Coding agents

Quels enjeux?

David Salinas. 20/01/2025. Grenoble Meetup.

Coding agent

On the rise

Coding agent

On the rise

- Les coding agents se sont **dramatiquement** améliorés

Coding agent

On the rise

- Les coding agents se sont **dramatiquement** améliorés

I have a checkpoint into "checkpoint/", it is in litgpt format.

I want you to convert it to HuggingFace and write a script convert_script.py.

The model used is described in checkpoint/model_config.yaml, it is a Qwen model with ~50 parameters.

Please write code with the following interface:

```python
def load_litgpt(path):
    ...

def convert_to_huggingface(path):
    ...

def do_inference_litgpt(context: list[int], litgpt_model):
    ... # returns logits or string


def do_inference_huggingface(context: list[int], litgpt_model):
    ... # returns logits or string

# make sure that the following codes run iterates with "uv run python convert_script.py" until it does
litgpt_model = load_litgpt("checkpoint/")
hf_model = convert_to_huggingface("checkpoint/")

random_context = ... # todo draw random context

# makes sure that predictions matches, something like this:
assert do_inference_litgpt(context, litgpt_model) == do_inference_huggingface(context, hf_model)
```

This example worked in one shot for instance (!)

Coding agent

On the rise

- Les coding agents se sont **dramatiquement** améliorés
- Leur utilisation devient une part croissante de l’utilisation de LLMs

I have a checkpoint into "checkpoint/", it is in litgpt format.

I want you to convert it to HuggingFace and write a script convert_script.py.

The model used is described in checkpoint/model_config.yaml, it is a Qwen model with ~50 parameters.

Please write code with the following interface:

```
def load_litgpt(path):
    ...

def convert_to_huggingface(path):
    ...

def do_inference_litgpt(context: list[int], litgpt_model):
    ... # returns logits or string


def do_inference_huggingface(context: list[int], litgpt_model):
    ... # returns logits or string

# make sure that the following codes run iterates with "uv run python
convert_script.py" until it does
litgpt_model = load_litgpt("checkpoint/")
hf_model = convert_to_huggingface("checkpoint/")

random_context = ... # todo draw random context

# makes sure that predictions matches, something like this:
assert do_inference_litgpt(context, litgpt_model) ==
do_inference_huggingface(context, hf_model)
```

This example worked in one shot for instance (!)

Coding agent

On the rise

- Les coding agents se sont **dramatiquement** améliorés
- Leur utilisation devient une part croissante de l'utilisation de LLMs

I have a checkpoint into "checkpoint/", it is in litgpt format.

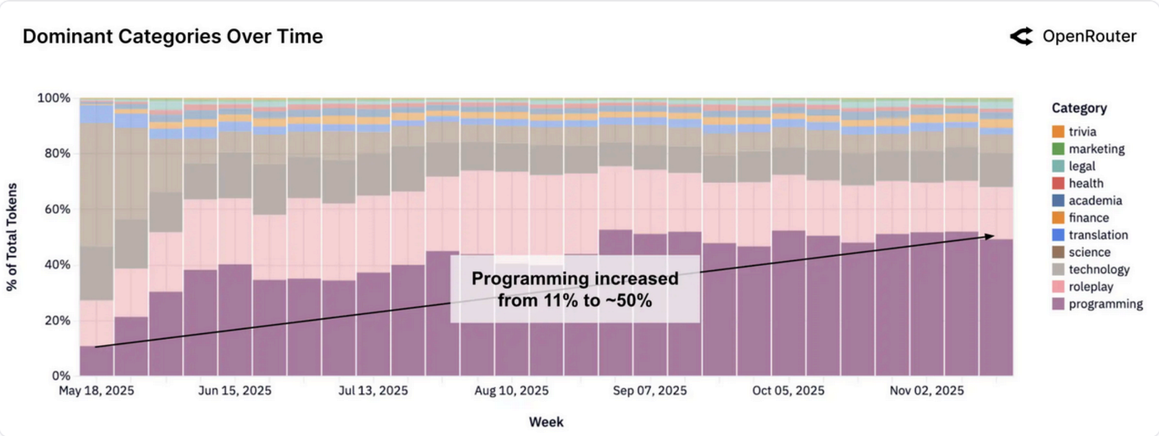
I want you to convert it to HuggingFace and write a script convert_script.py.

The model used is described in checkpoint/model_config.yaml, it is a Qwen model with ~50 parameters.

Please write code with the following interface:

```python
def load_litgpt(path):
    ...

def convert_to_huggingface(path):
    ...

def do_inference_litgpt(context: list[int], litgpt_model):
    ... # returns logits or string


def do_inference_huggingface(context: list[int], litgpt_model):
    ... # returns logits or string

# make sure that the following codes run iterates with "uv run python convert_script.py" until it does
litgpt_model = load_litgpt("checkpoint/")
hf_model = convert_to_huggingface("checkpoint/")

random_context = ... # todo draw random context

# makes sure that predictions matches, something like this:
assert do_inference_litgpt(context, litgpt_model) == do_inference_huggingface(context, hf_model)
```

This example worked in one shot for instance (!)



Programming as a dominant and growing category. The share of all LLM queries classified under programming has increased steadily, reflecting the rise of AI-assisted development workflows.

https://openrouter.ai/state-of-ai

Coding agent

On the rise

- Les coding agents se sont **dramatiquement** améliorés
- Leur utilisation devient une part croissante de l'utilisation de LLMs
- Aussi potentiellement très profitable:

I have a checkpoint into "checkpoint/", it is in litgpt format.

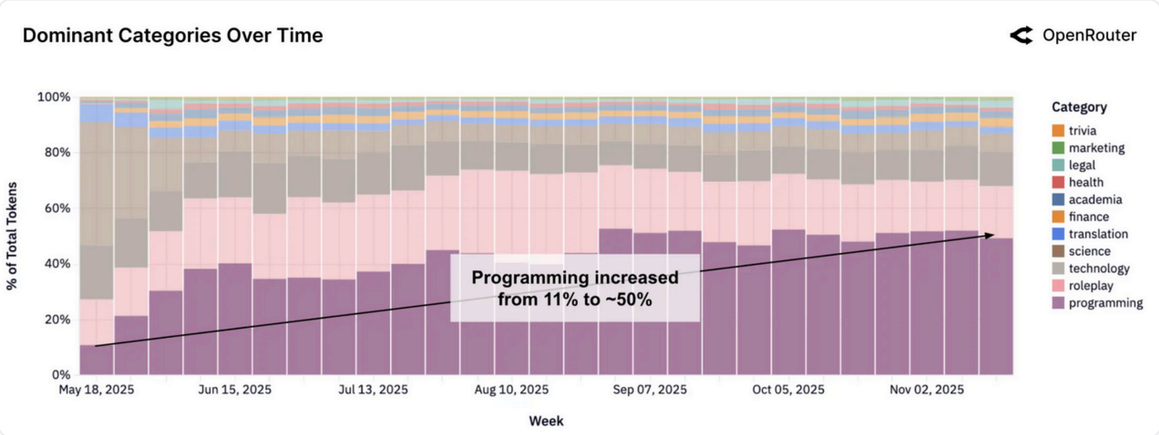
I want you to convert it to HuggingFace and write a script convert_script.py.

The model used is described in checkpoint/model_config.yaml, it is a Qwen model with ~50 parameters.

Please write code with the following interface:

```python
def load_litgpt(path):
    ...

def convert_to_huggingface(path):
    ...

def do_inference_litgpt(context: list[int], litgpt_model):
    ... # returns logits or string


def do_inference_huggingface(context: list[int], litgpt_model):
    ... # returns logits or string

# make sure that the following codes run iterates with "uv run python convert_script.py" until it does
litgpt_model = load_litgpt("checkpoint/")
hf_model = convert_to_huggingface("checkpoint/")

random_context = ... # todo draw random context

# makes sure that predictions matches, something like this:
assert do_inference_litgpt(context, litgpt_model) ==
do_inference_huggingface(context, hf_model)
```

This example worked in one shot for instance (!)



Programming as a dominant and growing category. The share of all LLM queries classified under programming has increased steadily, reflecting the rise of AI-assisted development workflows.

https://openrouter.ai/state-of-ai

Coding agent

On the rise

- Les coding agents se sont **dramatiquement** améliorés
- Leur utilisation devient une part croissante de l'utilisation de LLMs
- Aussi potentiellement très profitable:
  - €200/mois pour un abonnement Claude max

I have a checkpoint into "checkpoint/", it is in litgpt format.

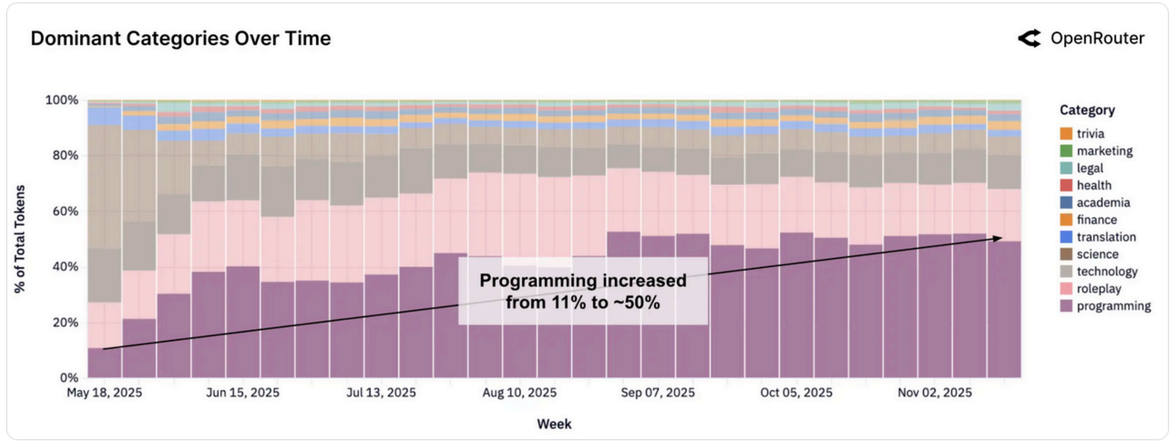
I want you to convert it to HuggingFace and write a script convert_script.py.

The model used is described in checkpoint/model_config.yaml, it is a Qwen model with ~50 parameters.

Please write code with the following interface:

```python
def load_litgpt(path):
    ...

def convert_to_huggingface(path):
    ...

def do_inference_litgpt(context: list[int], litgpt_model):
    ... # returns logits or string


def do_inference_huggingface(context: list[int], litgpt_model):
    ... # returns logits or string

# make sure that the following codes run iterates with "uv run python convert_script.py" until it does
litgpt_model = load_litgpt("checkpoint/")
hf_model = convert_to_huggingface("checkpoint/")

random_context = ... # todo draw random context

# makes sure that predictions matches, something like this:
assert do_inference_litgpt(context, litgpt_model) == do_inference_huggingface(context, hf_model)
```

This example worked in one shot for instance (!)



Programming as a dominant and growing category. The share of all LLM queries classified under programming has increased steadily, reflecting the rise of AI-assisted development workflows.

https://openrouter.ai/state-of-ai

Enjeux des coding agents

- Security
- Open-source vs propriétaire
- Evaluation
- Souveraineté

Security

Security

- Exécuter un coding agent nécessite de régler de nombreuses permissions :

Security

- Exécuter un coding agent nécessite de régler de nombreuses permissions :
  - Permission *de lire des fichiers* (pour lire le code)

Security

- Exécuter un coding agent nécessite de régler de nombreuses permissions :
  - Permission *de lire des fichiers* (pour lire le code)
  - Permission *d'écrire des fichiers* (pour écrire du code, des configurations, …)

Security

- Exécuter un coding agent nécessite de régler de nombreuses permissions :
  - Permission *de lire des fichiers* (pour lire le code)
  - Permission *d'écrire des fichiers* (pour écrire du code, des configurations, …)
  - Permission *d'exécuter des processus* (exécuter Python, effectuer des opérations bash)

Security

- Exécuter un coding agent nécessite de régler de nombreuses permissions :
  - Permission *de lire des fichiers* (pour lire le code)
  - Permission *d'écrire des fichiers* (pour écrire du code, des configurations, …)
  - Permission *d'exécuter des processus* (exécuter Python, effectuer des opérations bash)
  - Permission *d'exécuter du code* provenant de serveurs MCP

Security

- Exécuter un coding agent nécessite de régler de nombreuses permissions :
  - Permission *de lire des fichiers* (pour lire le code)
  - Permission *d'écrire des fichiers* (pour écrire du code, des configurations, …)
  - Permission *d'exécuter des processus* (exécuter Python, effectuer des opérations bash)
  - Permission *d'exécuter du code* provenant de serveurs MCP

Question: Quelqu'un sait-il ce qu'est le *YOLO* mode? 🤔

Security

- Exécuter un coding agent nécessite de régler de nombreuses permissions :
  - Permission *de lire des fichiers* (pour lire le code)
  - Permission *d'écrire des fichiers* (pour écrire du code, des configurations, …)
  - Permission *d'exécuter des processus* (exécuter Python, effectuer des opérations bash)
  - Permission *d'exécuter du code* provenant de serveurs MCP
- YOLO = You only live once (on ne vit qu'une fois)

Question: Quelqu'un sait-il ce qu'est le *YOLO* mode? 🤔

Security

- Exécuter un coding agent nécessite de régler de nombreuses permissions :
  - Permission *de lire des fichiers* (pour lire le code)
  - Permission *d'écrire des fichiers* (pour écrire du code, des configurations, …)
  - Permission *d'exécuter des processus* (exécuter Python, effectuer des opérations bash)
  - Permission *d'exécuter du code* provenant de serveurs MCP
- YOLO = You only live once (on ne vit qu'une fois)
  - => Accorder toutes les permissions et lancer l'exécution

Question: Quelqu'un sait-il ce qu'est le *YOLO* mode? 🤔

Security

- Exécuter un coding agent nécessite de régler de nombreuses permissions :
  - Permission *de lire des fichiers* (pour lire le code)
  - Permission *d'écrire des fichiers* (pour écrire du code, des configurations, …)
  - Permission *d'exécuter des processus* (exécuter Python, effectuer des opérations bash)
  - Permission *d'exécuter du code* provenant de serveurs MCP
- YOLO = You only live once (on ne vit qu'une fois)
  - => Accorder toutes les permissions et lancer l'exécution
  - Mode très répandu et risque de sécurité *énorme,* y compris avec les modèles open-weights

Question: Quelqu'un sait-il ce qu'est le *YOLO* mode? 🤔

Security

What to do

Security

What to do

- Option 1 : Utiliser un bac à sable (sandbox), par exemple Claude Code peut être utilisé dans un navigateur dans un environnement isolé

Security

What to do

- Option 1 : Utiliser un bac à sable (sandbox), par exemple Claude Code peut être utilisé dans un navigateur dans un environnement isolé
- Option 2 : Bien comprendre les risques, appliquer toutes les précautions recommandées

Security

What to do

- Option 1 : Utiliser un bac à sable (sandbox), par exemple Claude Code peut être utilisé dans un navigateur dans un environnement isolé
- Option 2 : Bien comprendre les risques, appliquer toutes les précautions recommandées
  - N'activer que les serveurs MCP de confiance, n'autoriser l'accès qu'à des fichiers spécifiques, …

Security

What to do

- Option 1 : Utiliser un bac à sable (sandbox), par exemple Claude Code peut être utilisé dans un navigateur dans un environnement isolé
- Option 2 : Bien comprendre les risques, appliquer toutes les précautions recommandées
  - N'activer que les serveurs MCP de confiance, n'autoriser l'accès qu'à des fichiers spécifiques, …
  - Même définir les permissions manuellement est probablement une mauvaise idée, de nombreuses erreurs sont commises, de nombreux systèmes sont compromis

Security

What to do

- Option 1 : Utiliser un bac à sable (sandbox), par exemple Claude Code peut être utilisé dans un navigateur dans un environnement isolé
- Option 2 : Bien comprendre les risques, appliquer toutes les précautions recommandées
  - N'activer que les serveurs MCP de confiance, n'autoriser l'accès qu'à des fichiers spécifiques, …
  - Même définir les permissions manuellement est probablement une mauvaise idée, de nombreuses erreurs sont commises, de nombreux systèmes sont compromis
- Development de systèmes déployés “on-prem”

Security

What to do

- Option 1 : Utiliser un bac à sable (sandbox), par exemple Claude Code peut être utilisé dans un navigateur dans un environnement isolé
- Option 2 : Bien comprendre les risques, appliquer toutes les précautions recommandées
  - N'activer que les serveurs MCP de confiance, n'autoriser l'accès qu'à des fichiers spécifiques, …
  - Même définir les permissions manuellement est probablement une mauvaise idée, de nombreuses erreurs sont commises, de nombreux systèmes sont compromis
- Development de systèmes déployés “on-prem”
  - Mistral, Cohere, Poolside

Security

What to do

- Option 1 : Utiliser un bac à sable (sandbox), par exemple Claude Code peut être utilisé dans un navigateur dans un environnement isolé
- Option 2 : Bien comprendre les risques, appliquer toutes les précautions recommandées
  - N'activer que les serveurs MCP de confiance, n'autoriser l'accès qu'à des fichiers spécifiques, …
  - Même définir les permissions manuellement est probablement une mauvaise idée, de nombreuses erreurs sont commises, de nombreux systèmes sont compromis
- Development de systèmes déployés “on-prem”
  - Mistral, Cohere, Poolside
  - Restriction d’accès à internet, meilleure sécurité

Security

What to do

- Option 1 : Utiliser un bac à sable (sandbox), par exemple Claude Code peut être utilisé dans un navigateur dans un environnement isolé
- Option 2 : Bien comprendre les risques, appliquer toutes les précautions recommandées
  - N'activer que les serveurs MCP de confiance, n'autoriser l'accès qu'à des fichiers spécifiques, …
  - Même définir les permissions manuellement est probablement une mauvaise idée, de nombreuses erreurs sont commises, de nombreux systèmes sont compromis
- Development de systèmes déployés “on-prem”
  - Mistral, Cohere, Poolside
  - Restriction d’accès à internet, meilleure sécurité
  - => Gros marché

Open-source vs proprietary

Open-source vs proprietary

- Pour être clair:

Open-source vs proprietary

- Pour être clair:
  - *Propriétaire*: modèle disponible à travers une API

Open-source vs proprietary

- Pour être clair:
  - *Propriétaire*: modèle disponible à travers une API
    - Example: GPT-4, GPT-4, Claude-Opus, Gemini, …

Open-source vs proprietary

- Pour être clair:
  - *Propriétaire*: modèle disponible à travers une API
    - Example: GPT-4, GPT-4, Claude-Opus, Gemini, …
  - *Open-weight*: modele dont les paramètres sont disponible au téléchargement

Open-source vs proprietary

- Pour être clair:
  - *Propriétaire*: modèle disponible à travers une API
    - Example: GPT-4, GPT-4, Claude-Opus, Gemini, …
  - *Open-weight*: modele dont les paramètres sont disponible au téléchargement
    - Example: modèles Llama, Mistral, Deepseek, Qwen, GPT-OSS-120B

Open-source vs proprietary

- Pour être clair:
  - *Propriétaire*: modèle disponible à travers une API
    - Example: GPT-4, GPT-4, Claude-Opus, Gemini, …
  - *Open-weight*: modele dont les paramètres sont disponible au téléchargement
    - Example: modèles Llama, Mistral, Deepseek, Qwen, GPT-OSS-120B
  - Open-Source: modèles dont le code et les données sont disponibles en plus des paramètres

Open-source vs proprietary

- Pour être clair:
  - *Propriétaire*: modèle disponible à travers une API
    - Example: GPT-4, GPT-4, Claude-Opus, Gemini, …
  - *Open-weight*: modele dont les paramètres sont disponible au téléchargement
    - Example: modèles Llama, Mistral, Deepseek, Qwen, GPT-OSS-120B
  - Open-Source: modèles dont le code et les données sont disponibles en plus des paramètres
    - Example:

Open-source vs proprietary

- Pour être clair:
  - *Propriétaire*: modèle disponible à travers une API
    - Example: GPT-4, GPT-4, Claude-Opus, Gemini, …
  - *Open-weight*: modele dont les paramètres sont disponible au téléchargement
    - Example: modèles Llama, Mistral, Deepseek, Qwen, GPT-OSS-120B
  - Open-Source: modèles dont le code et les données sont disponibles en plus des paramètres
    - Example:
      - Base: GPT-Neo-x, Bloom, MPT, Pythia, Falcon, Polythias, OLMo2, …

Open-source vs proprietary

- Pour être clair:
  - *Propriétaire*: modèle disponible à travers une API
    - Example: GPT-4, GPT-4, Claude-Opus, Gemini, …
  - *Open-weight*: modele dont les paramètres sont disponible au téléchargement
    - Example: modèles Llama, Mistral, Deepseek, Qwen, GPT-OSS-120B
  - Open-Source: modèles dont le code et les données sont disponibles en plus des paramètres
    - Example:
      - Base: GPT-Neo-x, Bloom, MPT, Pythia, Falcon, Polythias, OLMo2, …
      - Instruction-tuned: SmolLM, StarCoder, Olmo3, …

Open-source vs proprietary

Un modèle pour tous n’est pas une bonne solution

> Building out **17 gigawatts** of capacity would require the equivalent of about **17 nuclear power** plants, each of which takes at least a decade to build. The OpenAI team says talks are underway with hundreds of infrastructure providers across North America, but there are no firm answers yet.
>
> The U.S. grid is already strained, gas turbines are sold out through 2028, nuclear is slow to deploy, and renewables are tied up in political roadblocks.
>
> “I am extremely bullish about nuclear, advanced fission, fusion,” Altman said. “We should build more ... a lot more of the current generation of fission plants, given the needs for dense, dense energy.”
>
> **Source:** https://www.cnbc.com/2025/09/26/openai-big-week-ai-arms-race.html

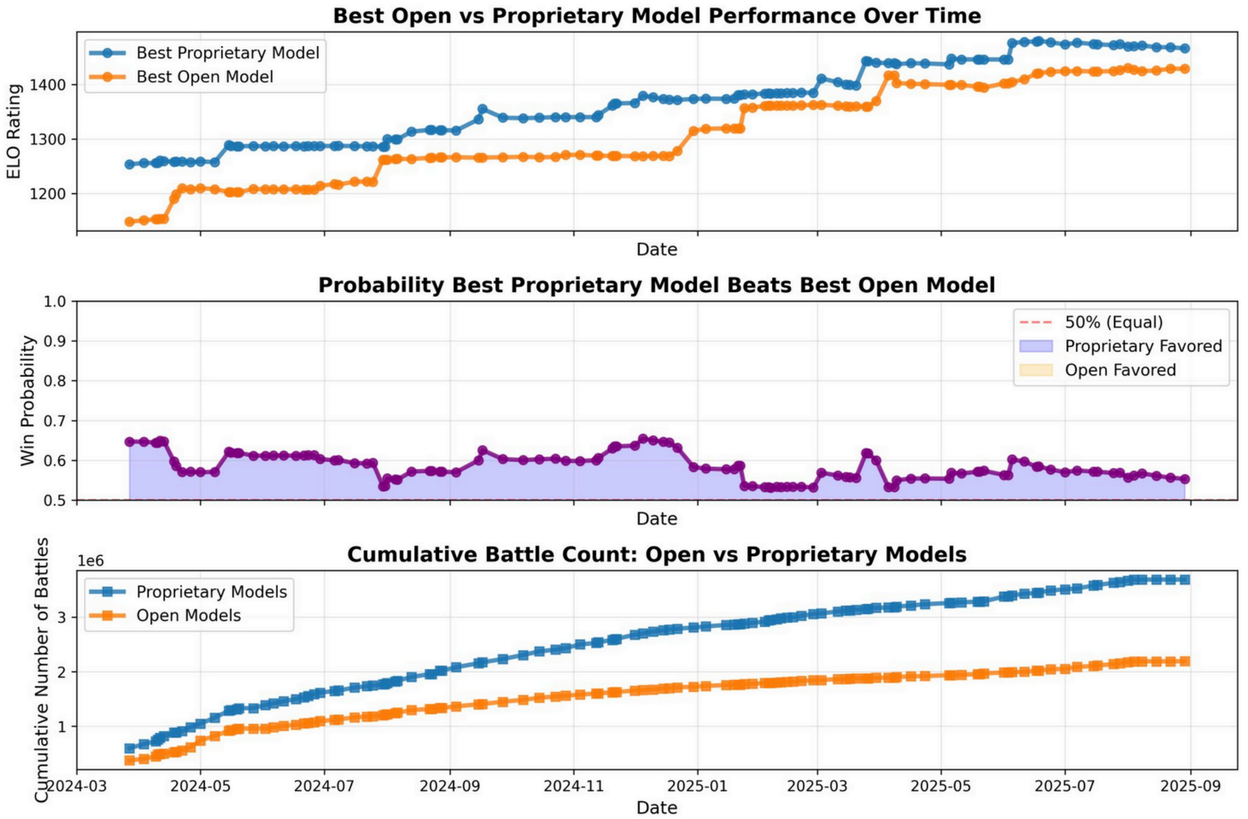
- Entrainer un modèle par entreprise serait un désastre
  - écologique
  - sociétal (concentration des pouvoirs)

Open-source vs proprietary

Quel écart?

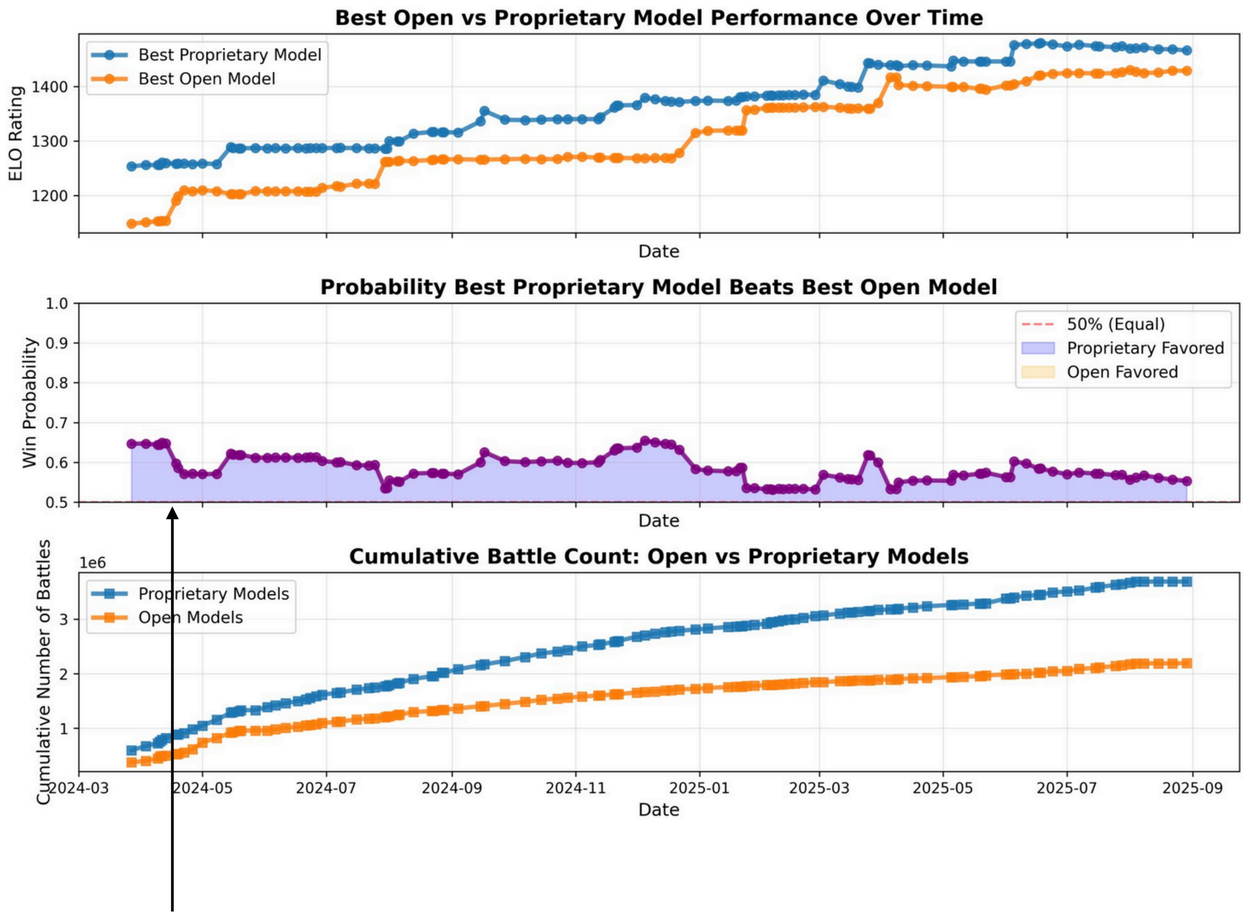
Open-source vs proprietary

Quel écart?



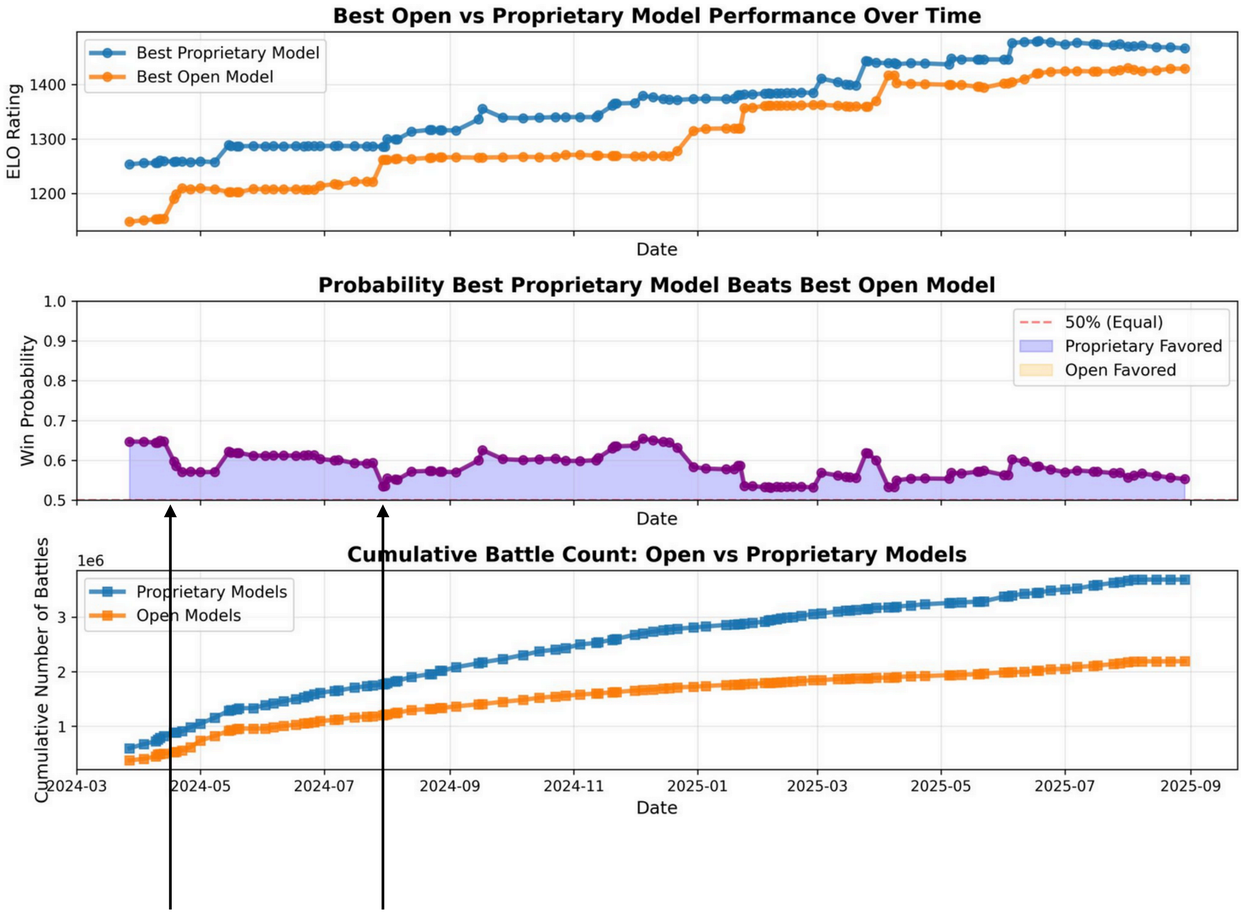
Open-source vs proprietary

Quel écart?



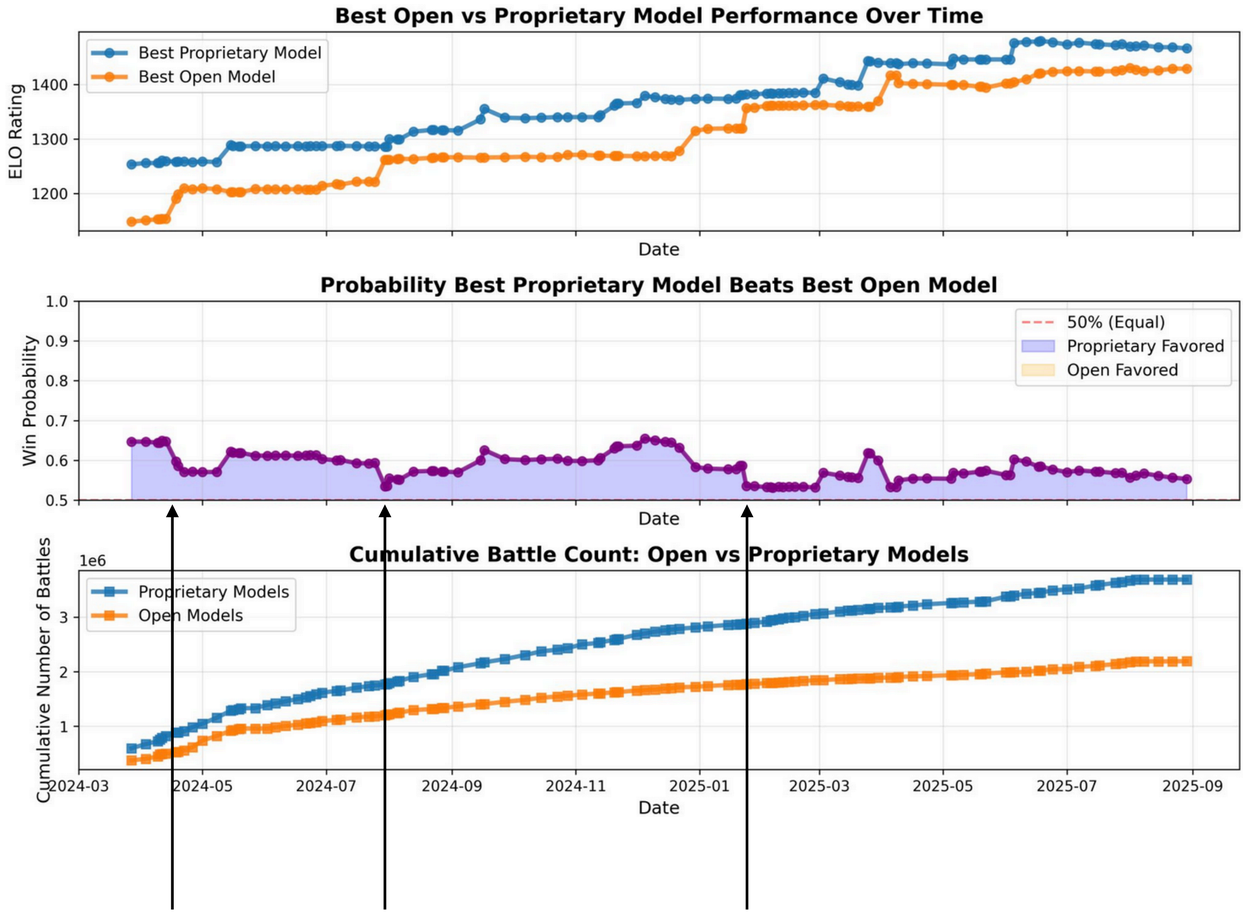
Open-source vs proprietary

Quel écart?



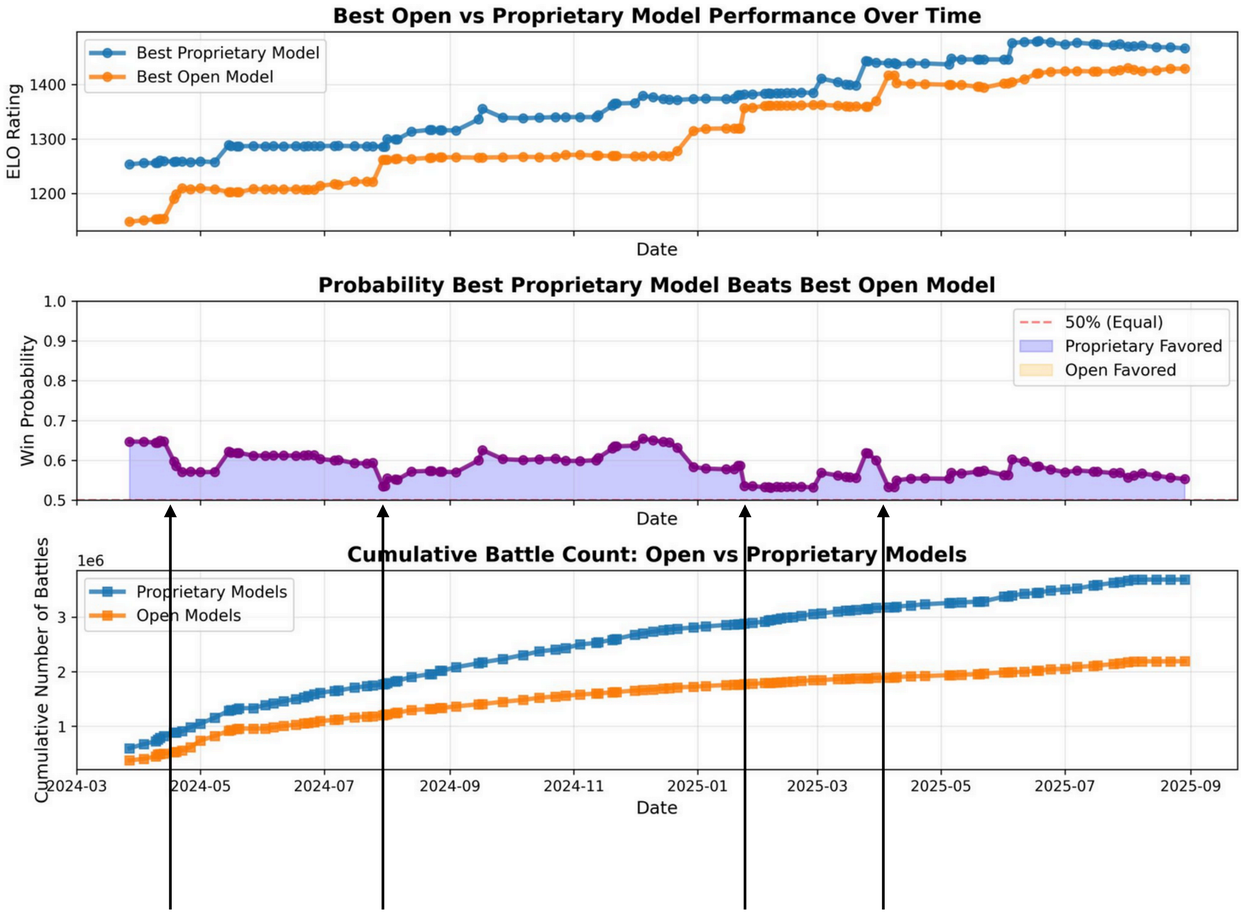
Open-source vs proprietary

Quel écart?



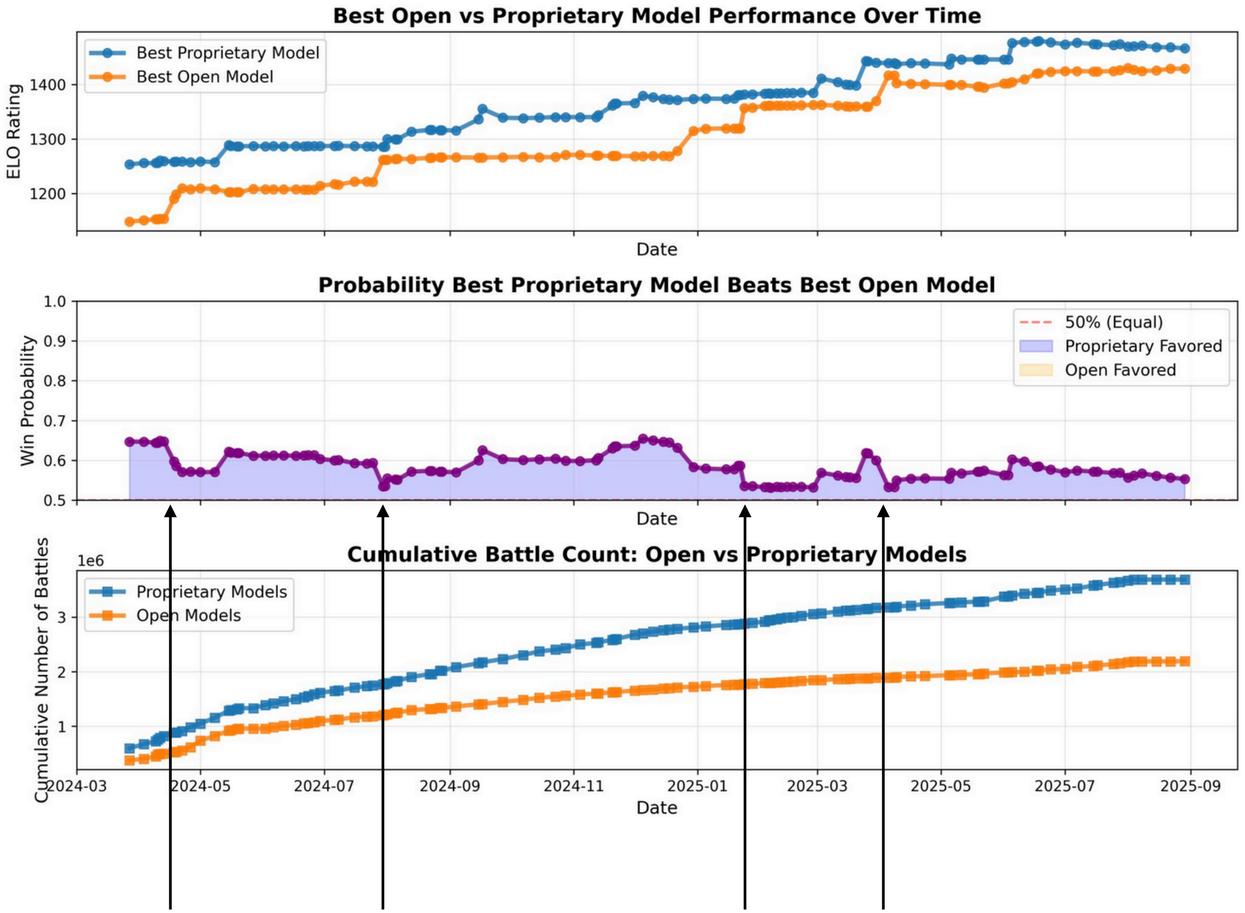
Open-source vs proprietary

Quel écart?



Open-source vs proprietary

Quel écart?



Savez vous quels modèles ouverts ont réduit l’écart? 🤔

Open-source vs proprietary

Quel raisons pour l’écart?

Open-source vs proprietary

Quel raisons pour l’écart?

- Ecart de moyens:

Open-source vs proprietary

Quel raisons pour l’écart?

- Ecart de moyens:
  - Humains:

Open-source vs proprietary

Quel raisons pour l’écart?

- Ecart de moyens:
  - Humains:
    - Salaire & taille d’équipes

Open-source vs proprietary

Quel raisons pour l’écart?

- Ecart de moyens:
  - Humains:
    - Salaire & taille d’équipes
  - GPUs:

Open-source vs proprietary

Quel raisons pour l’écart?

- Ecart de moyens:
  - Humains:
    - Salaire & taille d’équipes
  - GPUs:
    - Cluster gigantesques pour OpenAI, XAI, Google, …

Open-source vs proprietary

Quel raisons pour l’écart?

- Ecart de moyens:
  - Humains:
    - Salaire & taille d’équipes
  - GPUs:
    - Cluster gigantesques pour OpenAI, XAI, Google, …
    - Mais … clusters Européens significatifs dédiés à la recherche et aux startups 🤞

Open-source vs proprietary

Quel raisons pour l’écart?

- Ecart de moyens:
  - Humains:
    - Salaire & taille d’équipes
  - GPUs:
    - Cluster gigantesques pour OpenAI, XAI, Google, …
    - Mais … clusters Européens significatifs dédiés à la recherche et aux startups 🤞

Supercomputer	Country	Number of GPUs	GPU Type(s)
LUMI	Finland	~11,912	AMD MI250X
MareNostrum 5 (BSC)	Spain	4,480	NVIDIA H100 (Hopper)
JUPITER	Germany	~24,000	NVIDIA H100 (in GH200 Grace Hopper Superchips)
Leonardo	Italy	13,824	NVIDIA A100
Jean Zay	France	3,704	Mixed: 1,456 NVIDIA H100 + 416 NVIDIA A100 + 1,832 NVIDIA V100
Alps	Switzerland	10,752	NVIDIA H100 (in GH200 Grace Hopper Superchips)

European clusters 🇪🇺

Open-source vs proprietary

Quel raisons pour l’écart?

- Ecart de moyens:
  - Humains:
    - Salaire & taille d’équipes
  - GPUs:
    - Cluster gigantesques pour OpenAI, XAI, Google, …
    - Mais … clusters Européens significatifs dédiés à la recherche et aux startups 🤞
    - A relativiser avec XAI par exemple: 200,000 H100 reporté pour Colossus (!)

Supercomputer	Country	Number of GPUs	GPU Type(s)
LUMI	Finland	~11,912	AMD MI250X
MareNostrum 5 (BSC)	Spain	4,480	NVIDIA H100 (Hopper)
JUPITER	Germany	~24,000	NVIDIA H100 (in GH200 Grace Hopper Superchips)
Leonardo	Italy	13,824	NVIDIA A100
Jean Zay	France	3,704	Mixed: 1,456 NVIDIA H100 + 416 NVIDIA A100 + 1,832 NVIDIA V100
Alps	Switzerland	10,752	NVIDIA H100 (in GH200 Grace Hopper Superchips)

European clusters 🇪🇺

Open-source vs proprietary

Quel raisons pour l’écart?

- Ecart de moyens:
  - Humains:
    - Salaire & taille d’équipes
  - GPUs:
    - Cluster gigantesques pour OpenAI, XAI, Google, …
    - Mais … clusters Européens significatifs dédiés à la recherche et aux startups 🤞
    - A relativiser avec XAI par exemple: 200,000 H100 reporté pour Colossus (!)
- Ecart de règles:

Supercomputer	Country	Number of GPUs	GPU Type(s)
LUMI	Finland	~11,912	AMD MI250X
MareNostrum 5 (BSC)	Spain	4,480	NVIDIA H100 (Hopper)
JUPITER	Germany	~24,000	NVIDIA H100 (in GH200 Grace Hopper Superchips)
Leonardo	Italy	13,824	NVIDIA A100
Jean Zay	France	3,704	Mixed: 1,456 NVIDIA H100 + 416 NVIDIA A100 + 1,832 NVIDIA V100
Alps	Switzerland	10,752	NVIDIA H100 (in GH200 Grace Hopper Superchips)

European clusters 🇪🇺

Open-source vs proprietary

Quel raisons pour l’écart?

- Ecart de moyens:
  - Humains:
    - Salaire & taille d’équipes
  - GPUs:
    - Cluster gigantesques pour OpenAI, XAI, Google, …
    - Mais … clusters Européens significatifs dédiés à la recherche et aux startups 🤞
    - A relativiser avec XAI par exemple: 200,000 H100 reporté pour Colossus (!)
- Ecart de règles:
  - Les modèles Open Source se doivent d’essayer de respecter toutes les réglementations

Supercomputer	Country	Number of GPUs	GPU Type(s)
LUMI	Finland	~11,912	AMD MI250X
MareNostrum 5 (BSC)	Spain	4,480	NVIDIA H100 (Hopper)
JUPITER	Germany	~24,000	NVIDIA H100 (in GH200 Grace Hopper Superchips)
Leonardo	Italy	13,824	NVIDIA A100
Jean Zay	France	3,704	Mixed: 1,456 NVIDIA H100 + 416 NVIDIA A100 + 1,832 NVIDIA V100
Alps	Switzerland	10,752	NVIDIA H100 (in GH200 Grace Hopper Superchips)

European clusters 🇪🇺

Open-source vs proprietary

Quel raisons pour l'écart?

- Ecart de moyens:
  - Humains:
    - Salaire & taille d'équipes
  - GPUs:
    - Cluster gigantesques pour OpenAI, XAI, Google, …
    - Mais … clusters Européens significatifs dédiés à la recherche et aux startups ☝️
    - A relativiser avec XAI par exemple: 200,000 H100 reporté pour Colossus (!)
- Ecart de règles:
  - Les modèles Open Source se doivent d'essayer de respecter toutes les réglementations
  - Meta a utilisé un dataset de dump de livres piratés collectés sur Libgen (source: documents révélés lors du procès)

Supercomputer	Country	Number of GPUs	GPU Type(s)
LUMI	Finland	~11,912	AMD MI250X
MareNostrum 5 (BSC)	Spain	4,480	NVIDIA H100 (Hopper)
JUPITER	Germany	~24,000	NVIDIA H100 (in GH200 Grace Hopper Superchips)
Leonardo	Italy	13,824	NVIDIA A100
Jean Zay	France	3,704	Mixed: 1,456 NVIDIA H100 + 416 NVIDIA A100 + 1,832 NVIDIA V100
Alps	Switzerland	10,752	NVIDIA H100 (in GH200 Grace Hopper Superchips)

European clusters 🇪🇺

Open-source vs proprietary

Quel raisons pour l'écart?

- Ecart de moyens:
  - Humains:
    - Salaire & taille d'équipes
  - GPUs:
    - Cluster gigantesques pour OpenAI, XAI, Google, …
    - Mais … clusters Européens significatifs dédiés à la recherche et aux startups 🤞
    - A relativiser avec XAI par exemple: 200,000 H100 reporté pour Colossus (!)
- Ecart de règles:
  - Les modèles Open Source se doivent d'essayer de respecter toutes les réglementations
  - Meta a utilisé un dataset de dump de livres piratés collectés sur Libgen (source: documents révélés lors du procès)
- Ecart structurel:

Supercomputer	Country	Number of GPUs	GPU Type(s)
LUMI	Finland	~11,912	AMD MI250X
MareNostrum 5 (BSC)	Spain	4,480	NVIDIA H100 (Hopper)
JUPITER	Germany	~24,000	NVIDIA H100 (in GH200 Grace Hopper Superchips)
Leonardo	Italy	13,824	NVIDIA A100
Jean Zay	France	3,704	Mixed: 1,456 NVIDIA H100 + 416 NVIDIA A100 + 1,832 NVIDIA V100
Alps	Switzerland	10,752	NVIDIA H100 (in GH200 Grace Hopper Superchips)

European clusters 🇪🇺

Open-source vs proprietary

Quel raisons pour l'écart?

- Ecart de moyens:
  - Humains:
    - Salaire & taille d'équipes
  - GPUs:
    - Cluster gigantesques pour OpenAI, XAI, Google, …
    - Mais … clusters Européens significatifs dédiés à la recherche et aux startups 🤞
    - A relativiser avec XAI par exemple: 200,000 H100 reporté pour Colossus (!)
- Ecart de règles:
  - Les modèles Open Source se doivent d'essayer de respecter toutes les réglementations
  - Meta a utilisé un dataset de dump de livres piratés collectés sur Libgen (source: documents révélés lors du procès)
- Ecart structurel:
  - Modèles évalués principalement sur LMArena; plateforme devenue privée

Supercomputer	Country	Number of GPUs	GPU Type(s)
LUMI	Finland	~11,912	AMD MI250X
MareNostrum 5 (BSC)	Spain	4,480	NVIDIA H100 (Hopper)
JUPITER	Germany	~24,000	NVIDIA H100 (in GH200 Grace Hopper Superchips)
Leonardo	Italy	13,824	NVIDIA A100
Jean Zay	France	3,704	Mixed: 1,456 NVIDIA H100 + 416 NVIDIA A100 + 1,832 NVIDIA V100
Alps	Switzerland	10,752	NVIDIA H100 (in GH200 Grace Hopper Superchips)

European clusters 🇪🇺

Open-source vs proprietary

Quel raisons pour l'écart?

- Ecart de moyens:
  - Humains:
    - Salaire & taille d'équipes
  - GPUs:
    - Cluster gigantesques pour OpenAI, XAI, Google, …
    - Mais … clusters Européens significatifs dédiés à la recherche et aux startups ☝️
    - A relativiser avec XAI par exemple: 200,000 H100 reporté pour Colossus (!)
- Ecart de règles:
  - Les modèles Open Source se doivent d'essayer de respecter toutes les réglementations
  - Meta a utilisé un dataset de dump de livres piratés collectés sur Libgen (source: documents révélés lors du procès)
- Ecart structurel:
  - Modèles évalués principalement sur LMArena; plateforme devenue privée
  - Les modèles propriétaire ont accès à des quantités phénoménale de données pour post-entrainer leur modèle

Supercomputer	Country	Number of GPUs	GPU Type(s)
LUMI	Finland	~11,912	AMD MI250X
MareNostrum 5 (BSC)	Spain	4,480	NVIDIA H100 (Hopper)
JUPITER	Germany	~24,000	NVIDIA H100 (in GH200 Grace Hopper Superchips)
Leonardo	Italy	13,824	NVIDIA A100
Jean Zay	France	3,704	Mixed: 1,456 NVIDIA H100 + 416 NVIDIA A100 + 1,832 NVIDIA V100
Alps	Switzerland	10,752	NVIDIA H100 (in GH200 Grace Hopper Superchips)

European clusters 🇪🇺

Open-source vs proprietary

Quel raisons pour l'écart?

- Ecart de moyens:
  - Humains:
    - Salaire & taille d'équipes
  - GPUs:
    - Cluster gigantesques pour OpenAI, XAI, Google, …
    - Mais … clusters Européens significatifs dédiés à la recherche et aux startups ☝️
    - A relativiser avec XAI par exemple: 200,000 H100 reporté pour Colossus (!)
- Ecart de règles:
  - Les modèles Open Source se doivent d'essayer de respecter toutes les réglementations
  - Meta a utilisé un dataset de dump de livres piratés collectés sur Libgen (source: documents révélés lors du procès)
- Ecart structurel:
  - Modèles évalués principalement sur LMArena; plateforme devenue privée
  - Les modèles propriétaire ont accès à des quantités phénoménale de données pour post-entrainer leur modèle
  - "If you don't pay you are the product"

Supercomputer	Country	Number of GPUs	GPU Type(s)
LUMI	Finland	~11,912	AMD MI250X
MareNostrum 5 (BSC)	Spain	4,480	NVIDIA H100 (Hopper)
JUPITER	Germany	~24,000	NVIDIA H100 (in GH200 Grace Hopper Superchips)
Leonardo	Italy	13,824	NVIDIA A100
Jean Zay	France	3,704	Mixed: 1,456 NVIDIA H100 + 416 NVIDIA A100 + 1,832 NVIDIA V100
Alps	Switzerland	10,752	NVIDIA H100 (in GH200 Grace Hopper Superchips)

European clusters 🇪🇺

Un autre enjeu clé: la diversité linguistique

- Plutot que LMArena, considérez https://comparia.beta.gouv.fr/
  - Entièrement ouverte: code & données!
- Quel performance pour les modèles **ouverts**?
- Quel performance pour les modèles **Européens**?

Total modèles: 77 | **Total votes:** 235 000 | Mise à jour le 14/01/2026 | Télécharger les données | *Rechercher un modèle*

Rang	Modèle	Score de satisfaction BT	Confiance (±)	Total votes	Conso. moyenne (1000 tokens)	Taille (paramètres)	A
1	gemini-3-flash-preview	1132	-9/+1	174	N/A	**XL** - (estimation)	Pr
2	mistral-large-2512	1120	-3/+2	1442	50 Wh	**XL** - 675 Mds	M
3	mistral-medium-2508	1115	-1/+2	5387	N/A	**L** - (estimation)	Pr
4	gemini-2.5-flash	1104	-4/+1	3972	N/A	**XL** - (estimation)	Pr
5	gemini-3-pro-preview	1104	-4/+3	1974	N/A	**XL** - (estimation)	Pr
6	qwen3-max-2025-09-23	1103	-3/+3	2562	N/A	**XL** - (estimation)	Pr
7	gemini-2.0-flash	1100	-2/+3	8684	N/A	**XL** - (estimation)	Pr
8	deepseek-v3-0324	1091	-4/+2	4385	47 Wh	**XL** - 685 Mds	M
9	magistral-medium	1088	-6/+3	2207	N/A	**L** - (estimation)	Pr
10	gpt-5.2	1085	-10/+4	1080	N/A	**L** - (estimation)	Pr
11	gemma-3-27b	1083	-5/+2	7688	6 Wh	**S** - 27 Mds	D
12	grok-4.1-fast	1079	-10/+4	1461	N/A	**XL** - (estimation)	Pr
13	deepseek-chat-v3.1	1079	-8/+4	1588	47 Wh	**XL** - 685 Mds	M
14	deepseek-v3-chat	1078	-6/+4	5388	47 Wh	**XL** - 671 Mds	M
15	gpt-5.1	1074	-8/+5	2202	N/A	**L** - (estimation)	Pr
16	claude-4-5-sonnet	1073	-5/+4	5031	N/A	**XL** - (estimation)	Pr

Enjeux de souveraineté

Enjeux de souveraineté

- Les modèles Européens sont très loin en performance des modèles Chinois et Américain

Enjeux de souveraineté

- Les modèles Européens sont très loin en performance des modèles Chinois et Américain
- … à part Mistral

Enjeux de souveraineté

- Les modèles Européens sont très loin en performance des modèles Chinois et Américain
- … à part Mistral
- L’accès à cette technologie transformative dépend:

Enjeux de souveraineté

- Les modèles Européens sont très loin en performance des modèles Chinois et Américain
- … à part Mistral
- L’accès à cette technologie transformative dépend:
  - D’un seul acteur Européen

Enjeux de souveraineté

- Les modèles Européens sont très loin en performance des modèles Chinois et Américain
- … à part Mistral
- L’accès à cette technologie transformative dépend:
  - D’un seul acteur Européen
  - De deux pays potentiellement rivaux (l’accès à cette technologie peut être bloqué, cf l’Etats-Unis empêche l’export de GPU avancés vers la Chine)

Enjeux de souveraineté

- Les modèles Européens sont très loin en performance des modèles Chinois et Américain
- … à part Mistral
- L’accès à cette technologie transformative dépend:
  - D’un seul acteur Européen
  - De deux pays potentiellement rivaux (l’accès à cette technologie peut être bloqué, cf l’Etats-Unis empêche l’export de GPU avancés vers la Chine)
- Projets Européen ayant pour but de développer des modèles ouverts et améliorer l’ecosystème Européen

Enjeux de souveraineté

- Les modèles Européens sont très loin en performance des modèles Chinois et Américain
- … à part Mistral
- L’accès à cette technologie transformative dépend:
  - D’un seul acteur Européen
  - De deux pays potentiellement rivaux (l’accès à cette technologie peut être bloqué, cf l’Etats-Unis empêche l’export de GPU avancés vers la Chine)
- Projets Européen ayant pour but de développer des modèles ouverts et améliorer l’ecosystème Européen
  - EuroLLM, Apertus, OpenEuroLLM

OpenEuroLLM

OpenEuroLLM

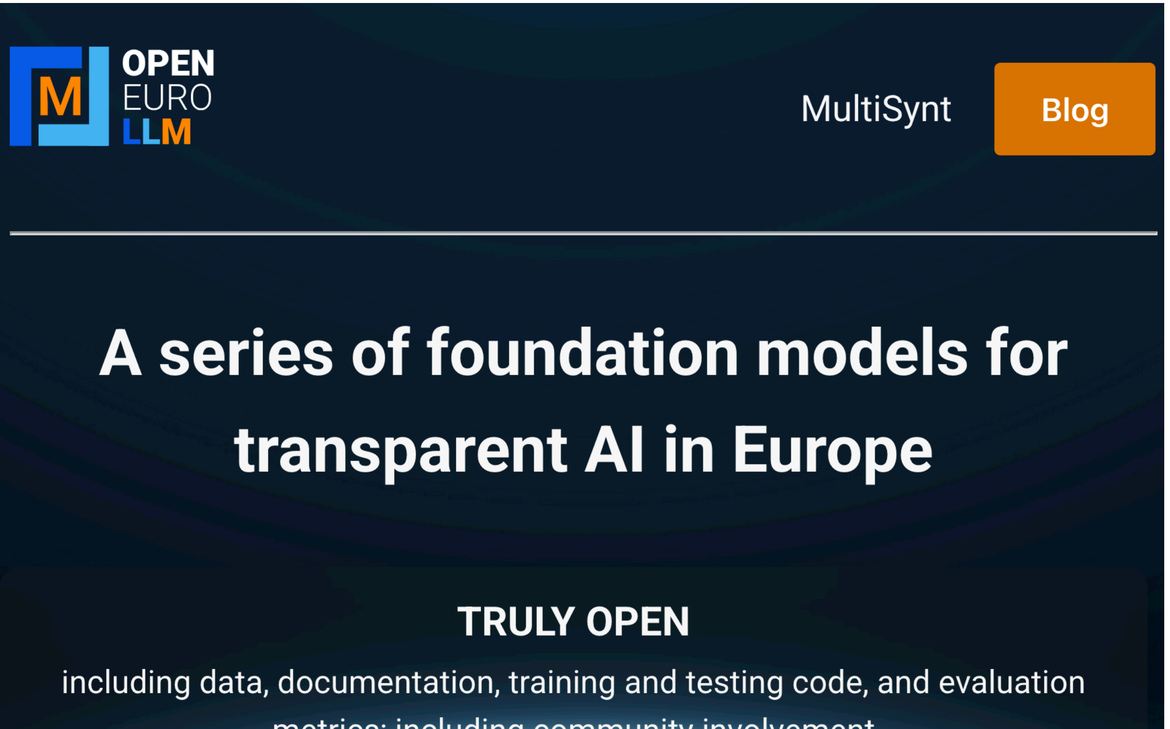
- Un effort pour construire des LLM multilingues à partir de zéro d'ici 2028, lancé en février 2025

OpenEuroLLM

- Un effort pour construire des LLM multilingues à partir de zéro d'ici 2028, lancé en février 2025
- https://openeurollm.eu/

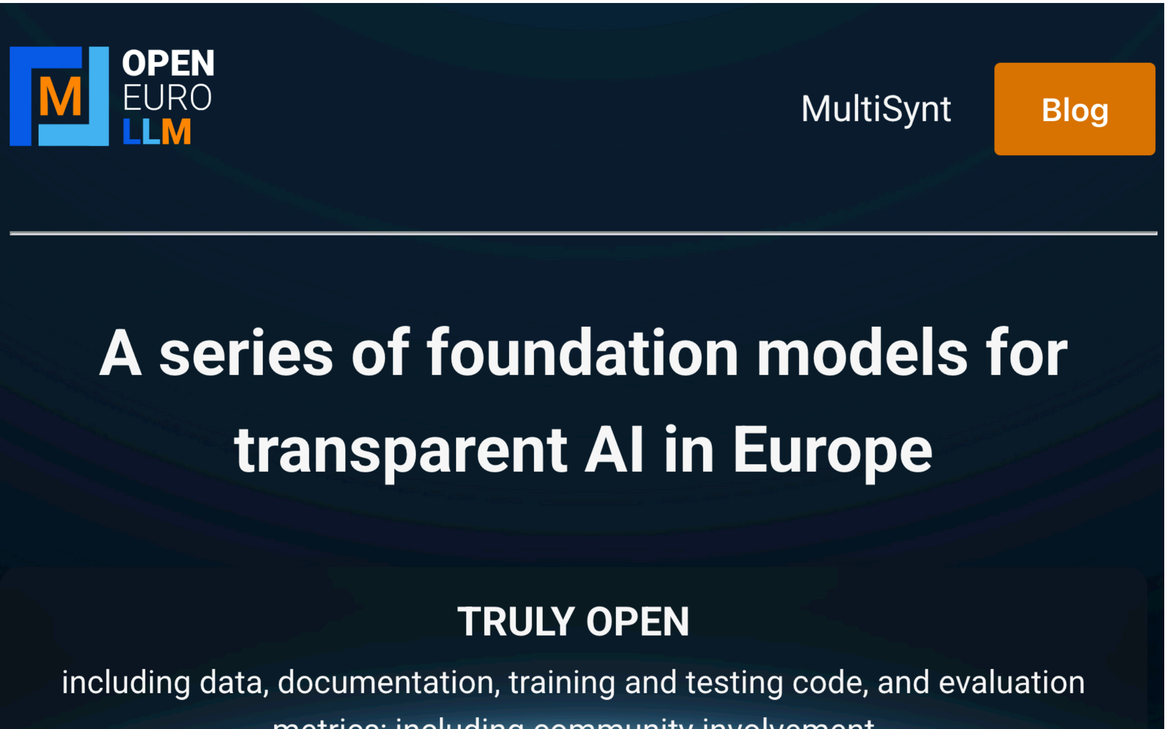
OpenEuroLLM

- Un effort pour construire des LLM multilingues à partir de zéro d'ici 2028, lancé en février 2025
- https://openeurollm.eu/



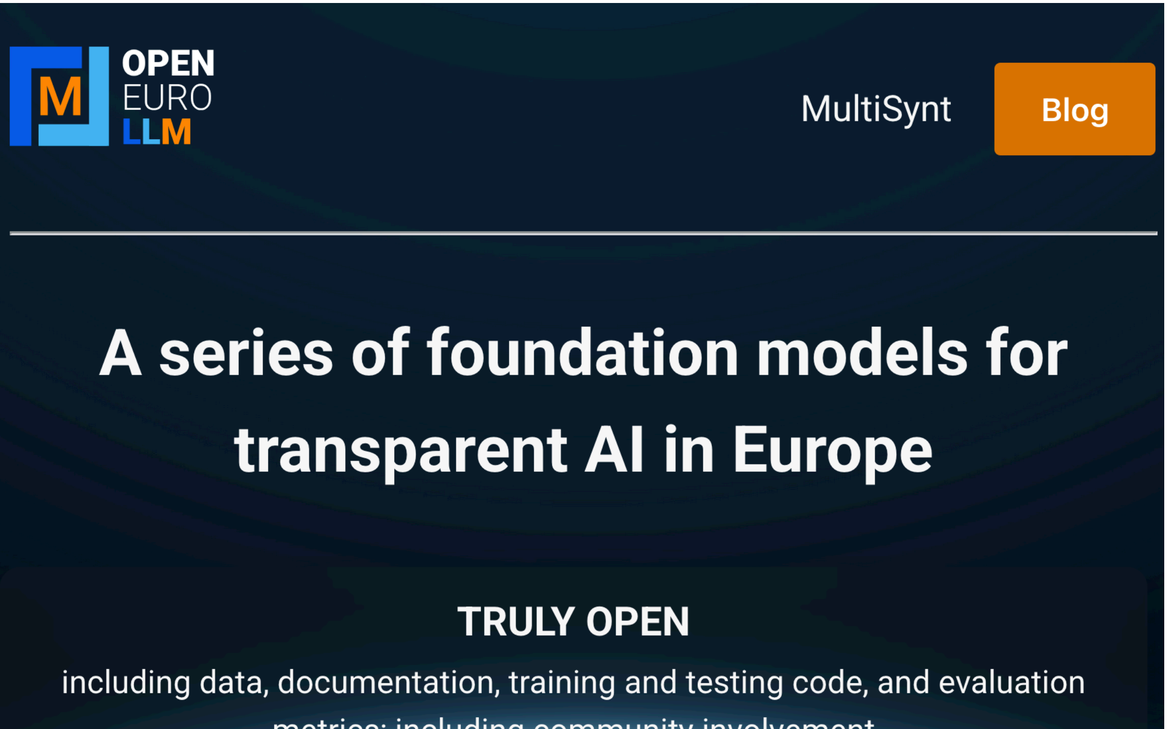
OpenEuroLLM

- Un effort pour construire des LLM multilingues à partir de zéro d'ici 2028, lancé en février 2025
- https://openeurollm.eu/
- Objectif : égaler l'état de l'art en anglais et améliorer significativement les performances sur les langues européennes



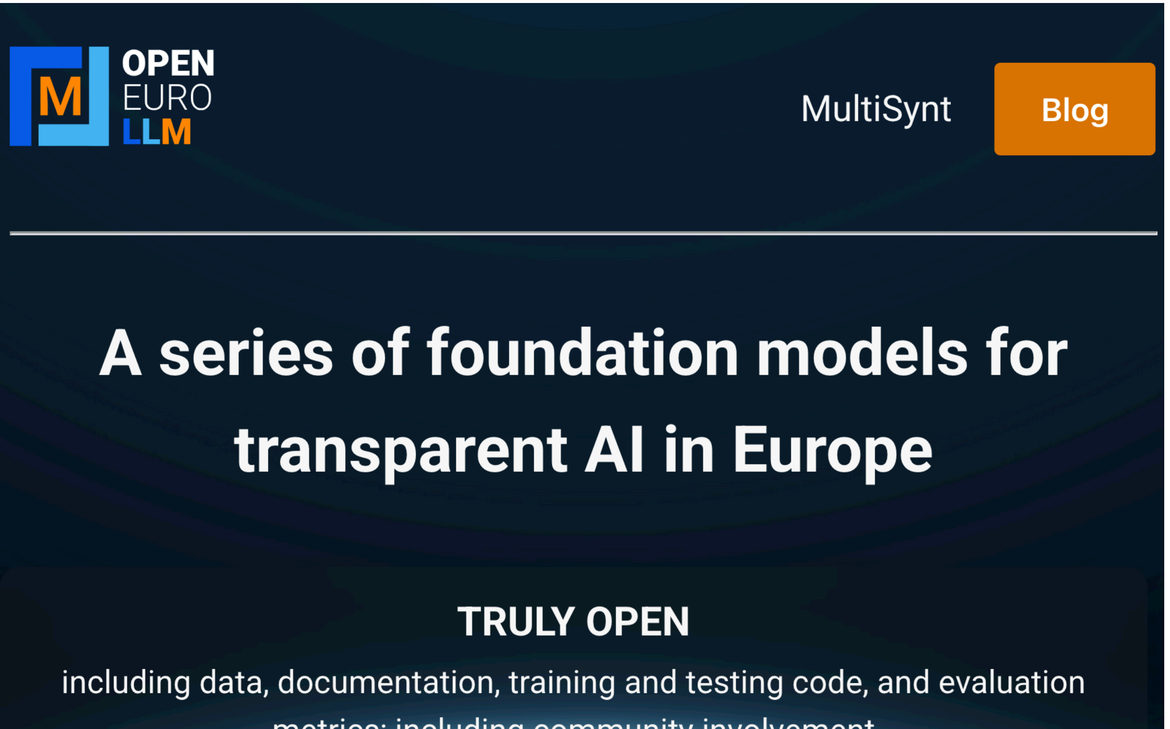
OpenEuroLLM

- Un effort pour construire des LLM multilingues à partir de zéro d'ici 2028, lancé en février 2025
- https://openeurollm.eu/
- Objectif : égaler l'état de l'art en anglais et améliorer significativement les performances sur les langues européennes
- Entièrement ouvert : poids, code et données



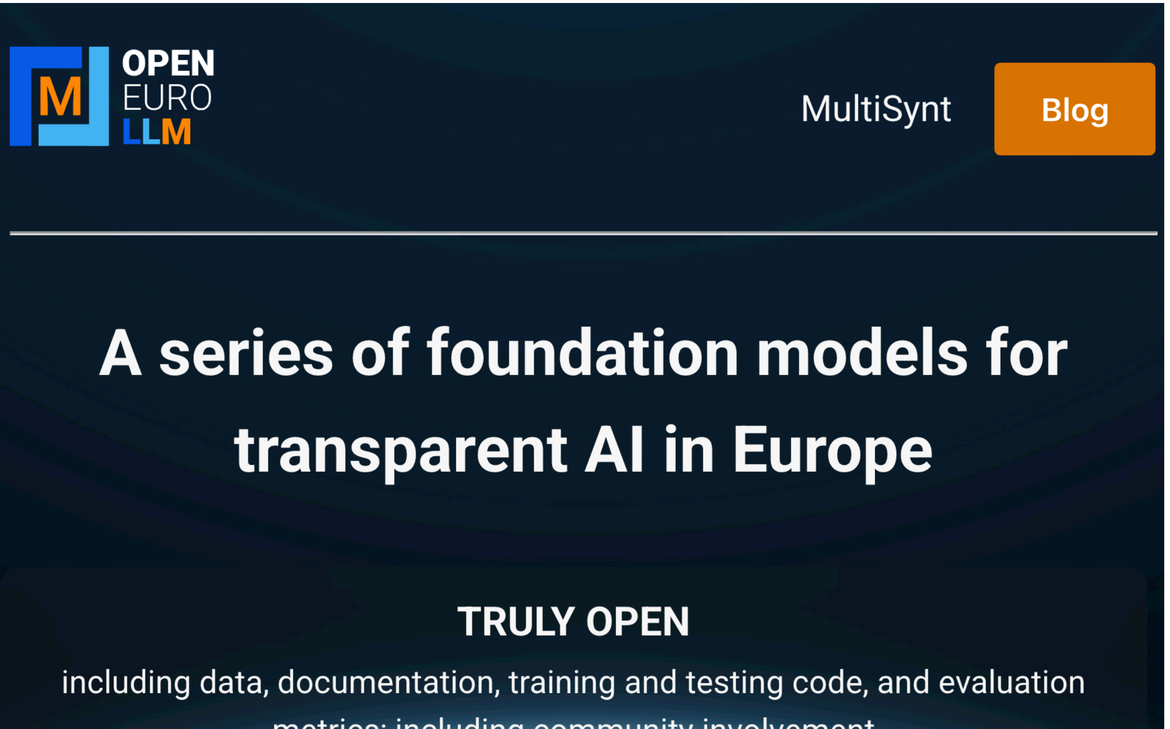
OpenEuroLLM

- Un effort pour construire des LLM multilingues à partir de zéro d'ici 2028, lancé en février 2025
- https://openeurollm.eu/
- Objectif : égaler l'état de l'art en anglais et améliorer significativement les performances sur les langues européennes
- Entièrement ouvert : poids, code et données
- Financement de 37,4 millions d’euros, également plusieurs millions d'heures GPU sur EuroHPC



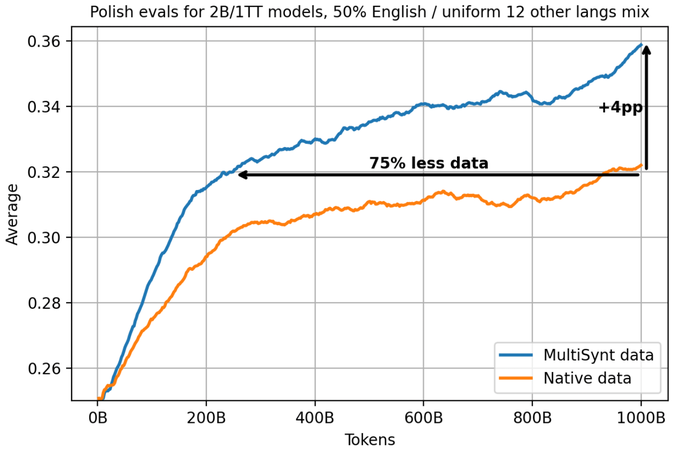
OpenEuroLLM

- Un effort pour construire des LLM multilingues à partir de zéro d'ici 2028, lancé en février 2025
- https://openeurollm.eu/
- Objectif : égaler l'état de l'art en anglais et améliorer significativement les performances sur les langues européennes
- Entièrement ouvert : poids, code et données
- Financement de 37,4 millions d’euros, également plusieurs millions d'heures GPU sur EuroHPC
- Jusqu’ici modèle entrainés avec 2B paramètres, entraînement en cours d’un modèle 8B pour Mai prochain



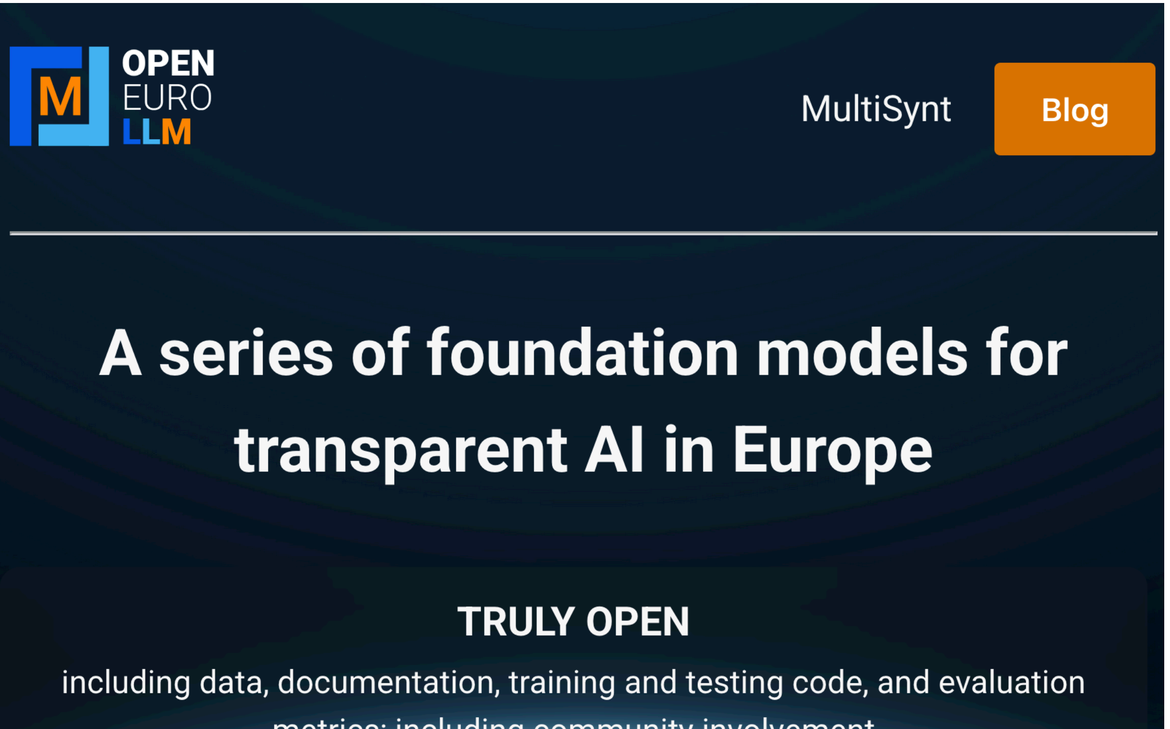
OpenEuroLLM

- Un effort pour construire des LLM multilingues à partir de zéro d'ici 2028, lancé en février 2025
- https://openeurollm.eu/
- Objectif : égaler l'état de l'art en anglais et améliorer significativement les performances sur les langues européennes
- Entièrement ouvert : poids, code et données
- Financement de 37,4 millions d’euros, également plusieurs millions d'heures GPU sur EuroHPC
- Jusqu’ici modèle entrainés avec 2B paramètres, entraînement en cours d’un modèle 8B pour Mai prochain



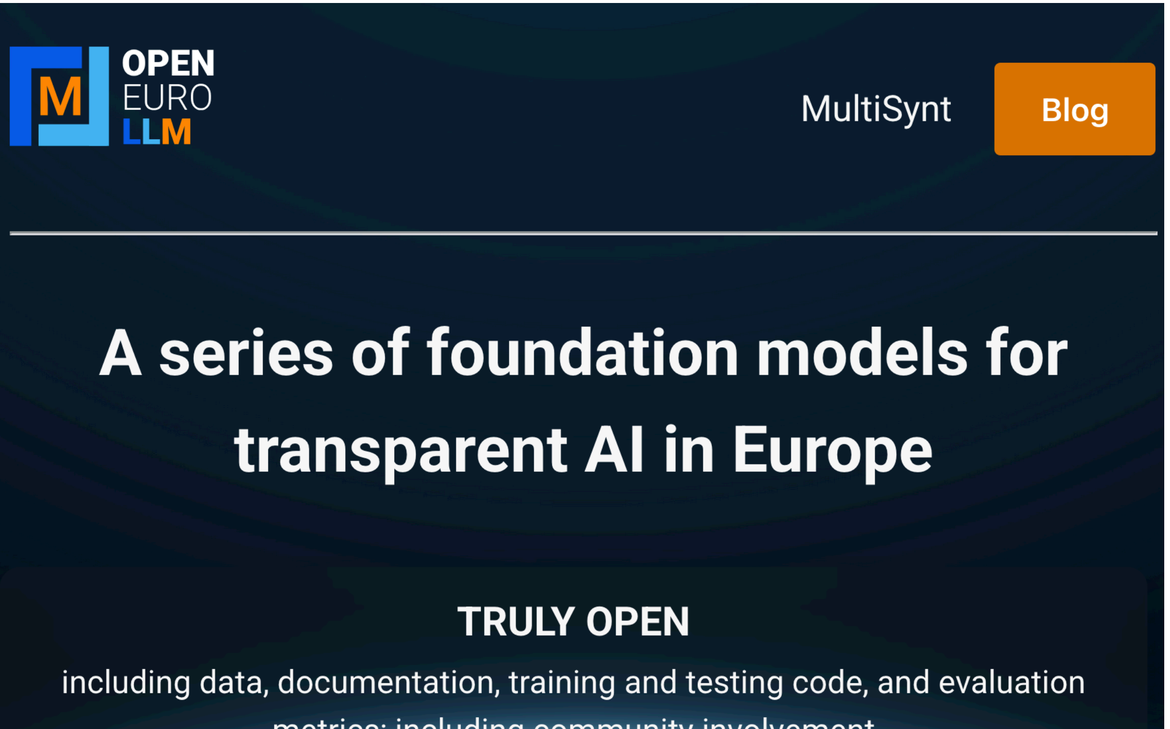
Average over belebele_pol_Latn_cf|0, global_mmlu_all_pol_cf|0, lumi_arc_pol_cf:challenge|0.

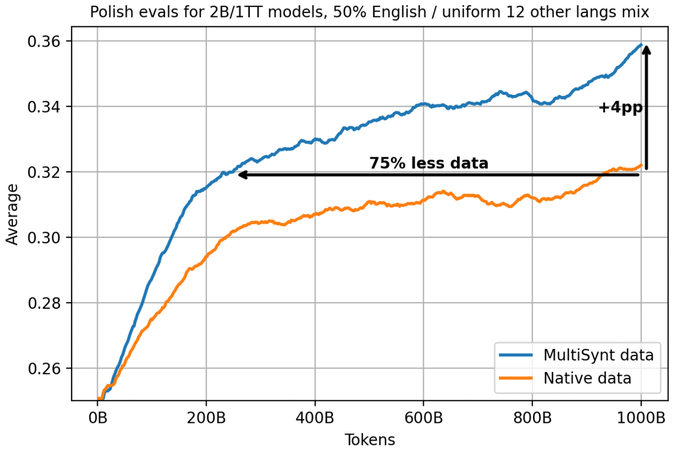
Utilisation de données synthétique pour les languages avec moins de ressources



OpenEuroLLM

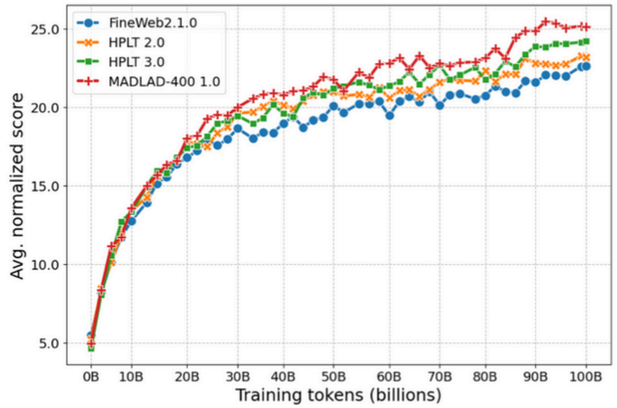
- Un effort pour construire des LLM multilingues à partir de zéro d'ici 2028, lancé en février 2025
- https://openeurollm.eu/
- Objectif : égaler l'état de l'art en anglais et améliorer significativement les performances sur les langues européennes
- Entièrement ouvert : poids, code et données
- Financement de 37,4 millions d’euros, également plusieurs millions d'heures GPU sur EuroHPC
- Jusqu’ici modèle entrainés avec 2B paramètres, entraînement en cours d’un modèle 8B pour Mai prochain





Utilisation de données synthétique pour les languages avec moins de ressources

Average over belebele_pol_Latn_cf|0, global_mmlu_all_pol_cf|0, lumi_arc_pol_cf:challenge|0.

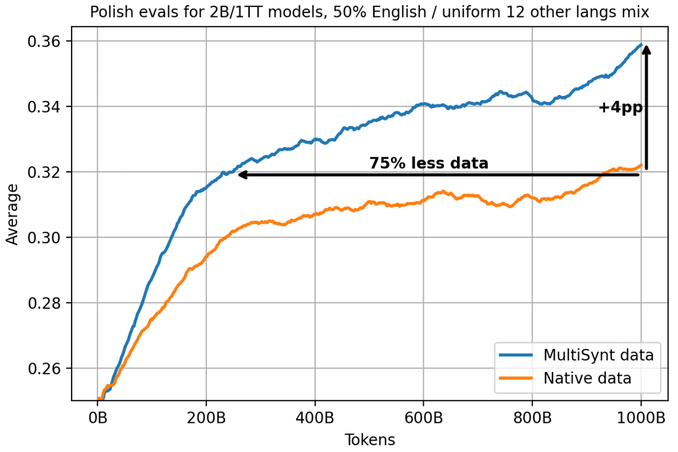


Acquisition (crawling) de jeux de données à grandes échelle pour les langues Européennes

Figure 2: Comparison of models pretrained on FineWeb, HPLT 2.0, 3.0, and MADLAD-400.

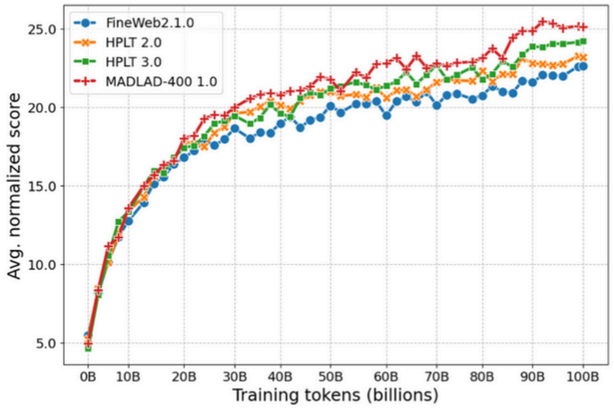
OpenEuroLLM

- Un effort pour construire des LLM multilingues à partir de zéro d'ici 2028, lancé en février 2025
- https://openeurollm.eu/
- Objectif : égaler l'état de l'art en anglais et améliorer significativement les performances sur les langues européennes
- Entièrement ouvert : poids, code et données
- Financement de 37,4 millions d’euros, également plusieurs millions d'heures GPU sur EuroHPC
- Jusqu’ici modèle entrainés avec 2B paramètres, entraînement en cours d’un modèle 8B pour Mai prochain

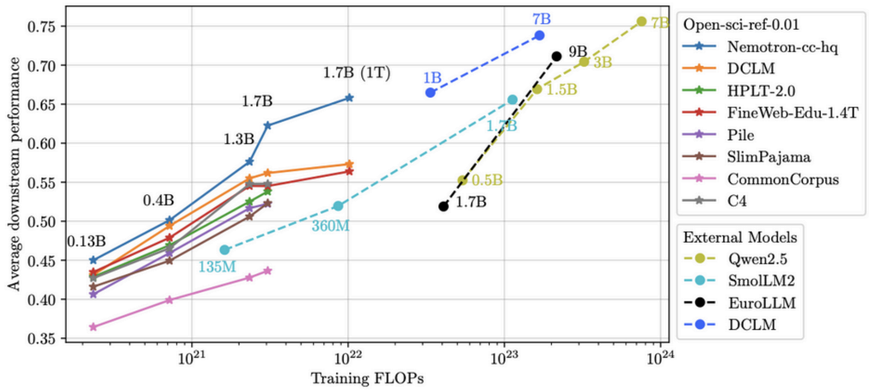


Average over belebele_pol_Latn_cf|0, global_mmlu_all_pol_cf|0, lumi_arc_pol_cf:challenge|0.

Utilisation de données synthétique pour les languages avec moins de ressources



Acquisition (crawling) de jeux de données à grandes échelle pour les langues Européennes



Scaling laws obtenues à petites échelles permettant de comparer des jeux de données.

OPEN EURO LLM

MultiSynt

Blog

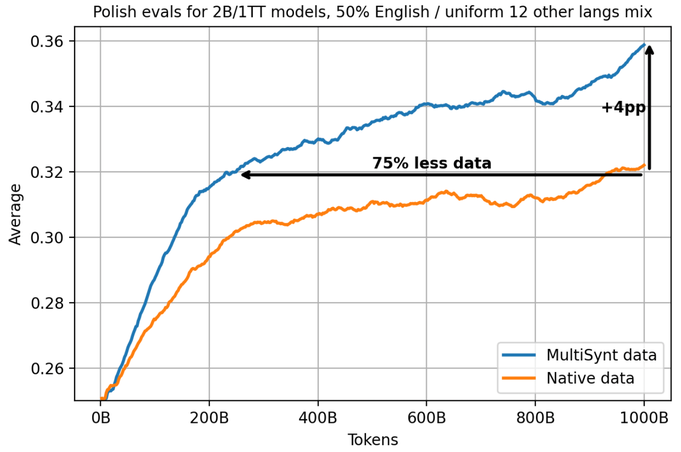
A series of foundation models for transparent AI in Europe

TRULY OPEN

including data, documentation, training and testing code, and evaluation

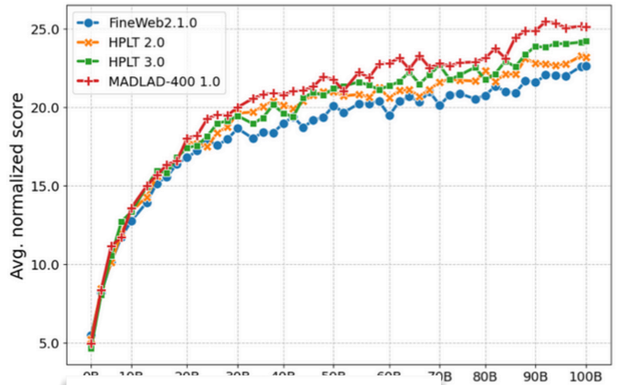
OpenEuroLLM

- Un effort pour construire des LLM multilingues à partir de zéro d'ici 2028, lancé en février 2025
- https://openeurollm.eu/
- Objectif : égaler l'état de l'art en anglais et améliorer significativement les performances sur les langues européennes
- Entièrement ouvert : poids, code et données
- Financement de 37,4 millions d’euros, également plusieurs millions d'heures GPU sur EuroHPC
- Jusqu’ici modèle entrainés avec 2B paramètres, entraînement en cours d’un modèle 8B pour Mai prochain



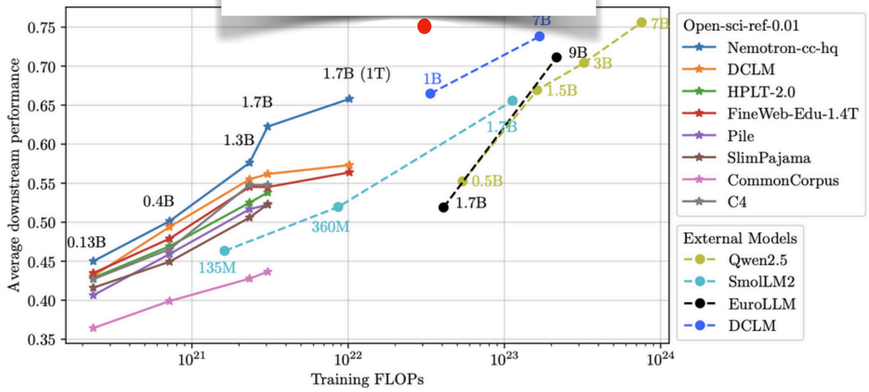
Average over belebele_pol_Latn_cf|0, global_mmlu_all_pol_cf|0, lumi_arc_pol_cf:challenge|0.

Utilisation de données synthétique pour les languages avec moins de ressources



Acquisition (crawling) de jeux de données à grandes échelle pour les langues Européennes

? A voir en Mai prochain 😅



Scaling laws obtenues à petites échelles permettant de comparer des jeux de données.

OPEN EURO LLM

MultiSynt

Blog

A series of foundation models for transparent AI in Europe

TRULY OPEN

including data, documentation, training and testing code, and evaluation

Conclusion

- Les agents de codage et les LLM : une technologie déjà transformative
- Enjeux critiques à adresser :
  - Sécurité : risques liés aux permissions et à l'exécution de code
  - Open vs propriétaire : transparence, contrôle, dépendance
  - Souveraineté : indépendance technologique européenne