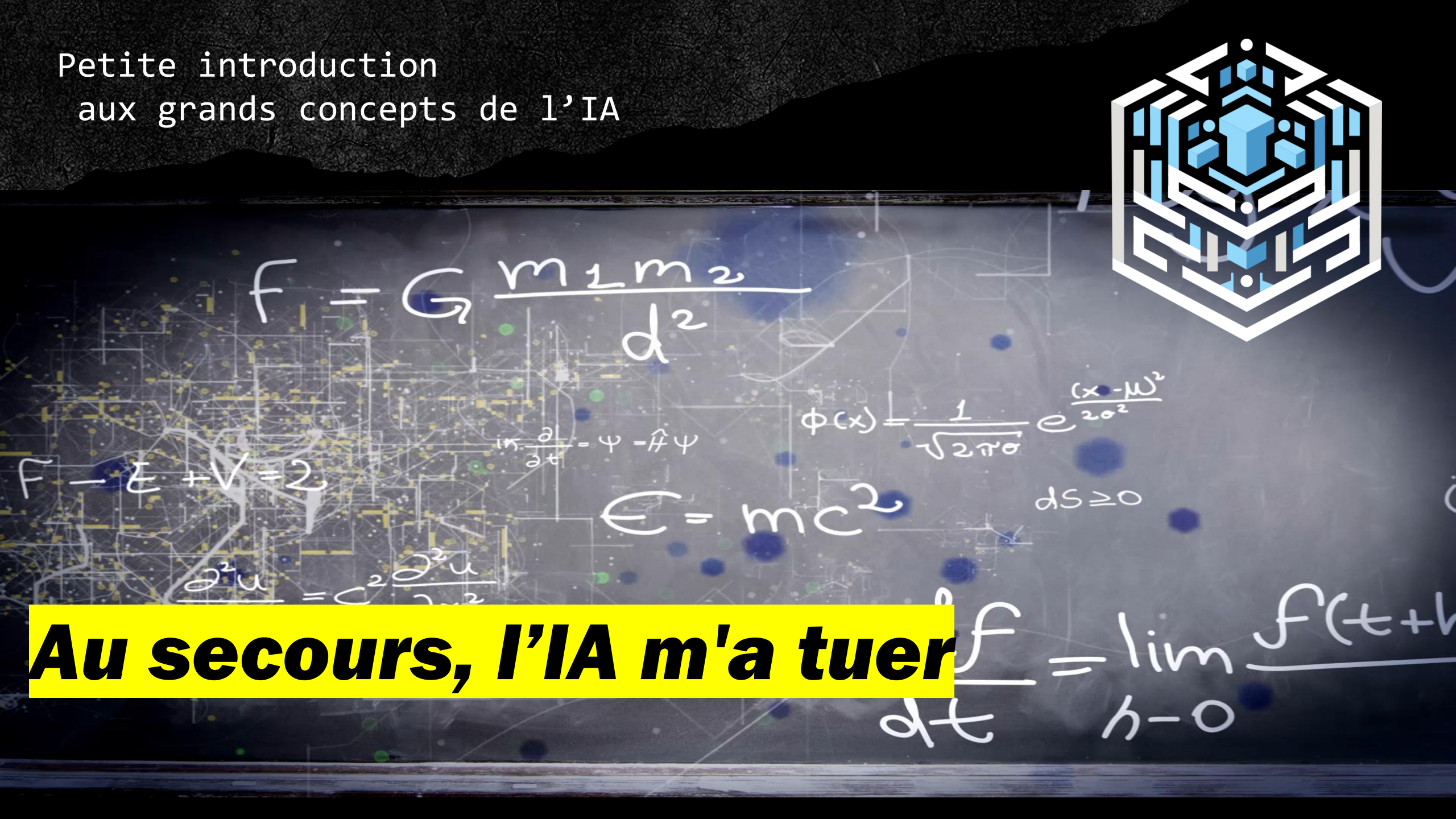
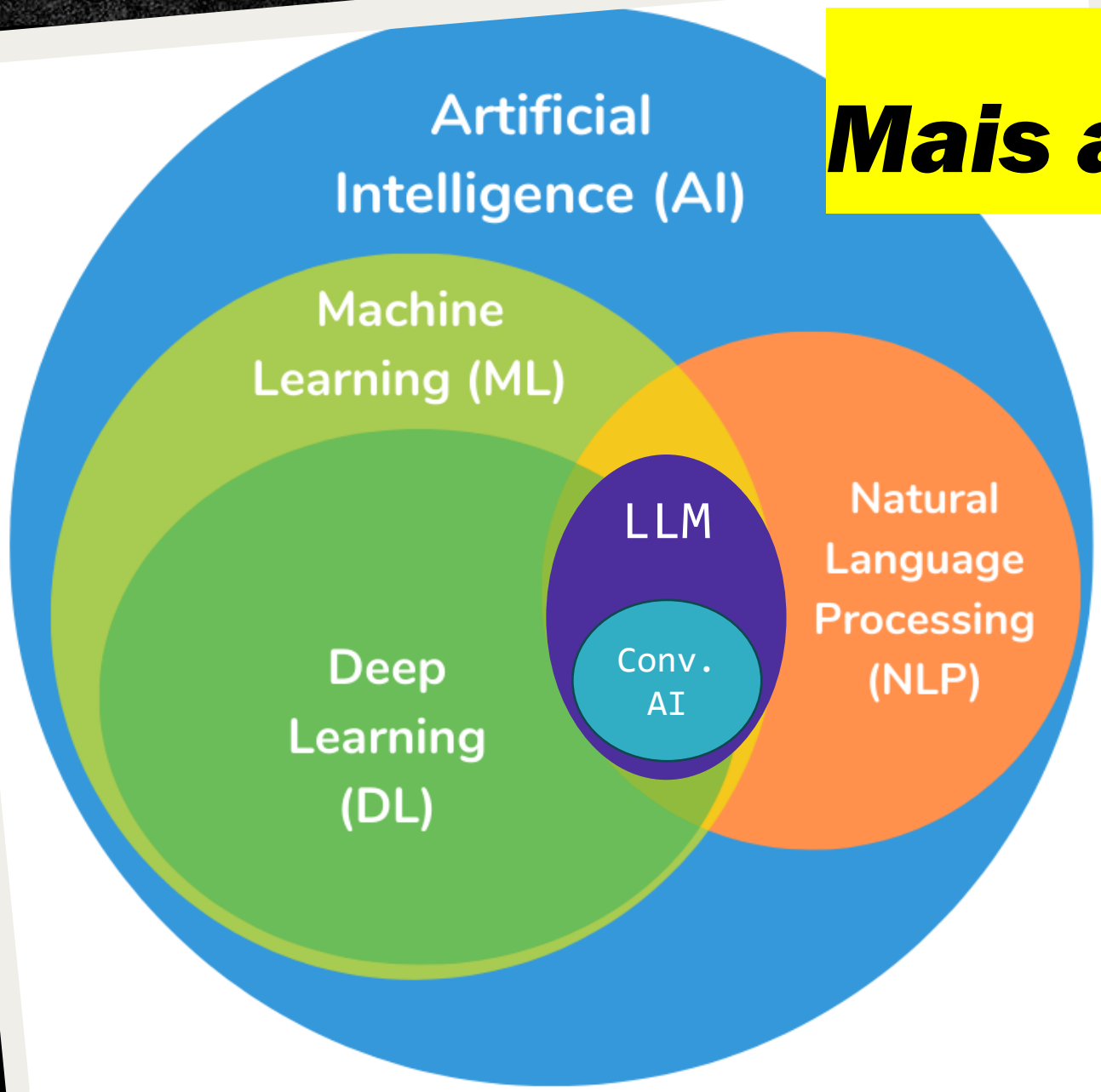
Petite introduction
aux grands concepts de l'IA



Au secours, l'IA m'a tuer



Mais au fait, c'est quoi l'IA ?

LLM, SLM ...

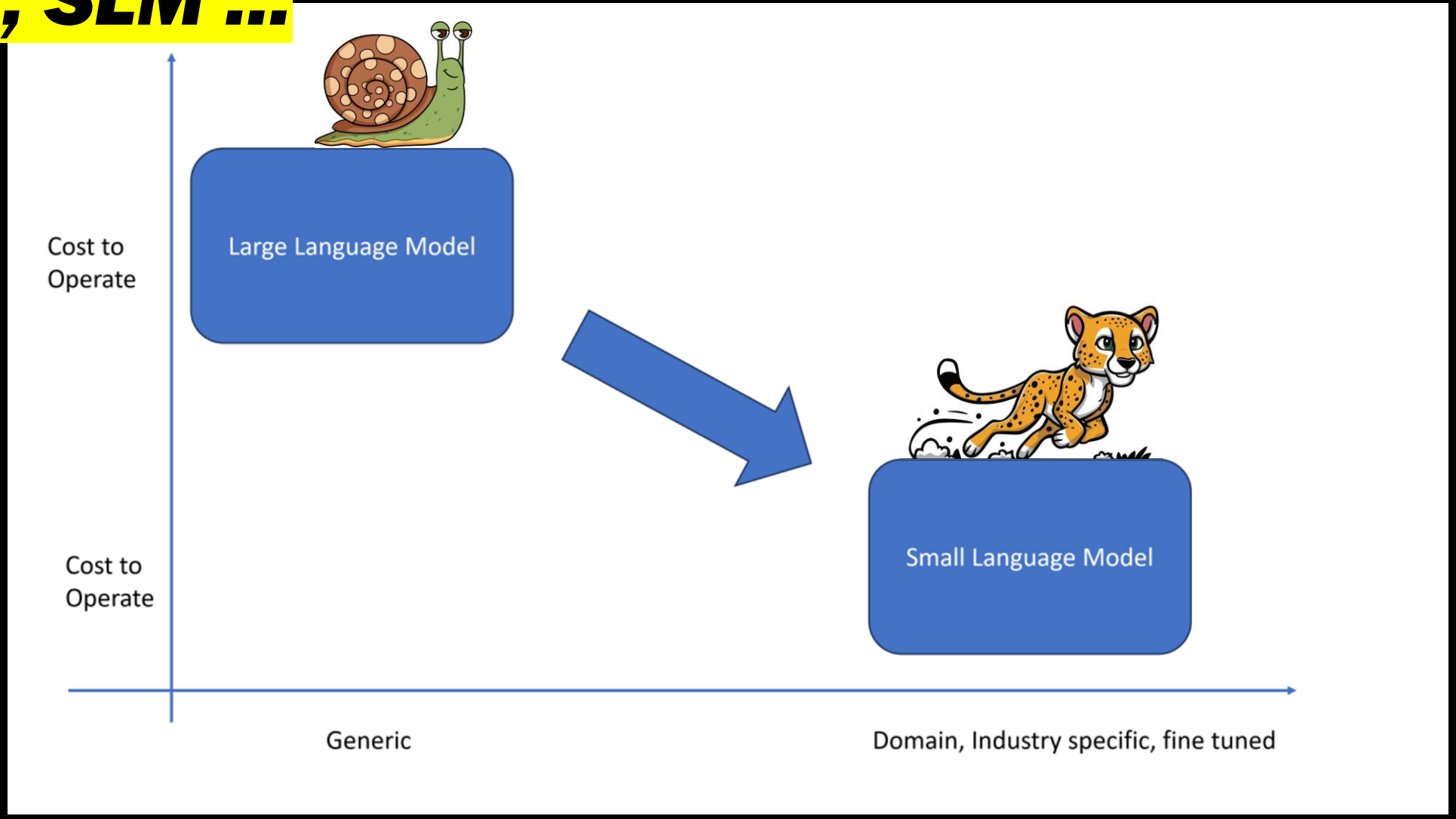
SLM: <= 12B params

- Les petits Google Gemma-3
- Les petits Meta Llama3.x
- Mistral Small
- Microsoft Phi-4

LLM:

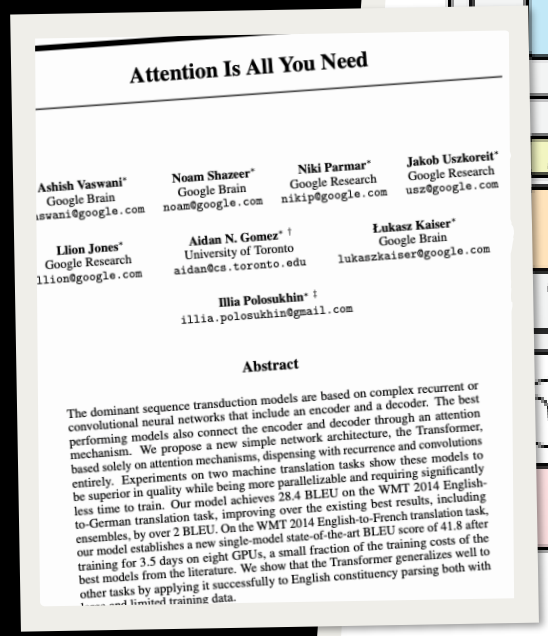
- OpenAI GPT-o1 / GPT-4o
- Xai Grok-3
- Deepseek V3 & R1
- Anthropic Claude 3.x
- Mistral Large
- Les gros Meta Llama-3.x

LLM, SLM ...


Cost to Operate
Large Language Model
Small Language Model
Cost to Operate
Generic
Domain, Industry specific, fine tuned

Et la théorie ?

- Un peu de géométrie : des vecteurs, mécanisme d'attention, etc
- Quelques layers de réseaux neuronaux
- Et le tour est joué…



Attention Is All You Need

Ashish Vaswani* Google Brain avaswani@google.com
Noam Shazeer* Google Brain noam@google.com
Niki Parmar* Google Research nikip@google.com
Jakob Uszkoreit* Google Research usz@google.com
Llion Jones* Google Research llion@google.com
Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu
Łukasz Kaiser* Google Brain lukaszkaiser@google.com
Illia Polosukhin* ‡ illia.polosukhin@gmail.com

Abstract

The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles, by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.8 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature. We show that the Transformer generalizes well to other tasks by applying it successfully to English constituency parsing both with

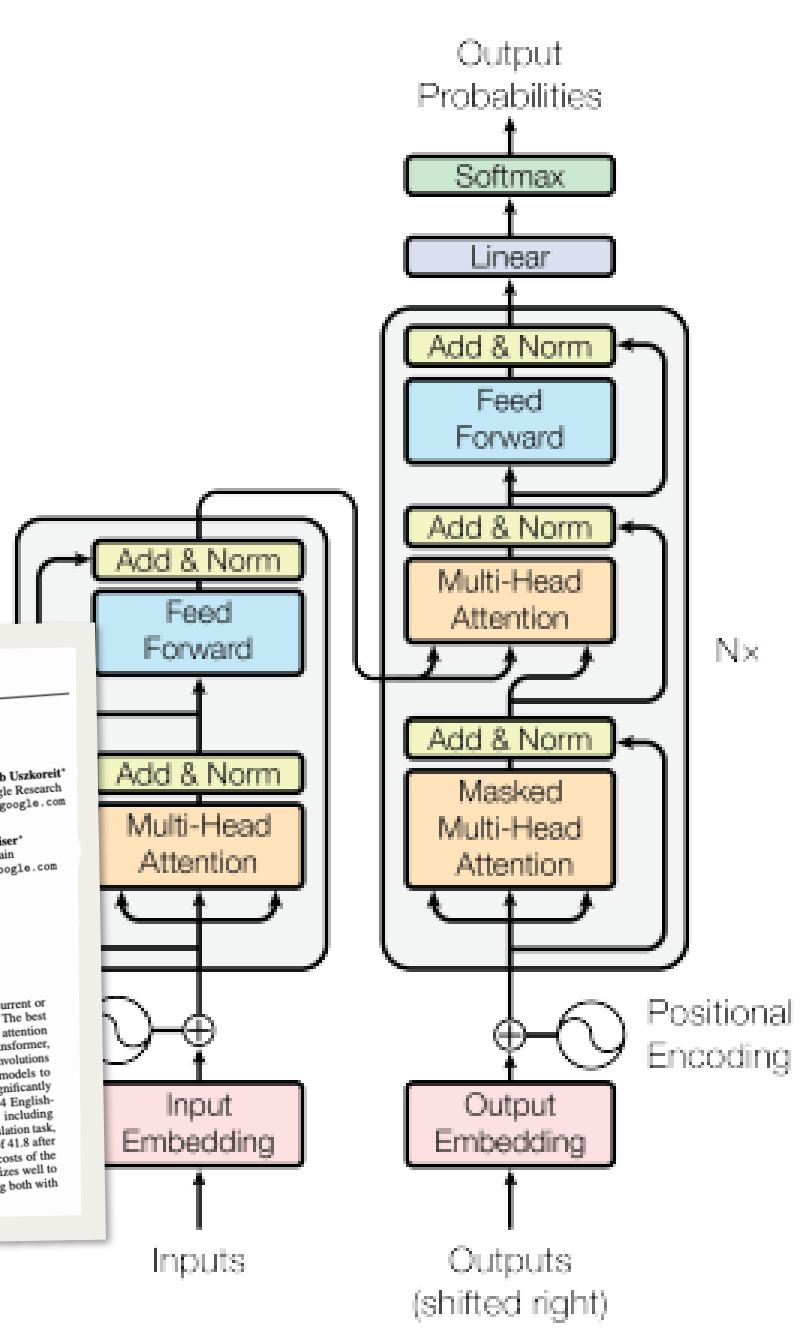
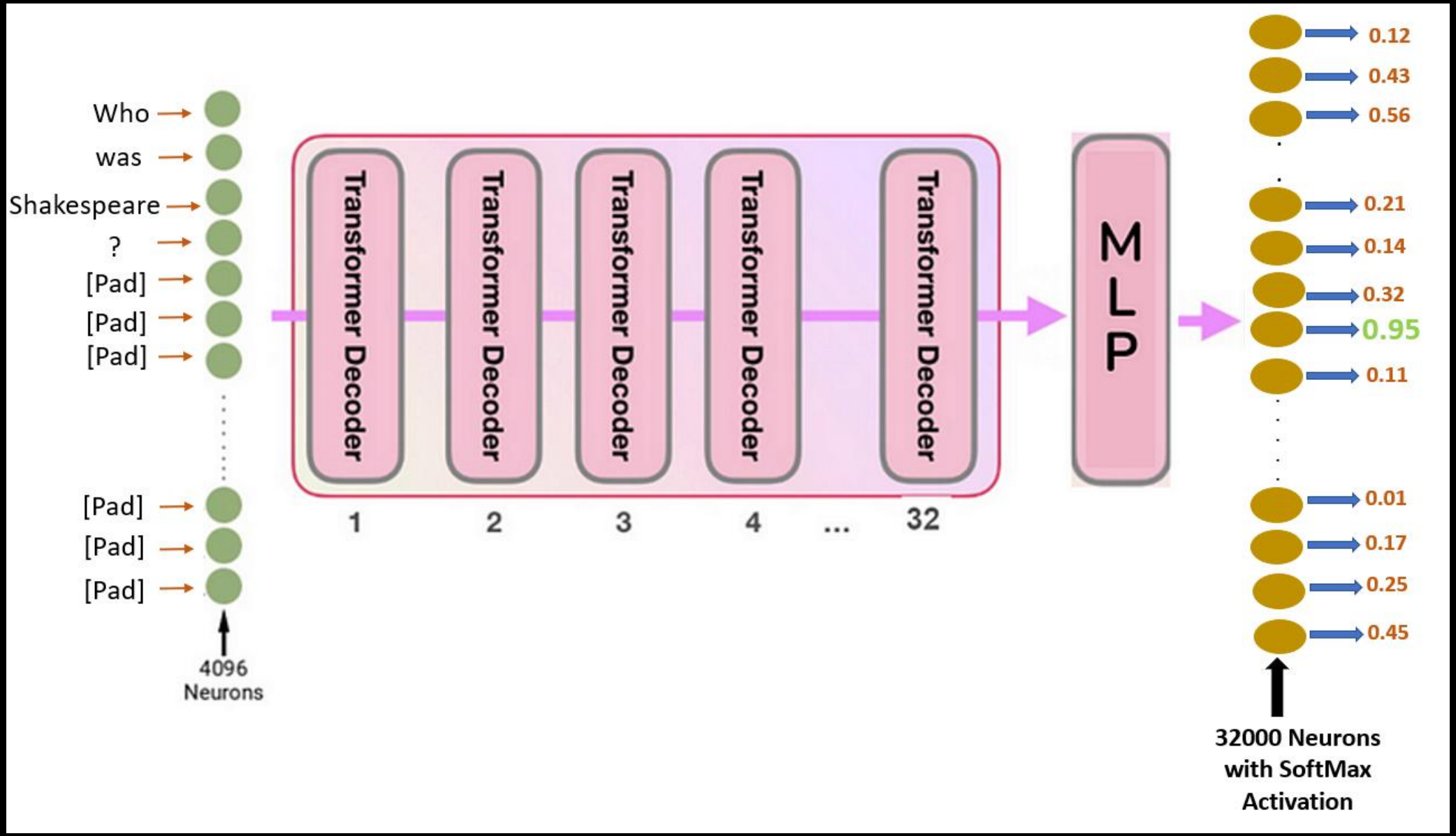


Figure 1: The Transformer - model architecture.

Le fonctionnement d'un modèle



La quantization

Floating Point

1231.4531

Integer

1231

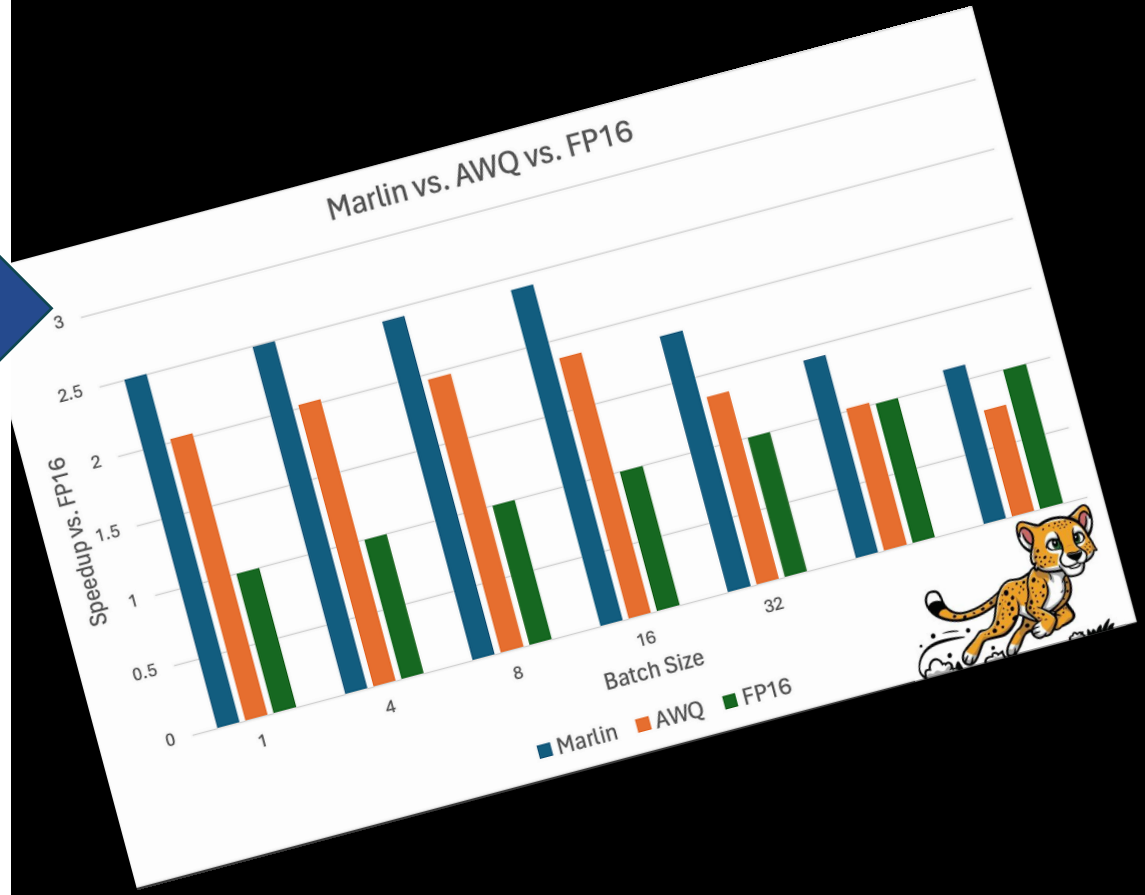
32 bit

0.21	-0.37	-2.54
4.5	4.37	-0.78
5.1	0.01	9.6

Quantization

8 bit

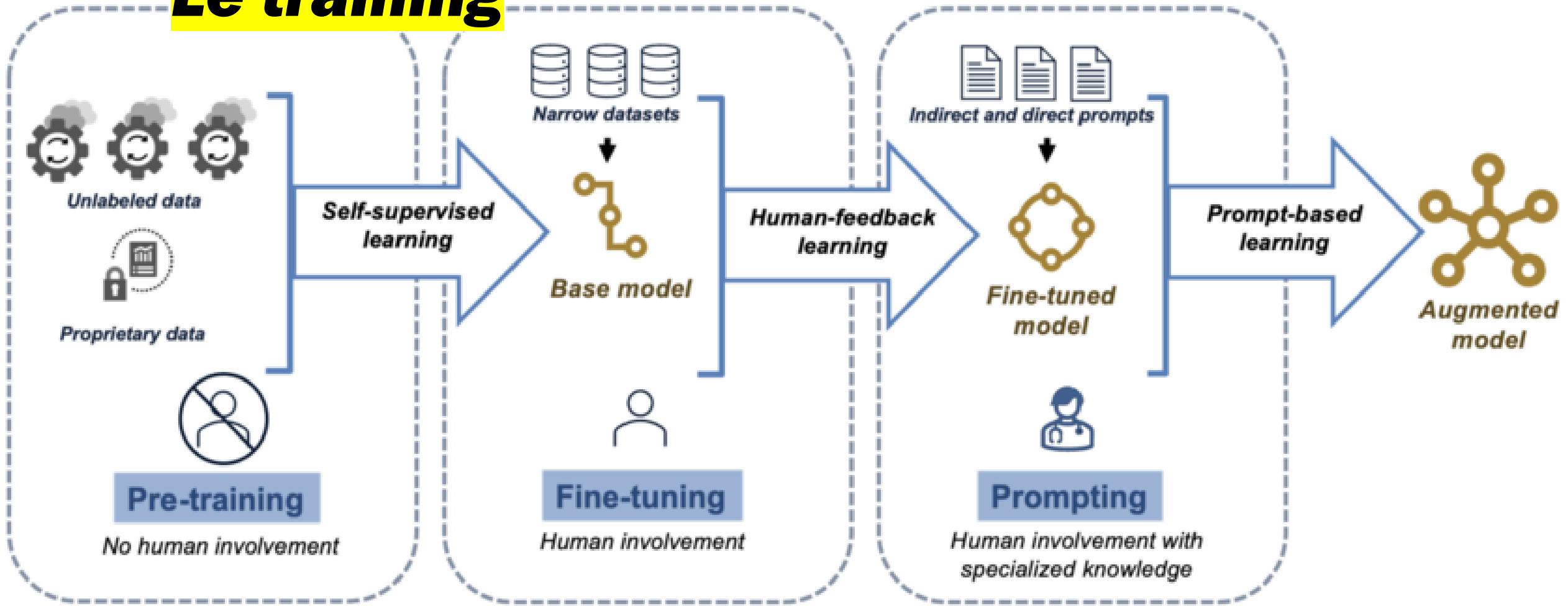
21	37	25
45	43	78
51	23	96



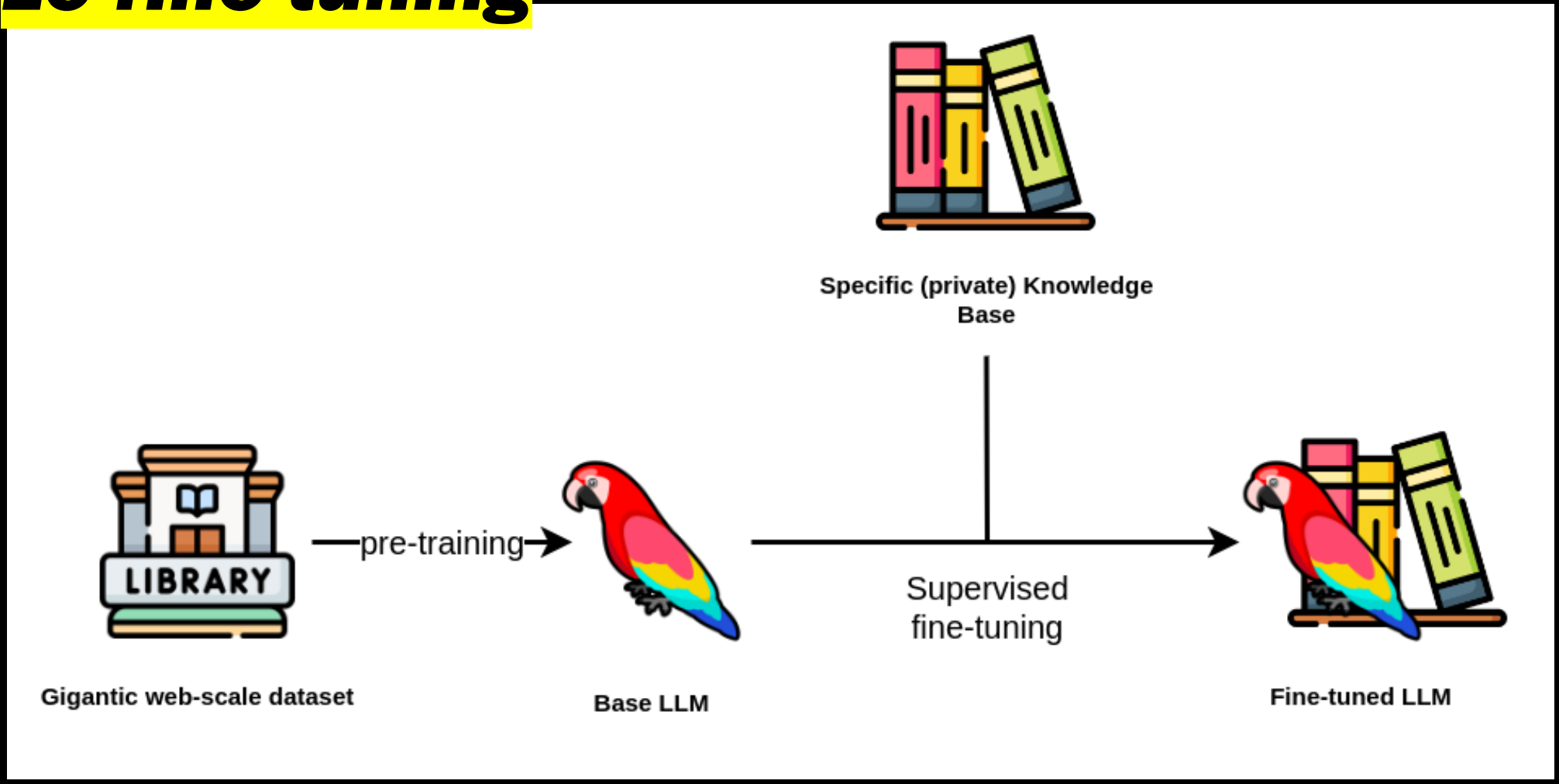
Faire tourner un modèle dans le cloud ?

Model	Memory	GPU	Price/mo (scaleway)
Llama-3.1 8B (q8)	> 9GB	Nvidia L4 (24GB) 	540€
Llama-3.3 70B (q4)	> 40GB	Nvidia L40S (48GB)	~1000€
Llama-3.3 70B (FP16)	> 150GB	2x Nvidia H100 (80GB) 	~4000€
Llama-3.1 405B (q8)	> 420GB	4x Nvidia H100-SXM (80GB)	~10000€

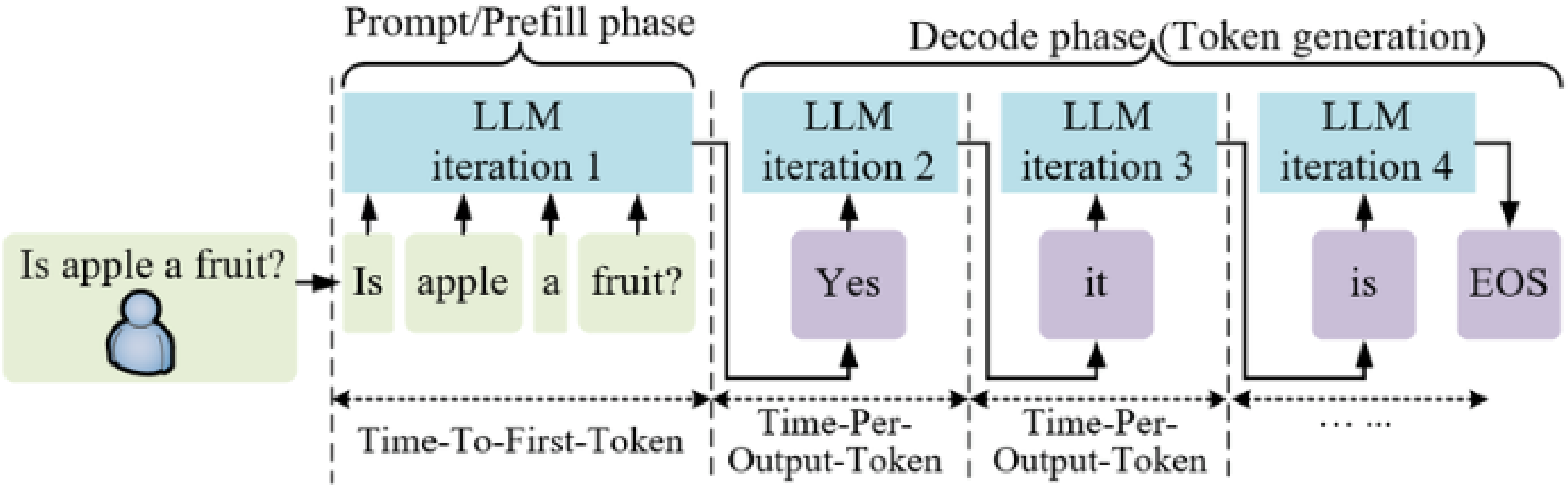
Le training


Unlabeled data
Proprietary data
Self-supervised learning
Narrow datasets
Base model
Human-feedback learning
Indirect and direct prompts
Fine-tuned model
Prompt-based learning
Augmented model
Pre-training
No human involvement
Fine-tuning
Human involvement
Prompting
Human involvement with specialized knowledge

Le fine tuning



L'inférence



LLM inference process illustration. (EOS: end-of-sequence).

Les limites des LLMs

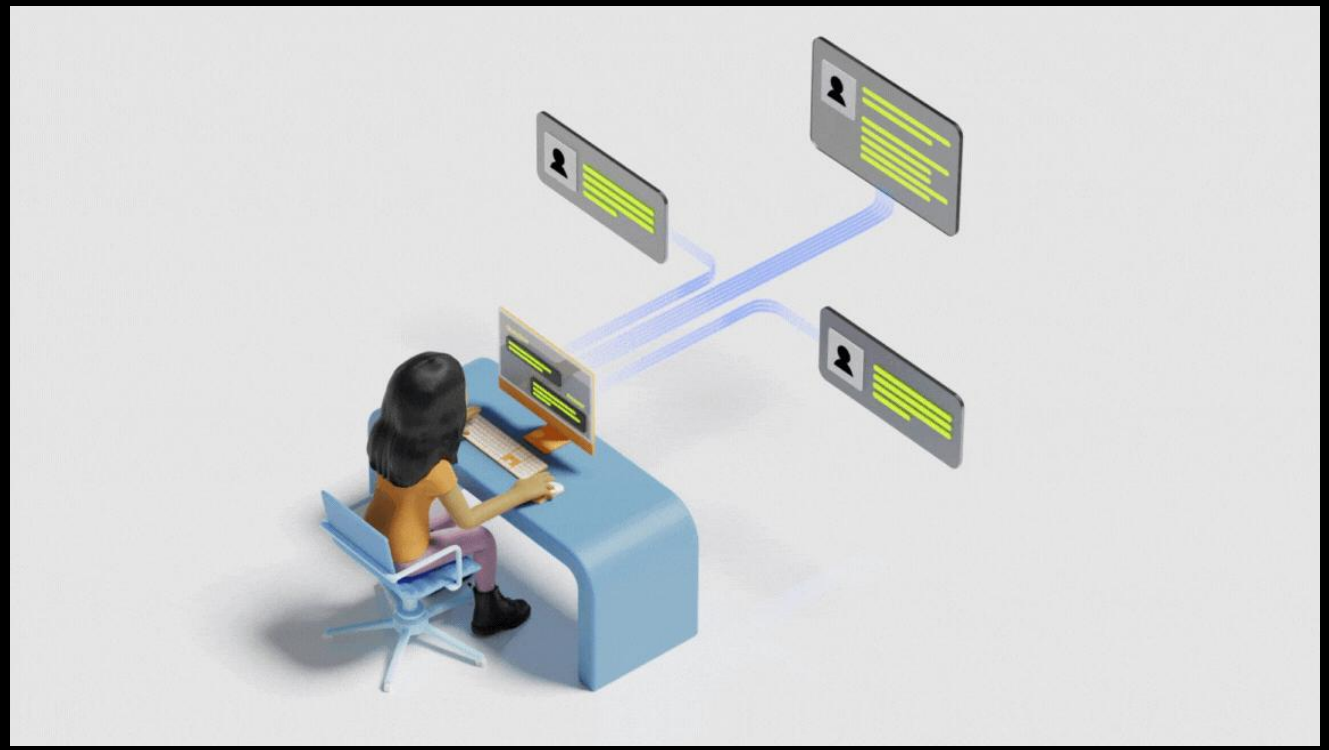
- Les modèles sont entraînés sur des données publiques.
- Les modèles ne sont pas à jour des dernières actualités.
- Les modèles performants sont gourmands en mémoire GPU.
- L'entraînement des LLMs est très coûteux.

Le RAG

Retrieval

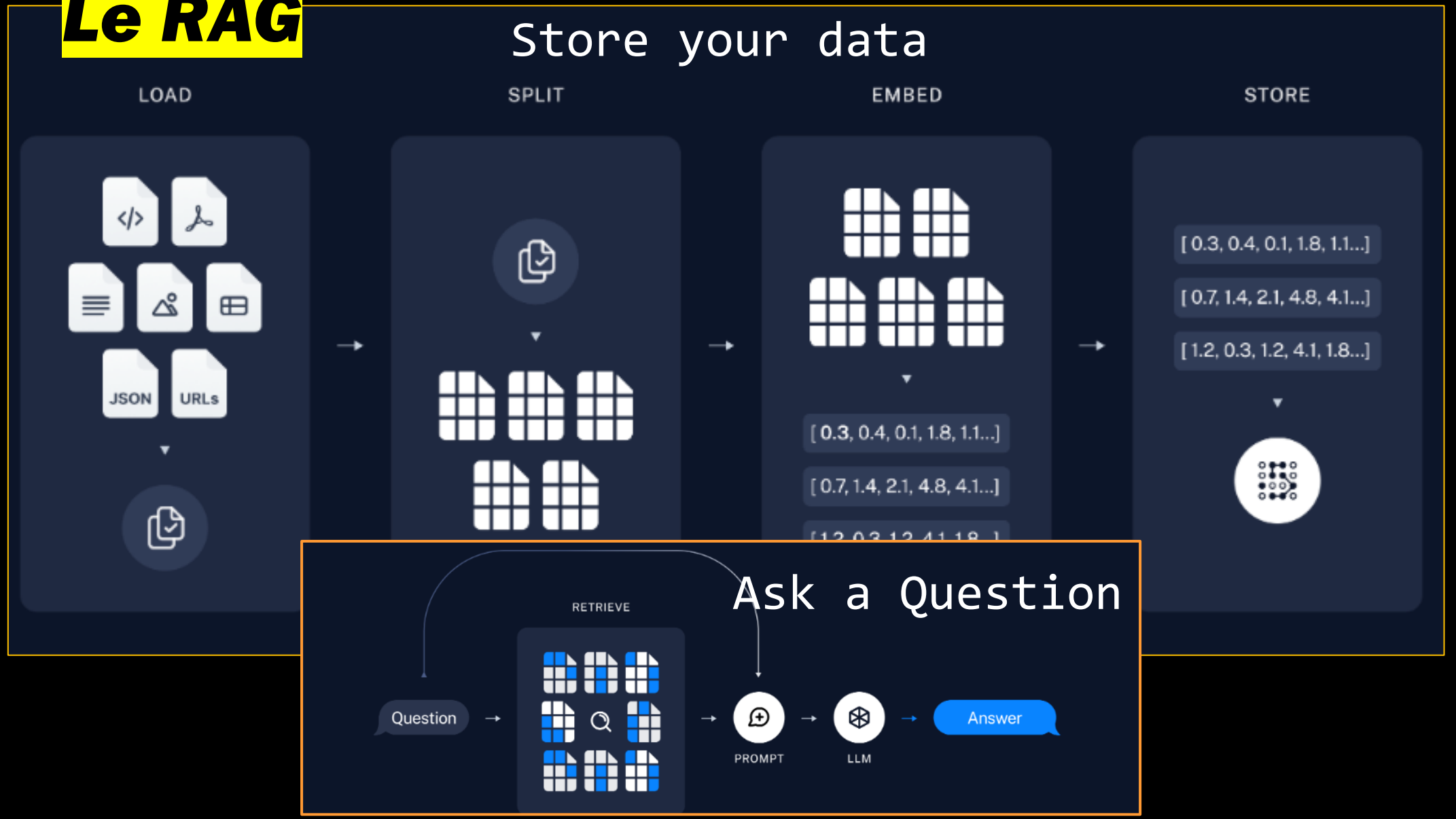
Augmented

Generation



Connecter un LLM à de la donnée

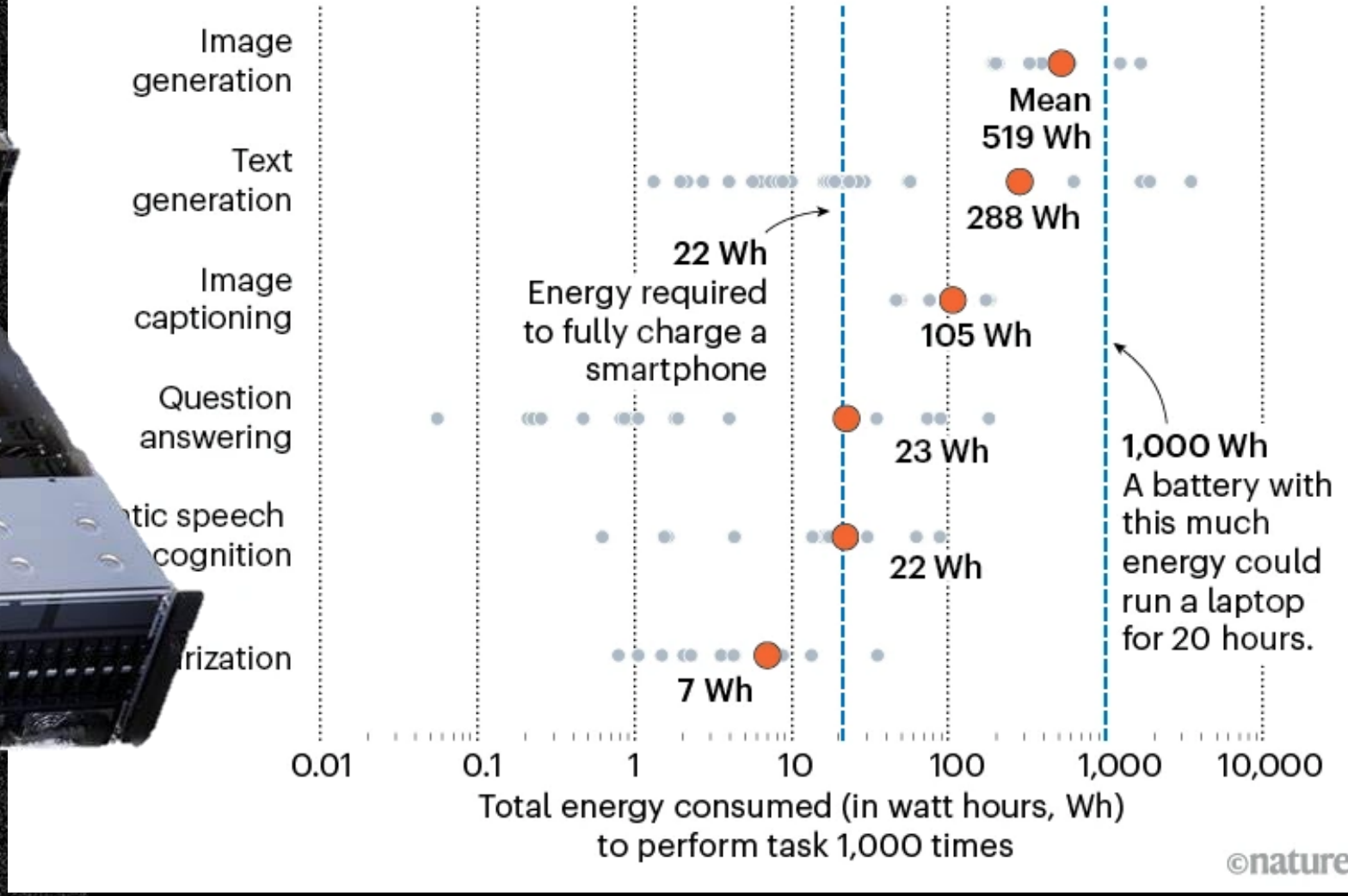
Le RAG


Store your data
LOAD
SPLIT
EMBED
STORE
JSON
URLs
[0.3, 0.4, 0.1, 1.8, 1.1...]
[0.7, 1.4, 2.1, 4.8, 4.1...]
[1.2, 0.3, 1.2, 4.1, 1.8...]
Ask a Question
RETRIEVE
Question
PROMPT
LLM
Answer

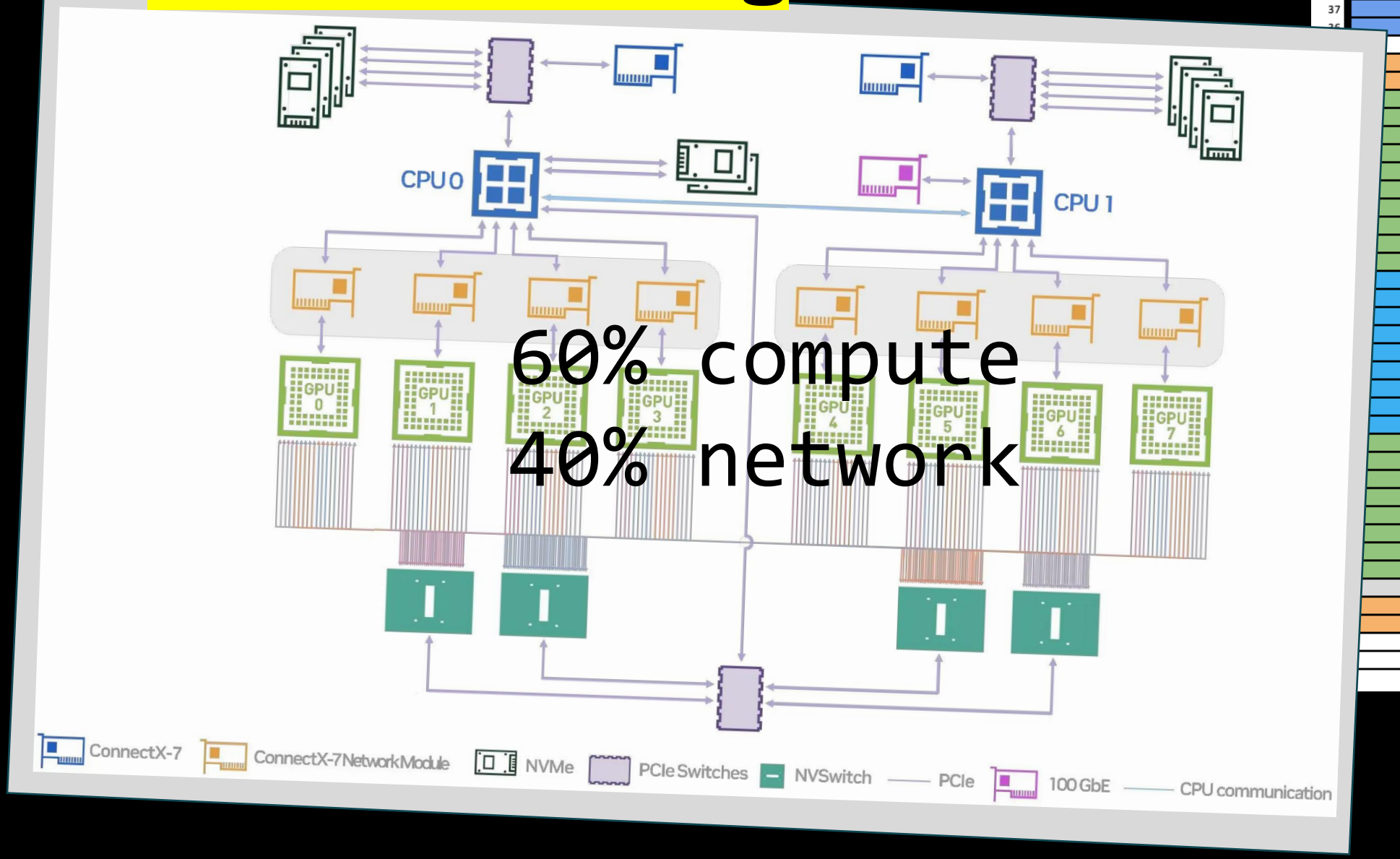
Le hardware

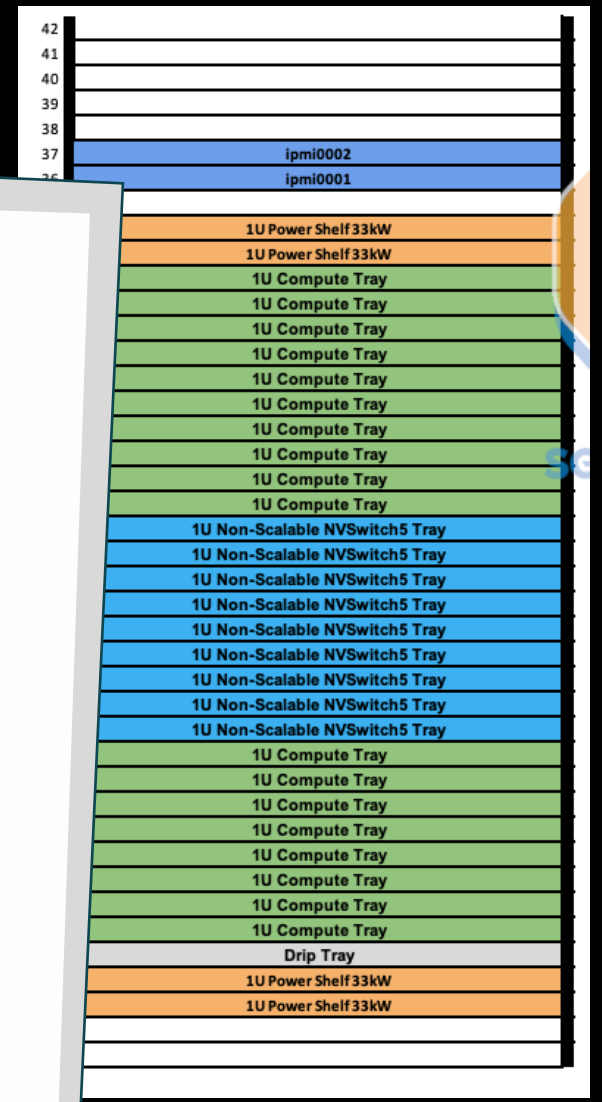
HOW MUCH ENERGY DOES AI USE?

The AI Energy Score project tested dozens of artificial-intelligence models to estimate how much energy they consume when performing various tasks. Plotting the energy required to perform a task 1,000 times shows that energy use varies greatly depending on the task and the model.

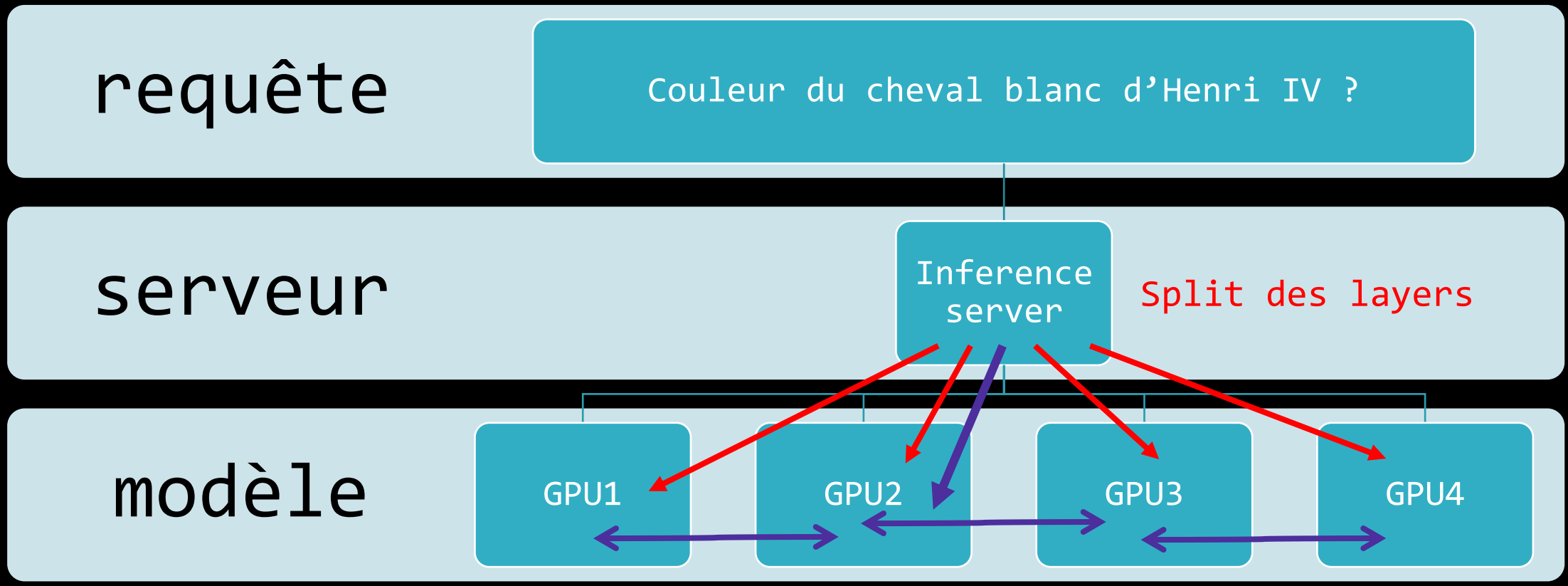


L'infra de training





L'infra d'inférence



Ou mise en commun de la RAM via bus proprio

(ou les deux)

Par ex :

141G x 4 = 564G de capacité de modèle

141G x 4 x 2 = 1128G de capacité de modèle

L’IA est un outil…

Vous n’allez pas être remplacés demain par une IA.
L’IA vous offre des gains de productivité…

Dans nos métiers :

Génération automatisée de configuration et de règles de firewall…

Root-Cause analysis sur des sources uniques ou multiples de logs

Aide au support technique (chatbot, smart search engine, filtrage d’astreinte intelligent, etc)

Le futur

- La fin des GPUs ? Cerebras, Groq, Etched…
- Des modèles plus petits et meilleurs : Google Gemma 3 (27B), Deepseek, Inception AI dLLM…
- Artificial General Intelligence (AGI)

  Level 3 → 90th percentile of skilled adults

  ou Human-Level AI (Yann Le Cun)

  → Years to go…

Et sinon, je fais quoi dans la vie ?

Nous facilitons l'intégration, sans compromis sur la sécurité, d'Intelligence Artificielle, dans vos applications métiers.

Une plateforme clé en main d'IA Européenne, souveraine, non-soumise au Cloud Act.

API OCR+RAG chiffré+LLM plug&play :
Smart Search, Chatbot, etc.

Questions ?

contact@vauban.cloud