La consommation d’un modèle

Pour chaque requête, la consommation d’IA en ressources dépend de ces trois principaux facteurs :

Taille du modèle

En milliards de paramètres

Longueur du texte

En jetons

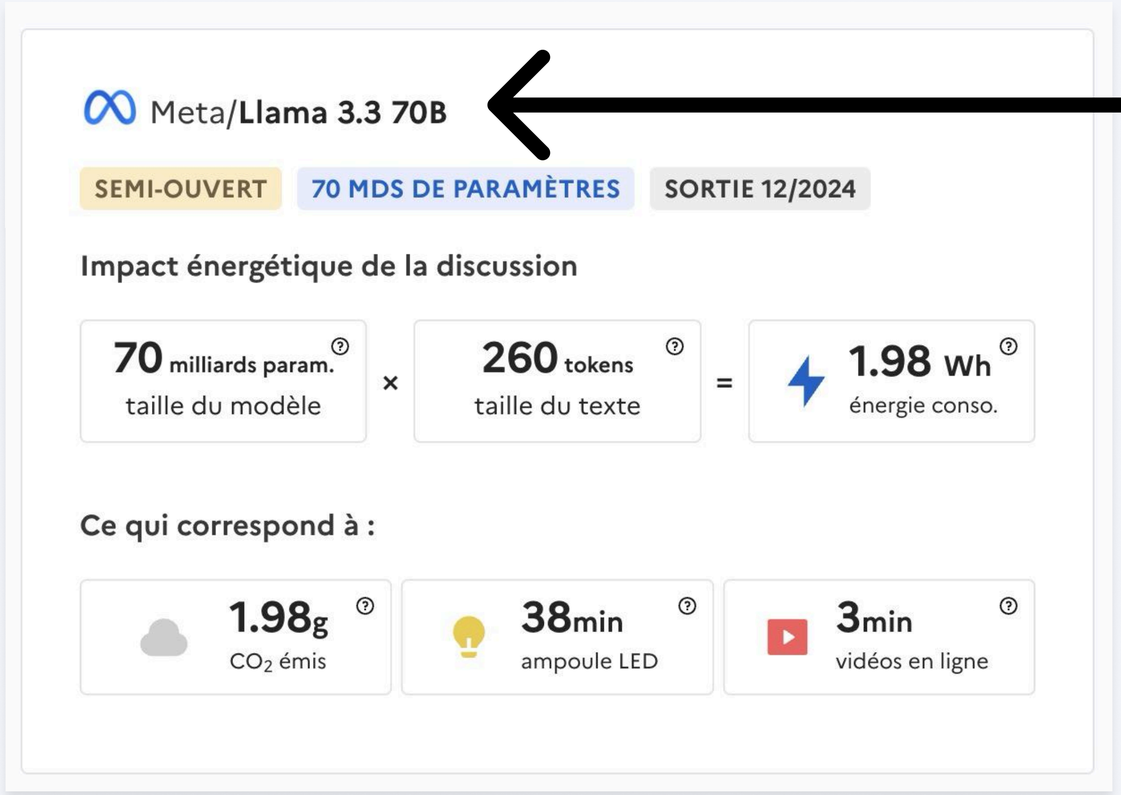
Architecture

Par exemple MoE ou MatFormer

Taille du modèle - facteurs de consommation

Qu'est-ce que la "taille" d'un modèle ?

Les duels de l'IA
par compar:IA



Quantité de paramètres

70B = 70 milliards de paramètres

Meta/**Llama 3.3 70B**

Un paramètre est comme une "synapse" artificiel. Plus il y a de paramètres, plus (en théorie) un modèle a de capacité d'apprendre. Néanmoins, ça ne veut pas dire qu'un modèle plus grand sera toujours plus "intelligent".

Taille du modèle - facteurs de consommation

Qu'est-ce que la "taille" d'un modèle ?

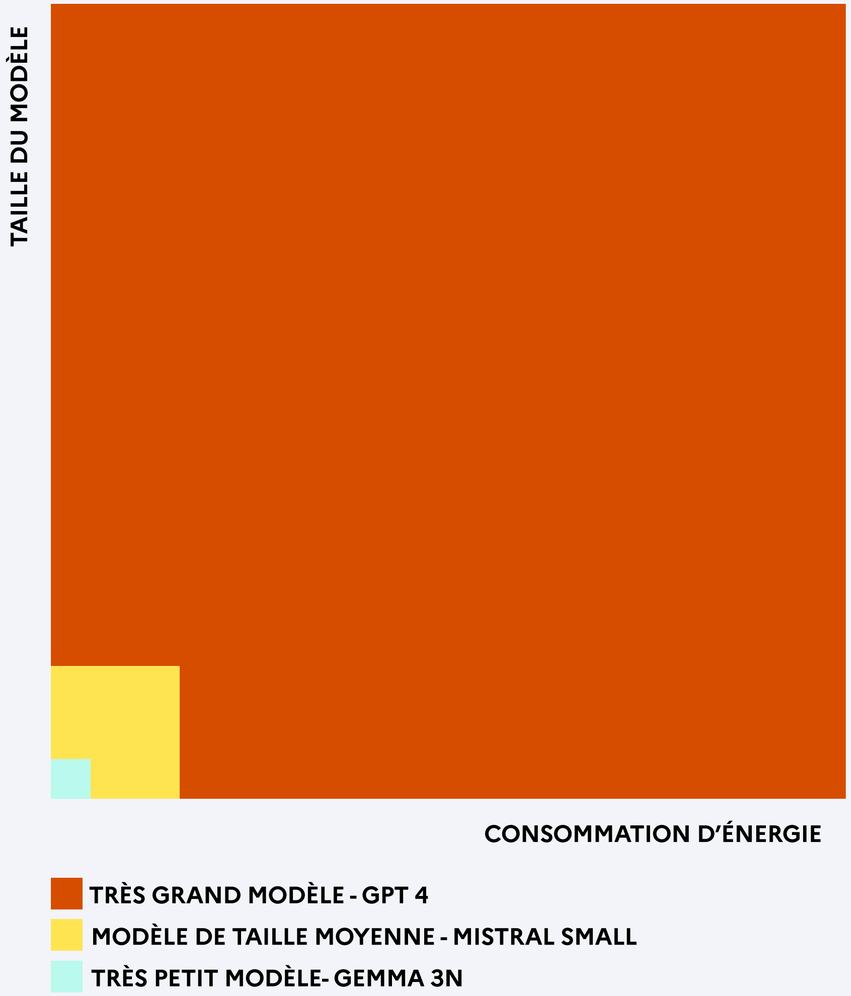
Les duels de l'IA
par compar:IA

Anthropic/**Claude 3.5 Sonnet**	→	**300** milliards de paramètres (env)	→ **XL**
Meta/**Llama 3.3 70B**	→	**70** milliards de paramètres	→ **L**
Mistral/**Mistral Small**	→	**24** milliards de paramètres	→ **M**
Google/**Gemma 3 4b**	→	**4** milliards de paramètres	→ **XS**

Taille du modèle - facteurs de consommation

Influence de la "taille"

Les duels de l'IA
par compar:IA



Un paramètre est comme une "synapse" artificielle. Plus il y a de paramètres, plus (en théorie) un modèle a de capacité d'apprendre. Néanmoins, ça ne veut pas dire qu'un modèle plus grand sera toujours plus "intelligent".

Par contre, **la taille a une grande influence sur la puissance de calcul nécessaire pour faire fonctionner le modèle et donc sur la consommation énergétique.**

Graphique - Conseil National du Numérique :
"20 cartes pour aborder l'impact énergétique de l'IA générative"

conseil national du numérique

Longueur du texte - facteurs de consommation

Les “jetons”

Les duels de l’IA
par compar:IA

La longueur du texte produit par un modèle de langage se quantifie en "jetons" (tokens).

Un jeton constitue une unité de base qui peut englober un ou plusieurs caractères, selon leur fréquence d'apparition ensemble dans la langue.

Par exemple :

Salut

1 JETON

Salut, ça va ?

5 JETONS

La jetonisation sert à :

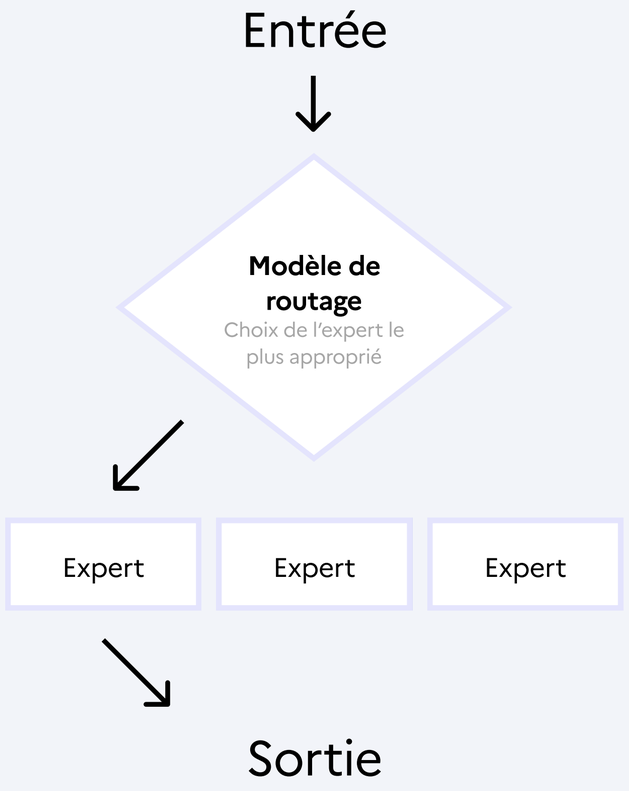
- Découper le texte en morceaux que l'IA peut comprendre
- Compter combien de "mots" l'IA peut traiter d'un coup
- Calculer le coût d'utilisation du modèle

Architectures - facteurs de consommation

Les “mélanges d’experts”

Les duels de l’IA
par compar:IA

Les modèles “mélanges d’experts” (Mixture of Experts) contiennent plusieurs experts, mais un seul est activé à chaque jeton.



Avantages

- Moins de puissance de calcul nécessaire
  - DONC consomme moins d’électricité
  - DONC coûte moins cher à utiliser

Désavantages

- Prend beaucoup de place dans la mémoire de la machine
- Plus complexe à développer
- Peut avoir des problèmes de généralisation

Exemples

Mistral/**Mistral 8x7B**

DeepSeek/**DeepSeek V3**

Meta/**Llama 4 Scout**

Les modèles de raisonnement

Les duels de l'IA
par compar:IA

Les modèles de raisonnement simulent une réflexion avant de donner la réponse définitive.

Parce que les modèles de raisonnement génèrent au total plus de jetons, les modèles de raisonnement consomment plus d'énergie à architecture et nombre de paramètres égal.

Modèle de raisonnement

Raisonnement

Okay, I need to help the user reformulate their sentence. The original sentence is: "Pour chaque requête, la consommation d'IA en ressources dépend de ces trois principaux facteurs." The user wants a better version of this.

First, I should understand the context. It's about AI resource consumption per query, influenced by three main factors. The user probably wants a more polished or professional tone.

Modèle "classique"

Voici quelques reformulations possibles, plus claires et plus élégantes, de la phrase "Pour chaque requête, la consommation d'IA en ressources dépend de ces trois principaux facteurs" :

Options privilégiées :

- **"La consommation de ressources de l'IA par requête est influencée par trois facteurs clés."** (Simple, direct et professionnel)
- **"Chaque requête entraîne une consommation de ressources de l'IA qui est déterminée par ces trois principaux facteurs."** (Plus formel et précis)

Impact environnemental : CO_2 et ressources hydriques

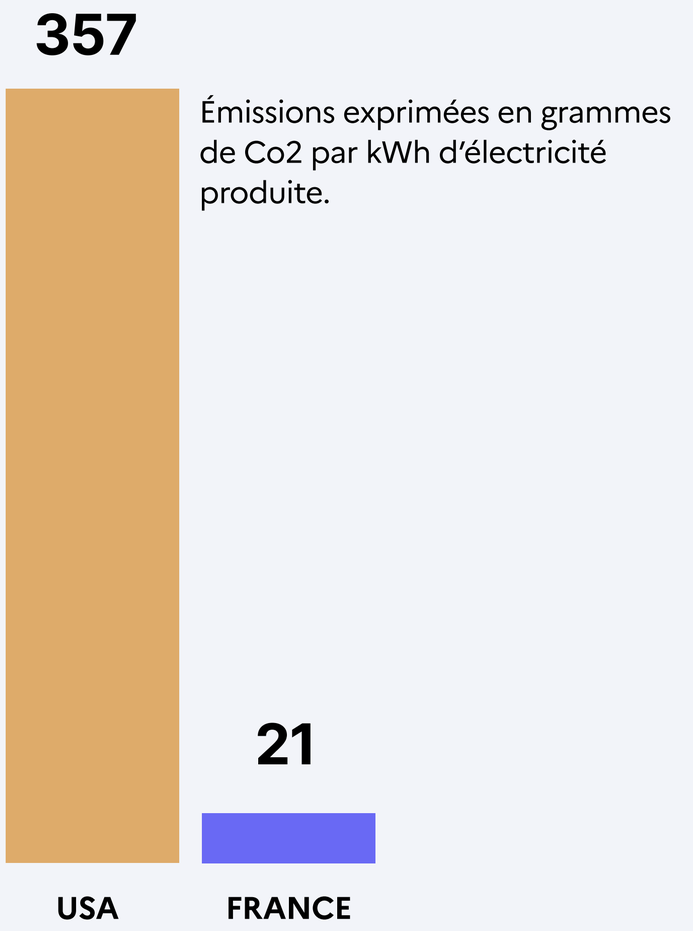
Localisation des serveurs

Pour évaluer l'impact carbone direct des centres de données, il est essentiel de considérer **l'intensité carbone du mix énergétique local** où ces infrastructures sont implantées, indépendamment de la localisation géographique des utilisateurs finaux.

Actuellement, les principaux acteurs du secteur étant des entreprises américaines, leurs centres de données se concentrent majoritairement sur le **territoire américain**.

Dans ce contexte, **la France présente un avantage concurrentiel significatif** pour l'implantation de futures infrastructures d'intelligence artificielle, grâce à son mix électrique largement décarboné reposant sur l'énergie nucléaire.

Les duels de l'IA
par compar:IA



Impact environnemental : CO_2 et ressources hydriques

Localisation des serveurs

À Santiago, au Chili, les géants technologiques américains construisent massivement des centres de données qui consomment des milliards de litres d'eau par an, aggravant la sécheresse.

Ce choix d'implantation illustre l'importance stratégique de la localisation des centres de données, les entreprises privilégiant les infrastructures fiables et la réglementation favorable du Chili au détriment des ressources hydriques locales.

La consommation de ressources ne doit donc pas être évaluée uniquement à l'échelle des ressources globales disponibles, mais mise en perspective avec les contraintes énergétiques et hydriques spécifiques de chaque région d'implantation.