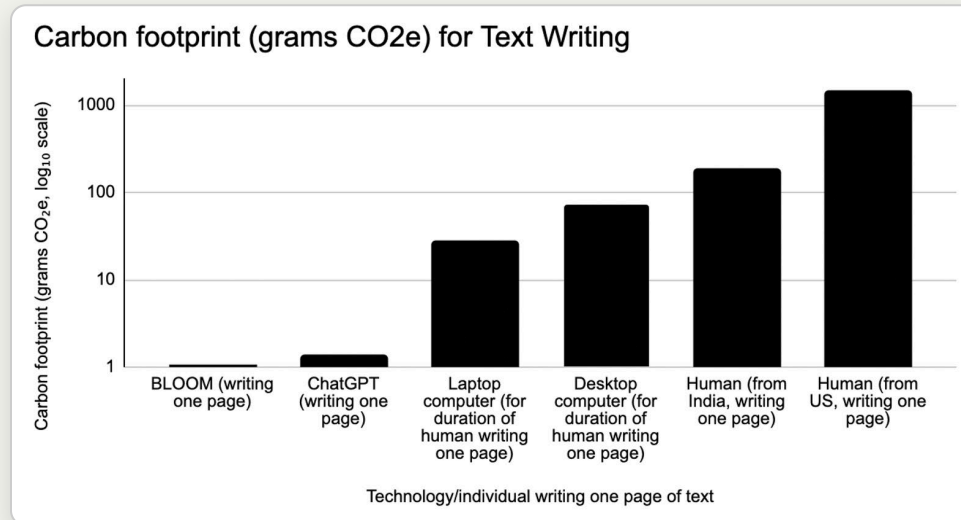
Petits modèles vs grand modèle : 60-70 % d'énergie économisée à précision égale sur les tâches routinières

	SLM petits modèles	LLM grand modèle
Qualité	égale sur le routinier, chute sur le raisonnement abstrait	stable, y compris sur le raisonnement difficile
Énergie & eau	60-70 % de moins	référence
Latence	réponse plus rapide	réponse plus lente

4

Un logiciel moins cher :
plus écologique, ou pas ?

À l'échelle de la tâche, le run assisté par IA émet 1 000× moins qu'un humain (États-Unis)



1 400 g vs
1,35 g

CO₂e pour écrire une page
(humain, États-Unis vs ChatGPT)
· illustrer une image : 5 500 g vs
2,2 g

CO₂e par page de texte, échelle log (figure 1 de l'article, CC BY 4.0)

À périmètres alignés, la réanalyse inverse le classement

CO ₂ e par page écrite	humain	IA
étude d'origine	1 411 g	1,4 g
réanalyse, périmètres alignés	9,3 g	11,7 g

réanalyse : bottom-up des deux côtés, matériel inclus, l'humain reste présent dans le scénario assisté (cadre EGDC)

La comparaison par tâche confronte deux **produits**. Une décision pose une autre question : l'écriture assistée **remplace** la pratique, ou s'y **ajoute** ?

Les effets indirects qui décident de l'impact net restent hors périmètre.

Trois formes de rebond :

direct : moins cher/appel → plus d'appels

indirect : l'IA remplace des pratiques → de nouveaux usages

temporel : les économies d'aujourd'hui → l'échelle de demain

L'impact indirect passe par la production logicielle elle-même



L'économie par tâche ne devient un gain global que si le nombre de tâches reste fixe

émissions du parc = émissions par tâche × nombre de tâches : la comparaison par tâche ne fixe que le premier facteur

À volume constant

le scénario comptable : autant de tâches qu'avant, l'économie par tâche se retrouve au niveau du parc (une année, tout le logiciel exécuté)

CO₂e du parc



Quand produire devient moins cher

ce qui arrive en pratique : le même total du parc, où l'inférence unitaire baisse mais plus de logiciel est produit et exécuté

dont inférence des runs



+ croissance du logiciel produit



Une grande partie de la demande est créée par les choix de conception



des fonctions IA ajoutées par défaut sur les outils du quotidien

« Le développement des réseaux ne vise pas à servir une demande préexistante : c'est lui qui stimule cette demande. »

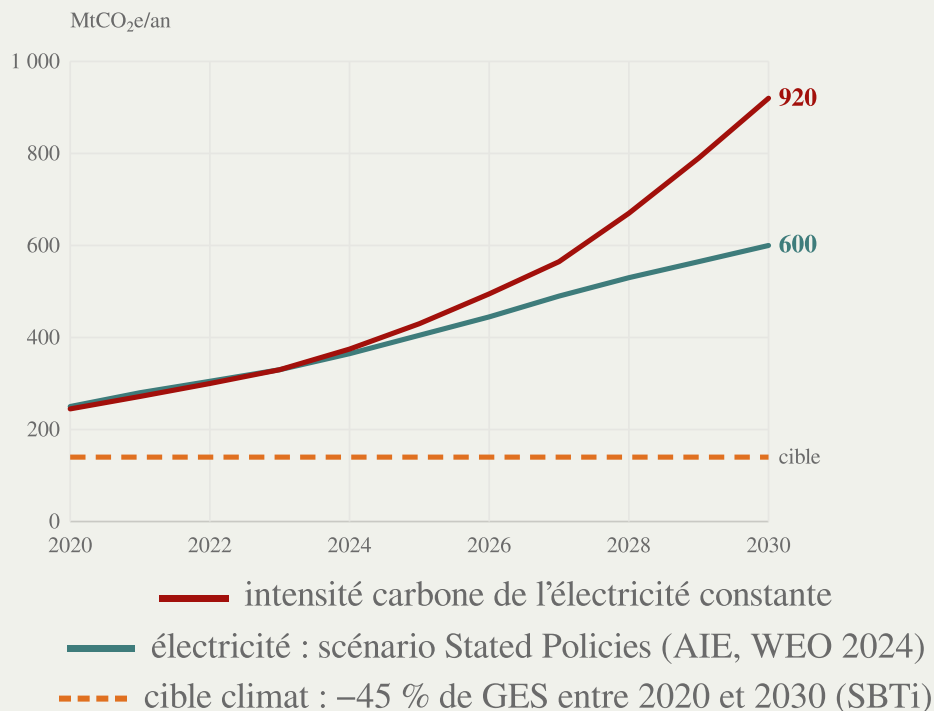
Morand et al. 2026, à propos du déploiement de la 5G · même logique pour les modèles de ML

Morand et al. 2026, *When rebound effect is not a side effect: analyzing sociotechnical contexts of digital technologies*

Morand, Ligozat & Névéol 2025, *The Environmental Impacts of Machine Learning Training Keep Rising, Evidencing Rebound Effect*

de Vries 2023, *The growing energy footprint of artificial intelligence*

Même en décarbonant l'électricité, les émissions des centres de données font $\times 2,5$ à $\times 3,7$ d'ici 2030



920

MtCO₂e/an en 2030, deux fois les émissions annuelles de la France

« Les trajectoires de déploiement actuelles sont donc incompatibles avec la double contrainte carbone, qu'importe les progrès technologiques »

Les émissions activées par les gains IA dépassent celles qu'ils évitent



optimisation
du réseau
électrique



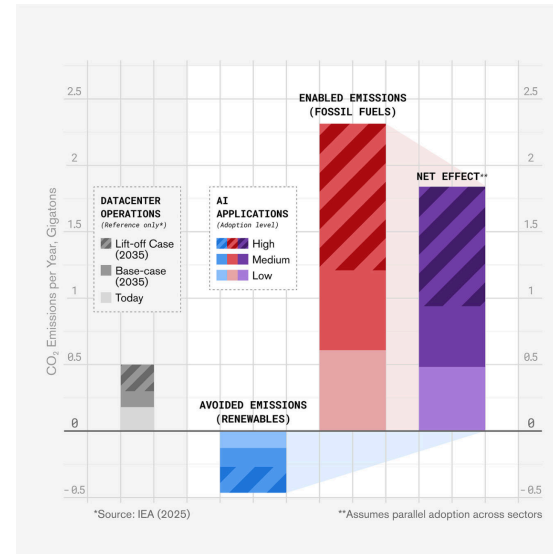
découverte de
matériaux et
de batteries



prévision
météo



gestion
énergétique
des bâtiments



émissions évitées (bleu : gains IA sur les renouvelables)
contre émissions activées (rouge : gains IA sur les fossiles) ;
en violet, l'effet net de l'adoption parallèle

+0,47-1,8 Gt CO₂/an

effet net, adoption parallèle (1,2-4,8 % des émissions énergie 2024) · activé : 0,6-2,4 Gt

Les économies par tâche sont réelles.
La croissance qu'elles activent aussi.

5

Mesurez votre propre
usage

EcoLogits estime l'énergie de chaque appel d'API LLM en Python

```
from ecologits import EcoLogits
from openai import OpenAI

EcoLogits.init(providers=["openai"])
client = OpenAI()
r = client.chat.completions.create(
    model="gpt-4o-mini", messages=[{"role": "user", "content": "write an email"}])
print(r.impacts.energy.value, "Wh")
```

3,96 Wh · 1,63 gCO₂eq · 14,9 mL d'eau · 6,62 µgSbeq



Angles morts : pas de cache ; jetons d'entrée non comptés

une requête « write an email » (calculateur EcoLogits)

CodeCarbon mesure le CO₂ de tout processus Python, matériel inclus

```
from codecarbon import EmissionsTracker

tracker = EmissionsTracker()
tracker.start()
train_my_model()
emissions = tracker.stop() # kg CO2eq
```

Ou usage en CLI

```
codecarbon monitor --no-api -- python train.py
```

Neuralwatt facture l'énergie réellement consommée, en kWh plutôt qu'en jetons

chaque réponse API renvoie :

```
{  
  "model": "deepseek-v4.1-flash",  
  "total_tokens": 13,  
  "energy_joules": 60.37,  
  ...  
}
```

10 \$/kWh

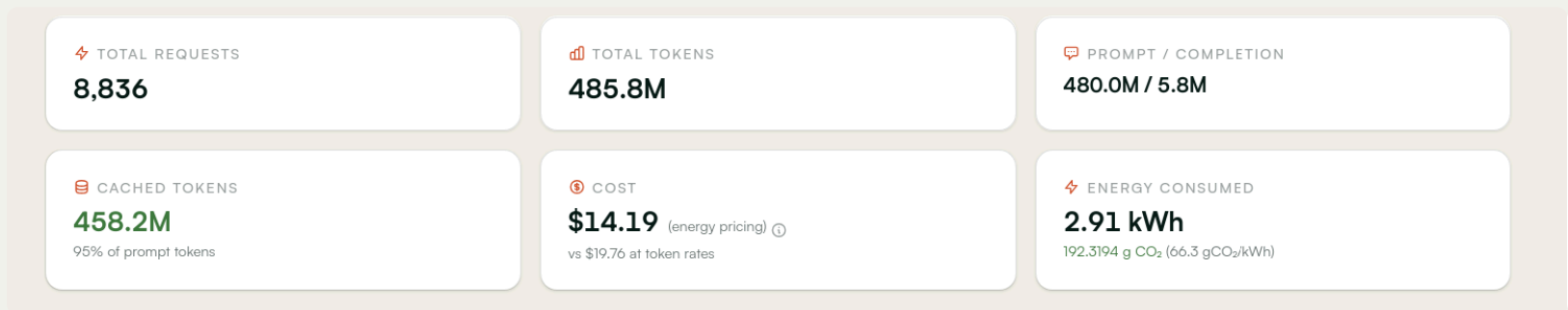
un tarif unique, tous modèles (moins en abonnement)

énergie mesurée par requête

DeepSeek V4.1 Flash **≈ 178 mWh/requête**

DeepSeek V4 Flash **≈ 263 mWh/requête**

Ma propre conso sur Neuralwatt.



95 % des jetons de prompt servis depuis le cache · 2,91 kWh et 192,3 g CO₂ au total

Chaque mesure est une fourchette : le même appel « write an email » pèse de 2,52 à 5,4 Wh

3,96 Wh

2,52 Wh

5,4 Wh

estimation centrale et fourchette affichée par le calculateur EcoLogits



Bon outil : tendances, ratios, ordres de grandeur



Mauvais outil : un inventaire carbone à déclarer

Une grille de frugalité : bien dimensionner le modèle, les tours, et mesurer d'abord

-
- ✓ Choix du modèle
 - ✓ Quand passer en agentique du tout
 - ✓ Ramification de sous-agents
 - ✓ Longueur des réponses
 - ✓ Mesurer avant de supposer
-

Petit point responsabilité



Ce qui compte pour moi :
l'usage en connaissance de cause

Merci

Tristan Coignon · tristan.coignon@inria.fr · tristan.coignon.fr

EcoLogits

l'énergie de chaque
appel d'API LLM
en Python

CodeCarbon

le CO₂ de tout
processus Python,
matériel inclus

Neuralwatt

un fournisseur qui
facture l'énergie
mesurée en kWh

Diapositives

cette présentation,
en ligne

Des requêtes comparables varient d'un facteur 70 en énergie mesurée

0,43 Wh requête courte GPT-4o vs
33+ Wh prompt long o3 / DeepSeek-R1

Une requête de texte coûte 0,24 à 4 Wh, dominée par la réponse

0,24 Wh

prompt Gemini médian, mesuré en production · jusqu'à 3,96 Wh pour un « write an email » non optimisé (calculateur EcoLogits)

Prompt (jetons d'entrée)



≈ 10 % de l'énergie

Réponse (jetons de sortie)



≈ 90 % de l'énergie

Ce que mesurent les barres : la part de l'énergie d'une requête, prompt vs réponse · proportions schématiques (non mesurées) : la réponse domine