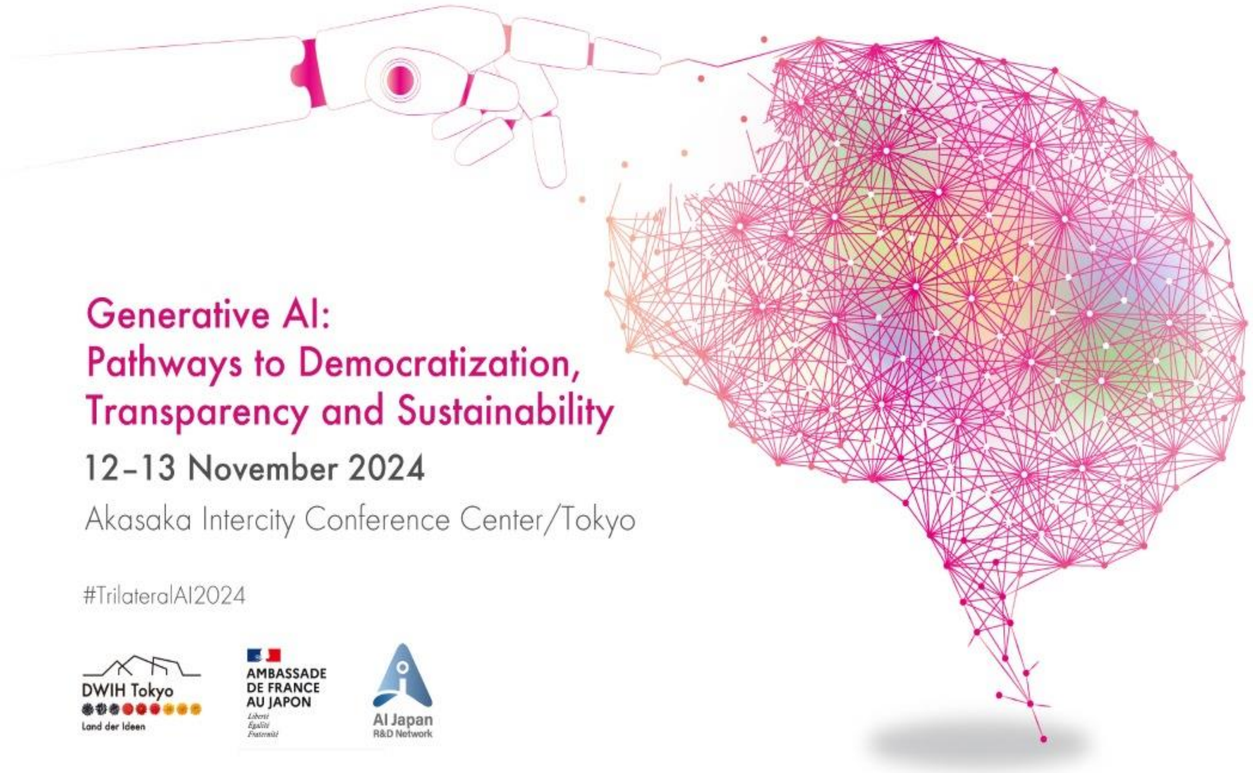
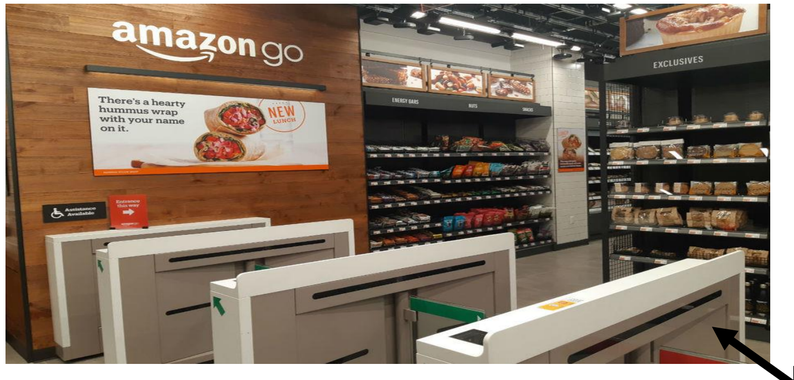
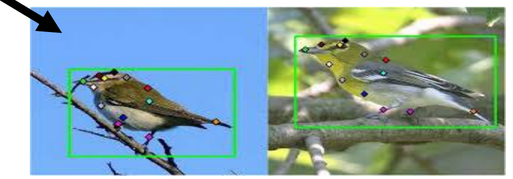
Generative and General-purpose AI : Strenghts and Weaknesses for AI Democratization

Céline Hudelot
Professor
MICS Laboratory
CentraleSupelec



AI is now everywhere !





CentraleSupélec | université PARIS-SACLAY

Democratization of AI

Four meanings and goals [Seger et al, 23]

Democratization of **AI Use**

« Making it easier for a wide range of people to access and use the technology, without coding experience »

Democratization of **AI Development**

« Helping a wider range of people contribute to AI design and development processes »

Democratization of **AI Profit**

« Facilitating the broad and equitable distribution of value accrued to organisations that build and control advanced AI capabilities »

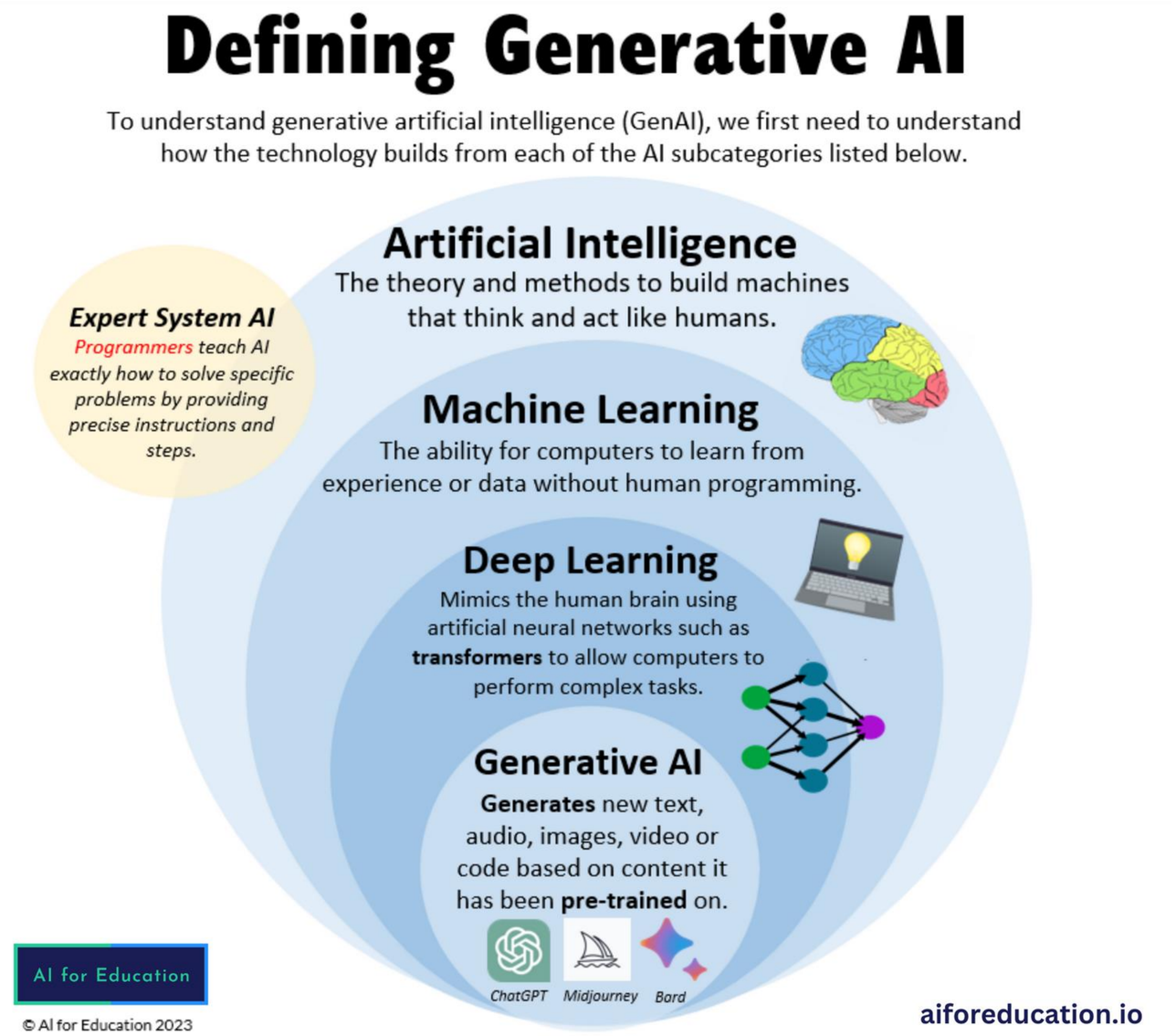
Democratization of **AI Governance**

« Ensure that decisions around questions such as usage, developement, and profits reflect the will and the preference of the society »

Seger et al., “Democratising AI: Multiple Meanings, Goals, and Methods”

Why Generative AI and General-purpose AI can be considered as Pathways to the Democratization of AI Use and AI Development ?

The Rise of the Generative AI Paradigm



The Rise of the Generative AI Paradigm

1949/1950

Turing test

1957-1980

Chomsky grammar

1990 - 2000

Statistical revolution

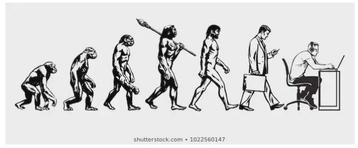
2010 -

Deep Learning

2023 -

ChatGPT - GPT4/5?

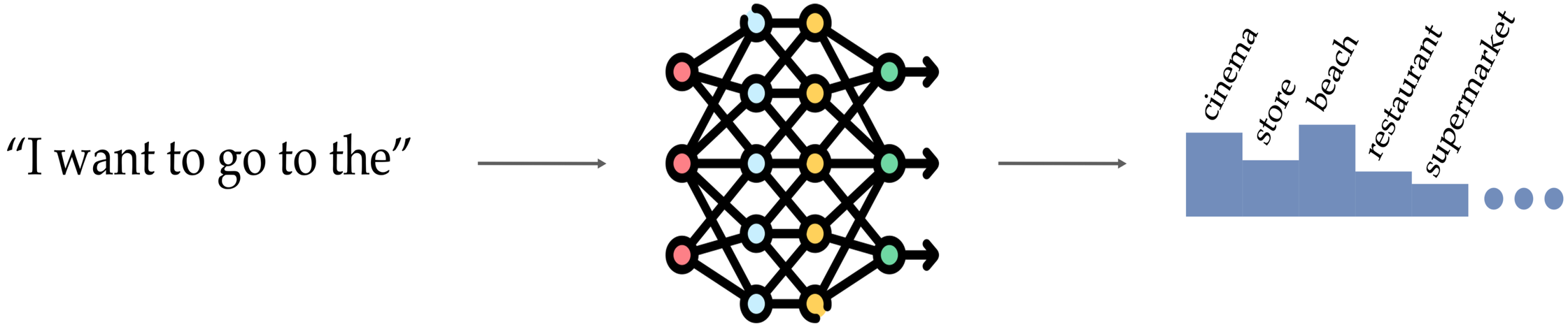
Increase of computing power and available data



Generative AI : Large Language Models

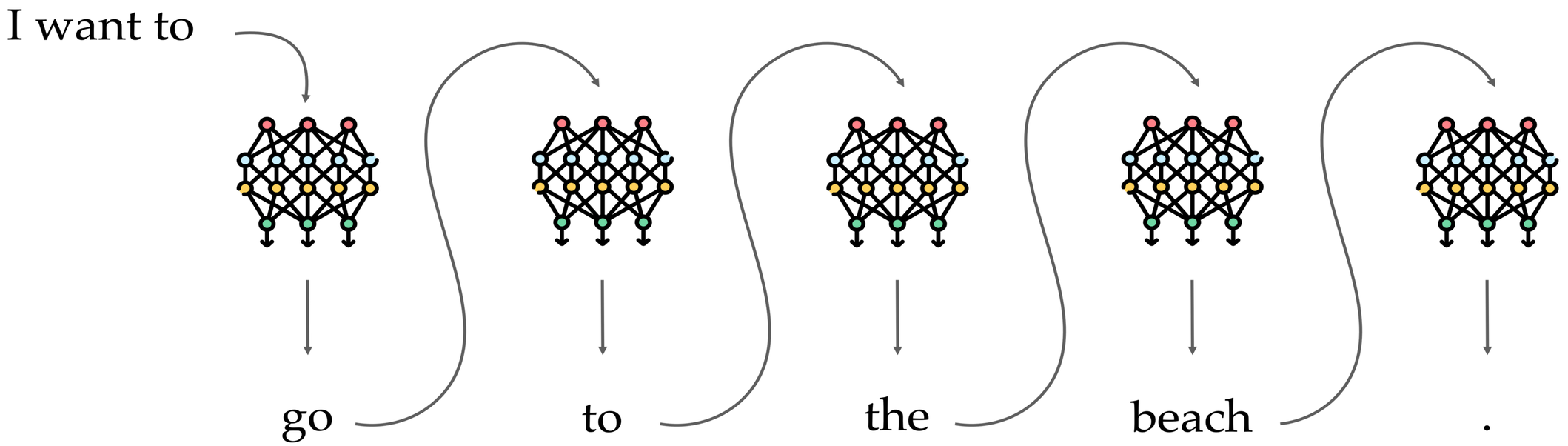
What is a Large Language Model (LLM)?

A LLM is a neural network designed to process and generate human text.



Generative AI : Large Language Models

How does a LLM generate text?

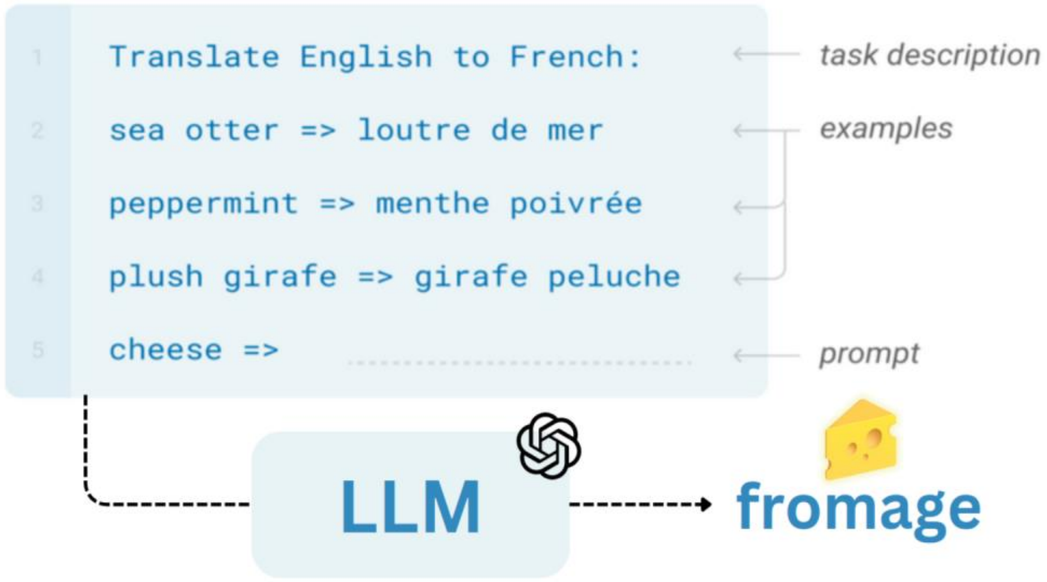


This is a very generic framework. How can we leverage this to perform multiple tasks?

Generative AI : LLMs

How can we leverage LLMs to perform many tasks?

Key ingredients : The era of In-context learning and prompt-engineering !



Brown et al., “Language models are Few-Shot Learners”

Prompt : input to a Generative AI model, that is used to guide its output.

In-context learning : paradigm that allows language models to learn tasks given only a few examples in the form of demonstration.

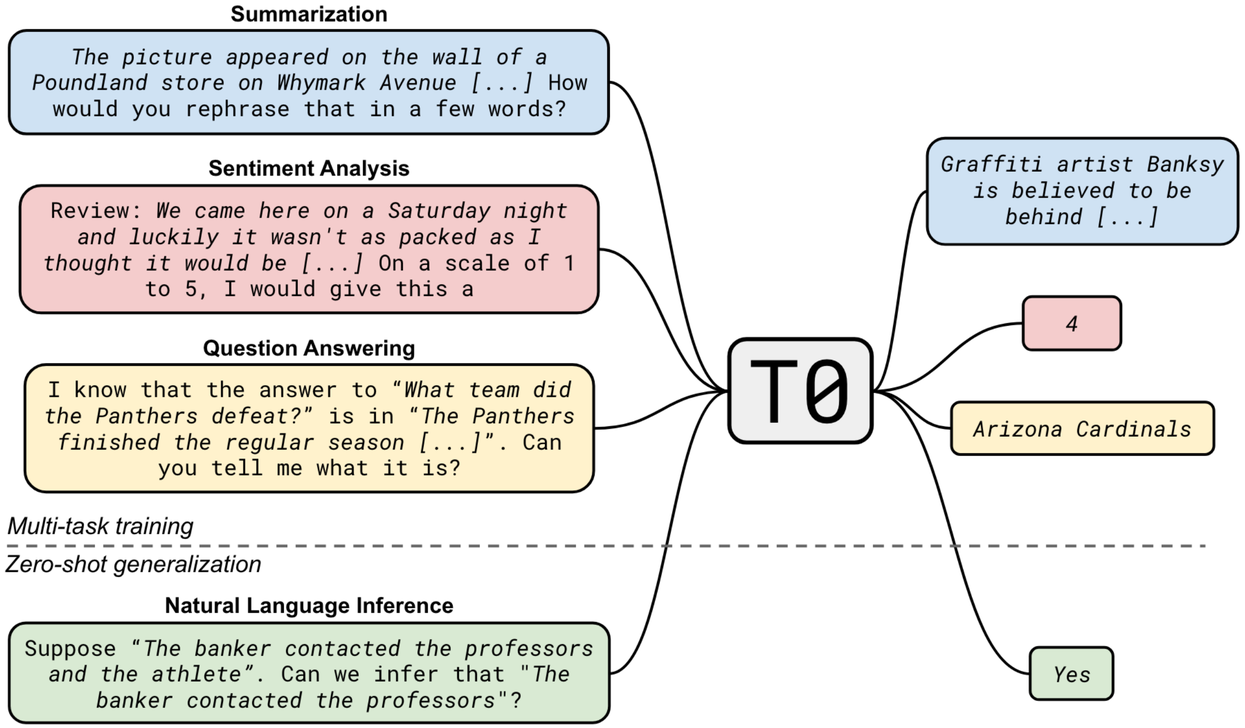
Prompt-engineering : developing and optimizing prompts to efficiently use language models (LMs) for a wide variety of applications and research topics.

Schulhoff et al., “The Prompt Report: A Systematic Survey of Prompting Techniques”

Generative AI : LLMs

How can we leverage LLMs to perform many tasks?

We can prompt them, using **natural language prompts**, for a **large set of tasks**.
We can **dialog** with them !



Prompting and In-context learning enable users and developers to use AI easily, only by prompting and for a large set of tasks.

An important key step to the Democratization of AI Use and AI development.

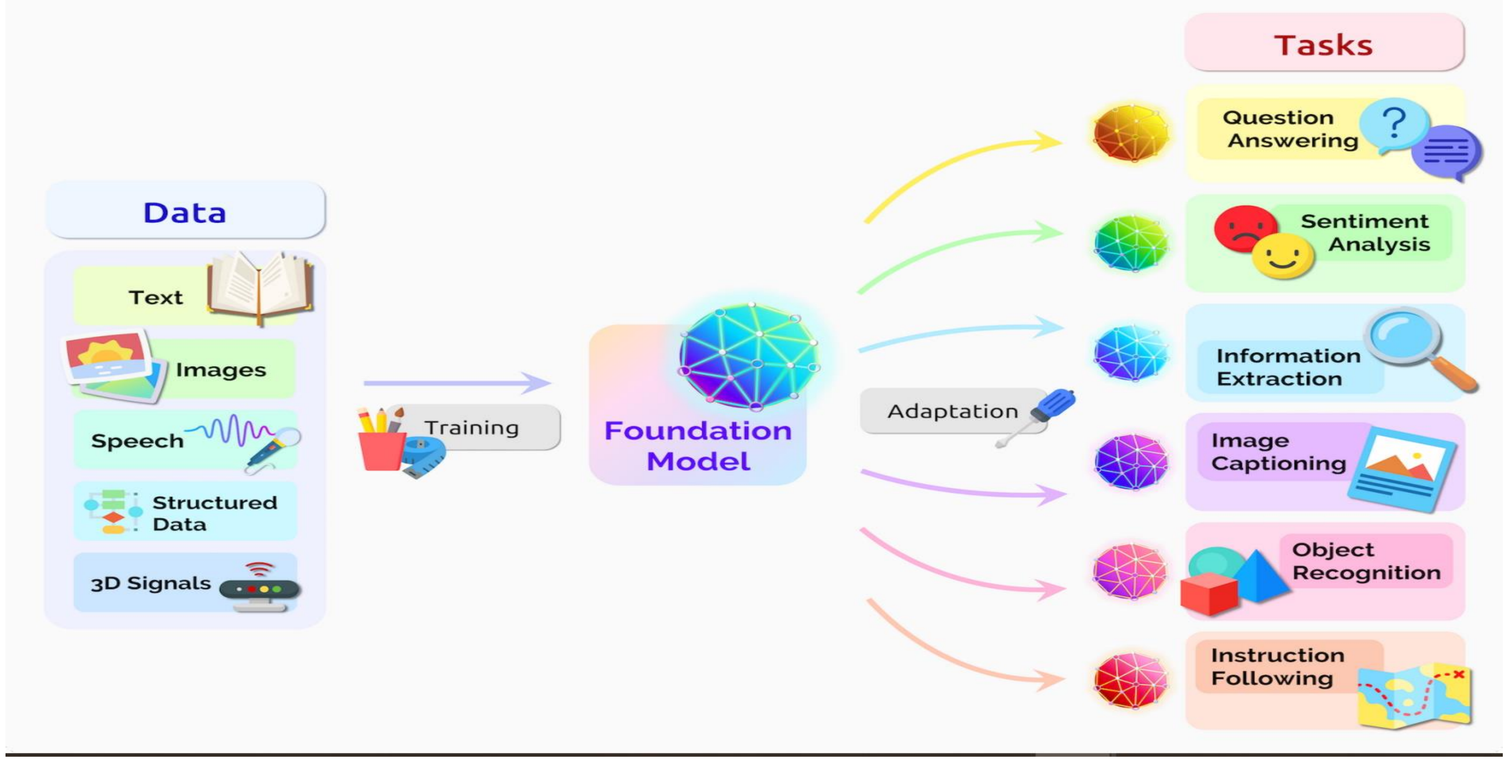
Sahn et al., "Multitask Prompted Training Enable Zero-Shot Task Generalization"

CentraleSupélec | université PARIS-SACLAY

Generative AI : the era of Foundations models

Not only for textual data. LLMs are just Foundations models

A new major paradigm for building AI system : train a model on broad data and adapt it to a wide range of downstrams tasks.



Foundation models are **another key ingredients to the Democratization of AI**

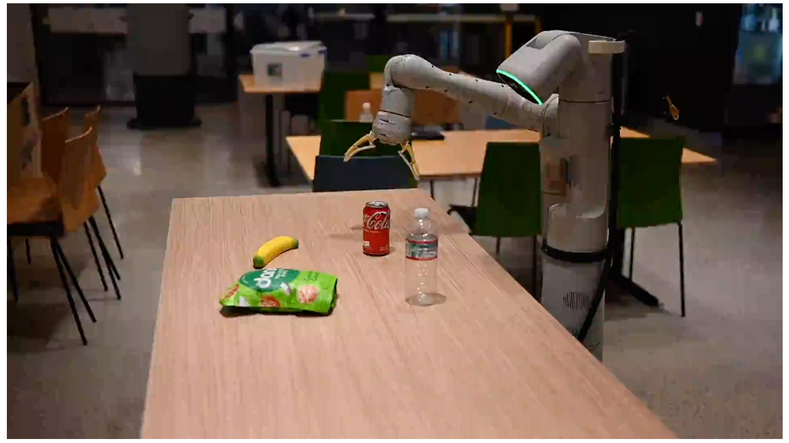
Bommasani et al., "**On the Opportunities and Risks of Foundation Models**"

Generative AI : the era of Foundations models

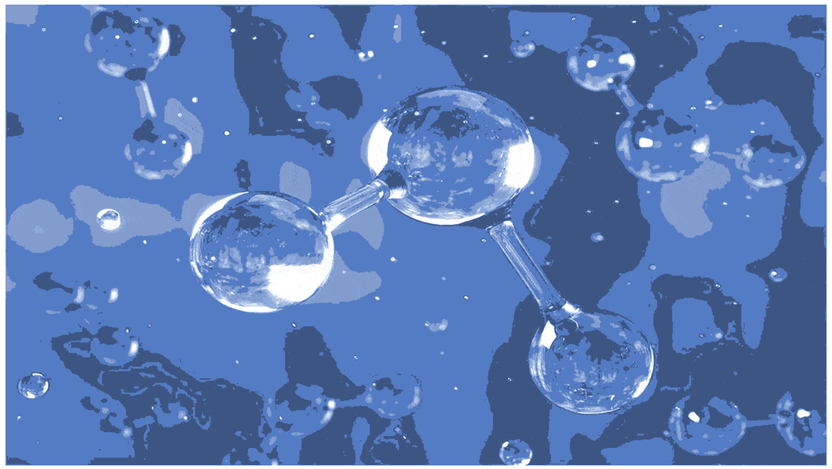
Examples of well-known foundation models

Segment Anything by Meta AI

RT-2: Vision-Language-Action Models by Google Deep Mind



MoLFormers, foundation models on chemicals by IBM

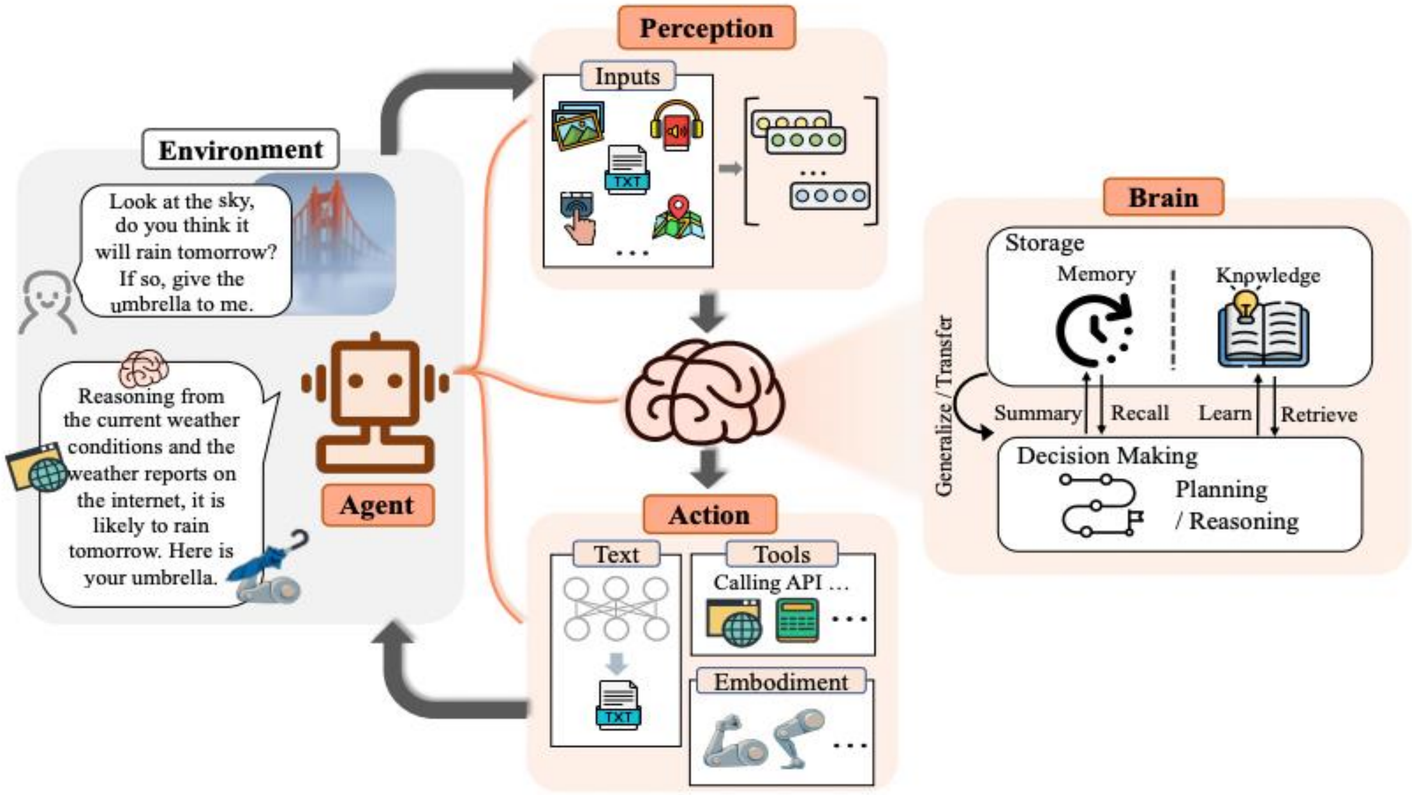


Generative AI : a new era for AI agents.

AI agent : can gather useful sensory information and interact with their environments in meaningful ways

Enable the orchestration of complex pipelines with LLM and foundations models applications executing complex tasks through the use of an architecture combining them with key modules like planning and memory.

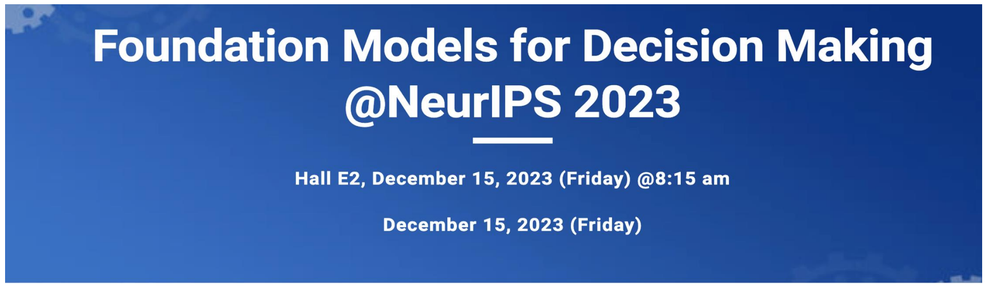
AI Agent paradigm is also a key ingredient to **the Democratization of AI (use and development)**



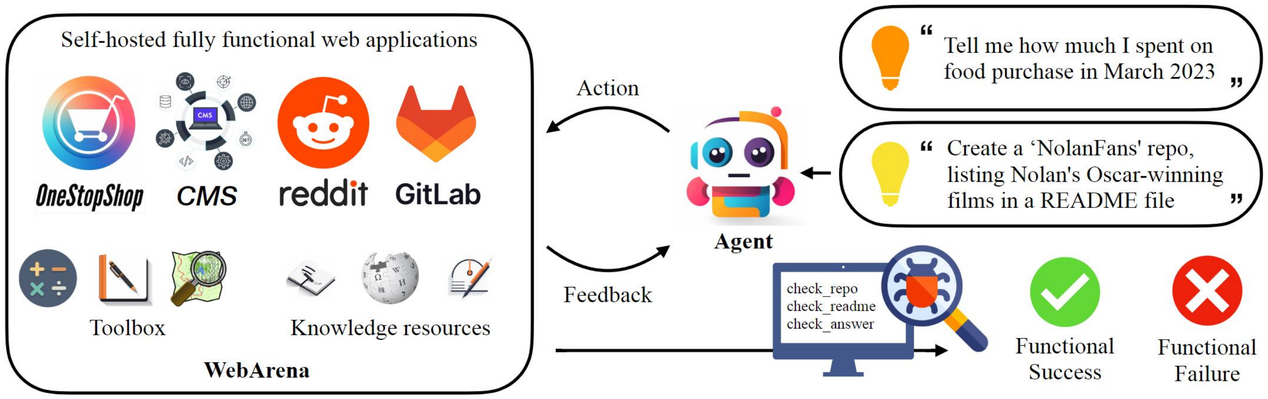
Xi et al., “The Rise and Potential of Large Language Model Based Agents: A Survey”

Generative AI : a new era for AI agents.

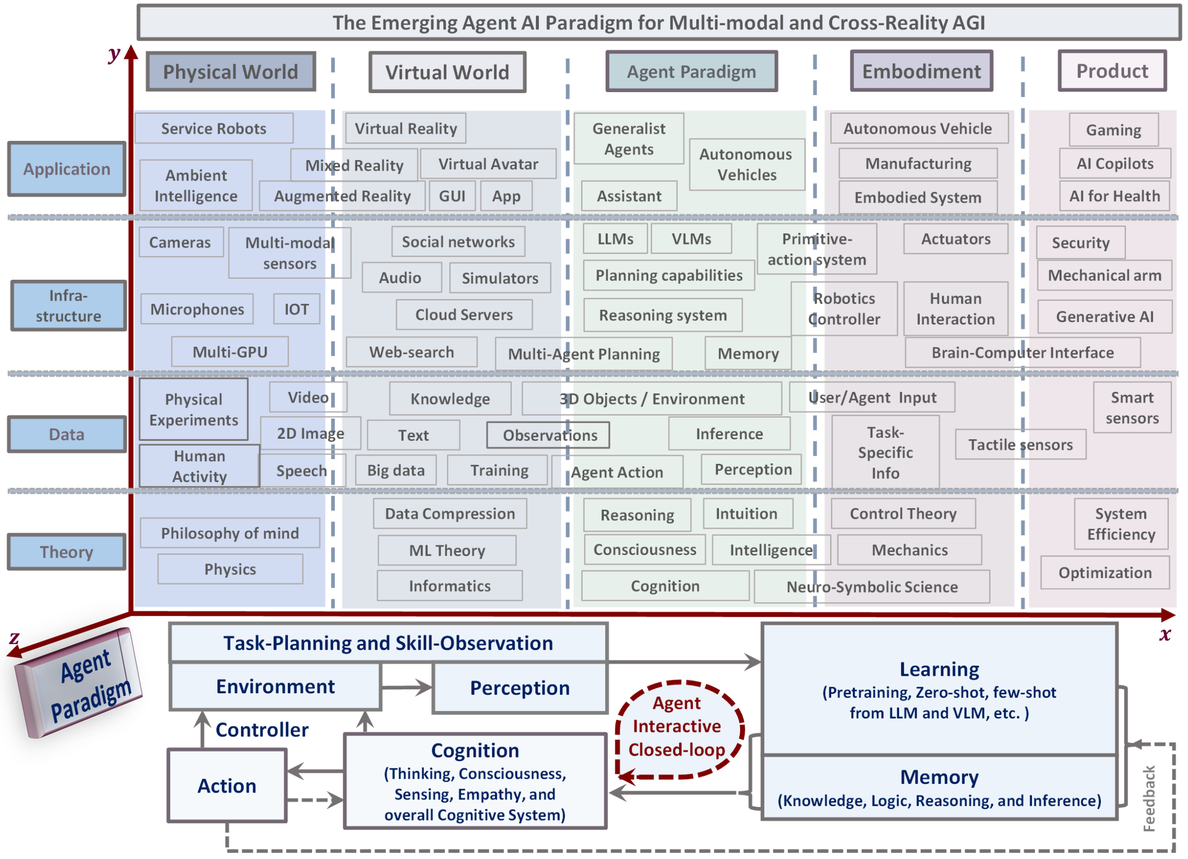
AI agent : A route towards Artificial General Intelligence (AGI)



Foundation Models and Decision Making come together to solve complex tasks at scale.



Zhou et al., "WebArena: A Realistic Web Environment for Building Autonomous Agents”

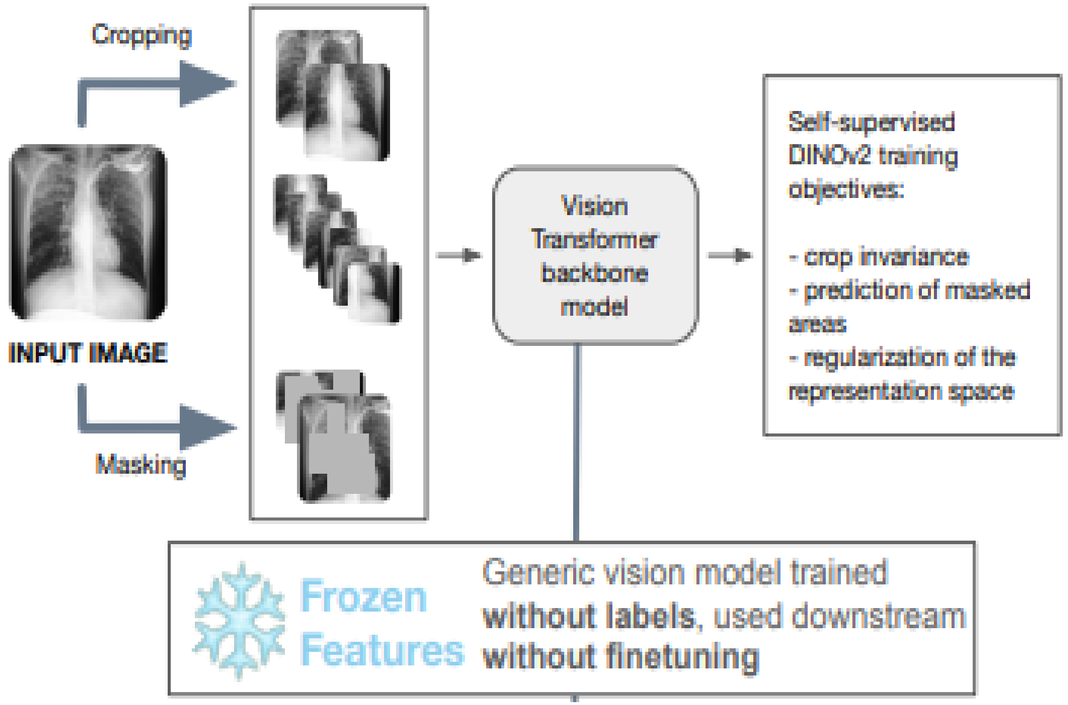


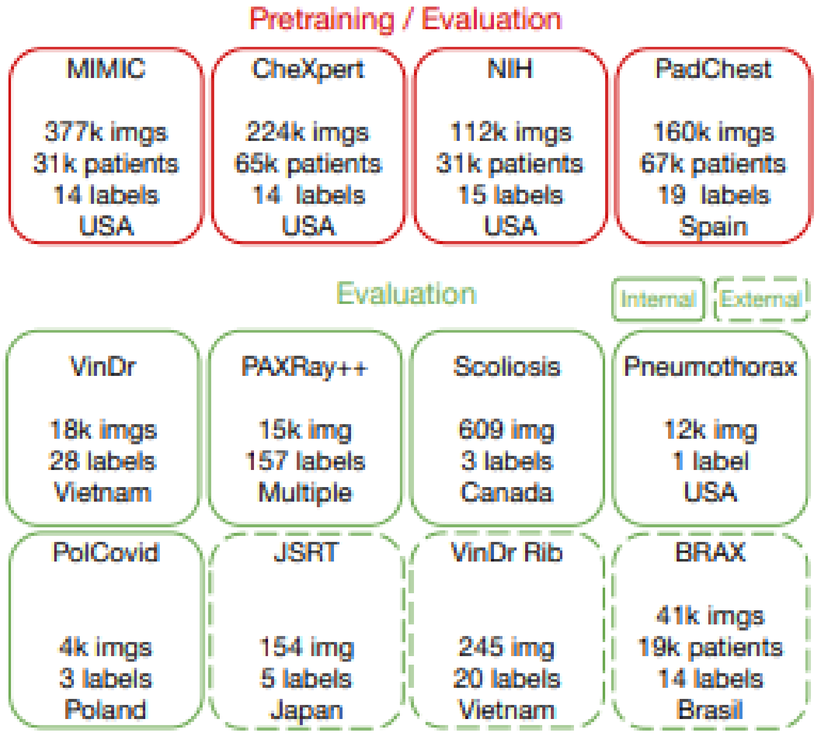
Durante et al., “Agent AI: Surveying the Horizons of Multimodal Interaction ”

CentraleSupélec | université PARIS-SACLAY

Use case study with a health foundation model

RayDINO, a X-ray foundation model

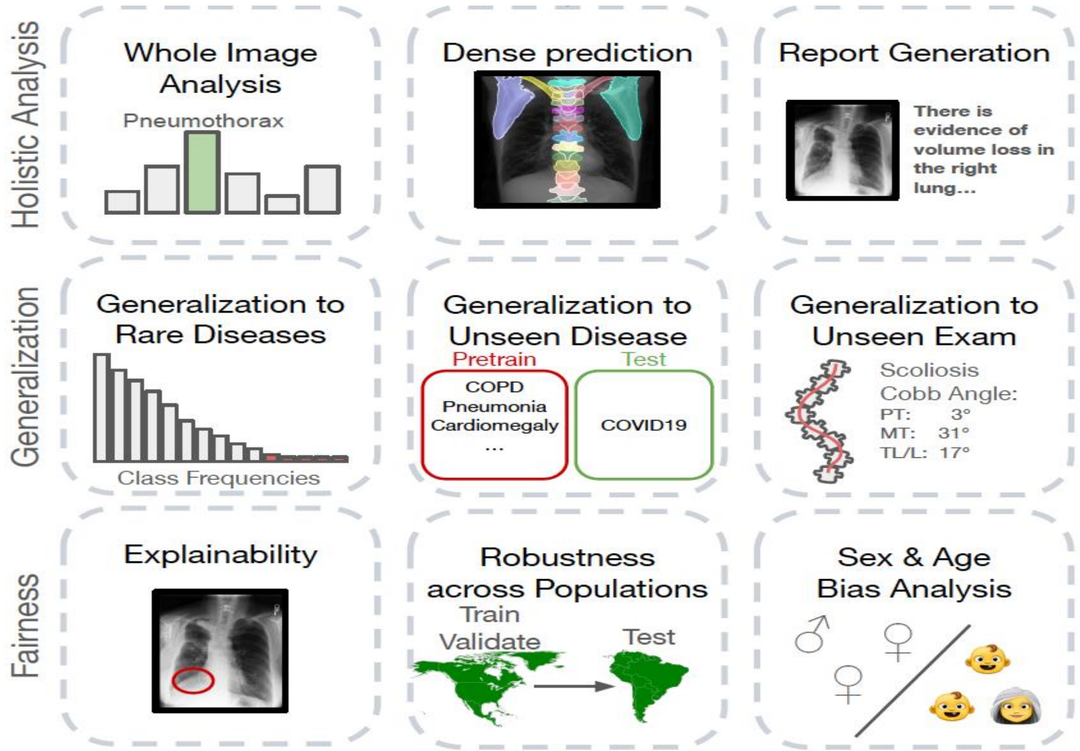


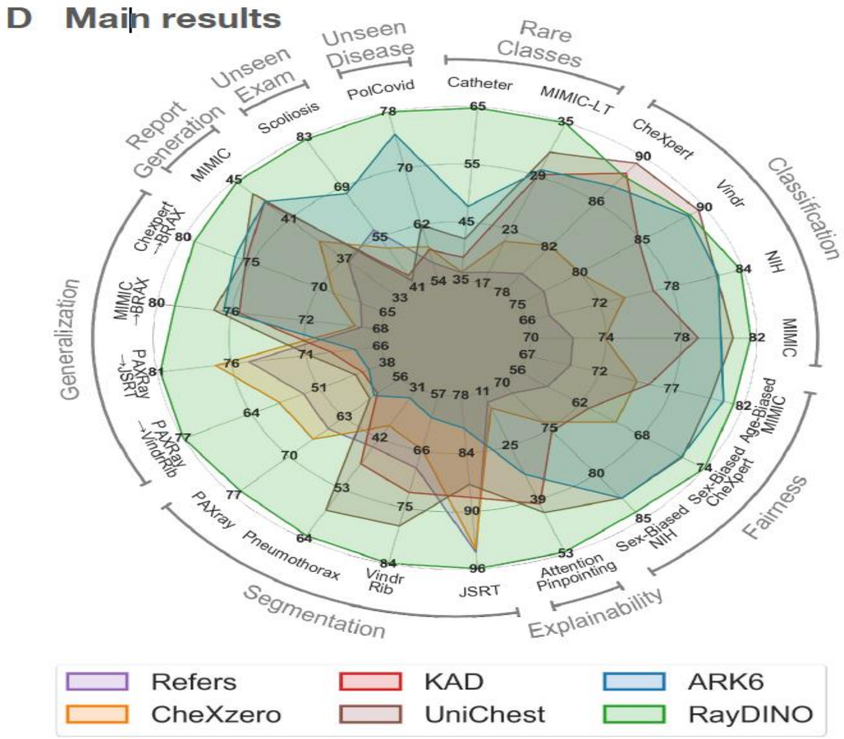


Moutakanni et al., "**Advancing human-centric AI for robust X-ray analysis through holistic self-supervised learning »**
"

Use case study with a health foundation model

RayDINO, a X-ray foundation model





Moutakanni et al., "**Advancing human-centric AI for robust X-ray analysis through holistic self-supervised learning »**

"

Use case study with a health foundation model

RayDINO, a X-ray foundation model

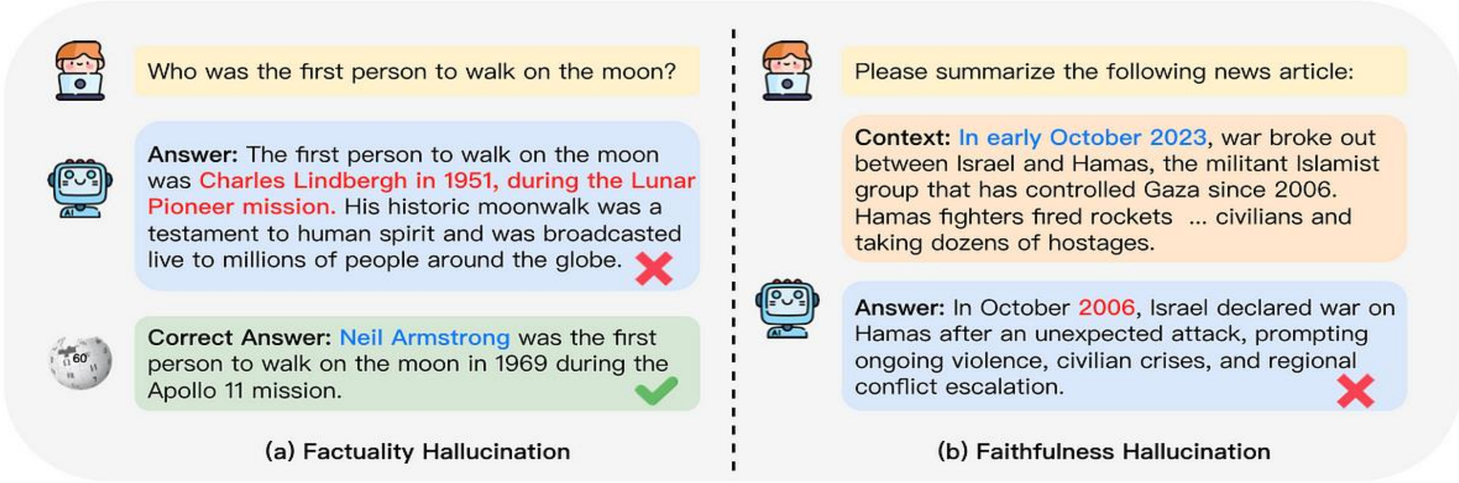
- Unified Approach for Comprehensive Radiology Interpretation: Classification, Segmentation, and Report Generation.
  - ✓ Make easier AI use and development.
- Ability to deal with rare health conditions that cannot be annotated on thousands of images.
  - ✓ A crucial assessment for the practical utility in routine clinical practice.
  - ✓ Potential for large-scale deployment.
- Out-of-Domain Task Performance (POLCOVID Dataset)
  - ✓ Potential for Exploring New Diseases and Exams.
- Robustness to Patient Demographics and their social origins
  - ✓ AI for all
- We also perform gender bias analysis
  - ✓ Ensure fairness and reliability across diverse applications and populations.

Moutakanni et al., “**Advancing human-centric AI for robust X-ray analysis through holistic self-supervised learning »**
”

Generative AI and General-purpose AI can be considered as Pathways to the Democratization of AI Use and AI Development

but

Accessibility and Ease of Use do not mean Adoption.

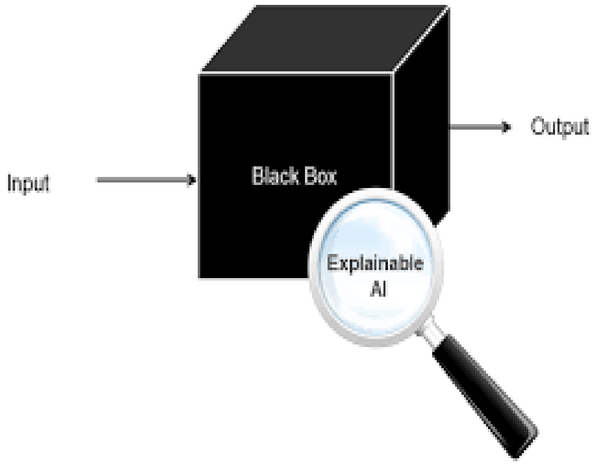


Hallucinations

Press release >

Generative AI: UNESCO study reveals alarming evidence of regressive gender stereotypes

Bias



Explainability, Transparency

CentraleSupélec | université PARIS-SACLAY

Generative AI and General-purpose AI can be considered as Pathways to the Democratization of AI Use and AI Development

but

Accessibility and Ease of Use do not mean Adoption.

Trustworthy AI is needed

Next session : transparency and explainability

Others issues for the Democratization of AI Development

« Helping a wider range of people contribute to AI design and development processes »

Goals

- Accelerate AI Innovation and Progress
- Enabling more people to participate in AI design and development : cater diverse needs and interests
- External evaluation

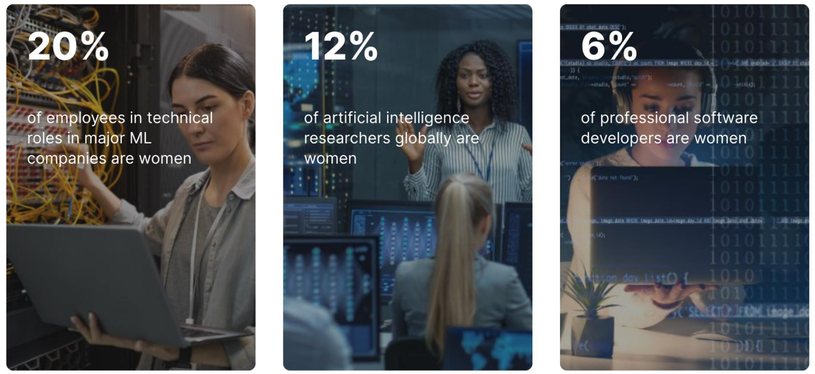
Seger et al., “Democratising AI: Multiple Meanings, Goals, and Methods”

Some issues

Big tech monopole

Huge models downsides

Figures of Women in AI

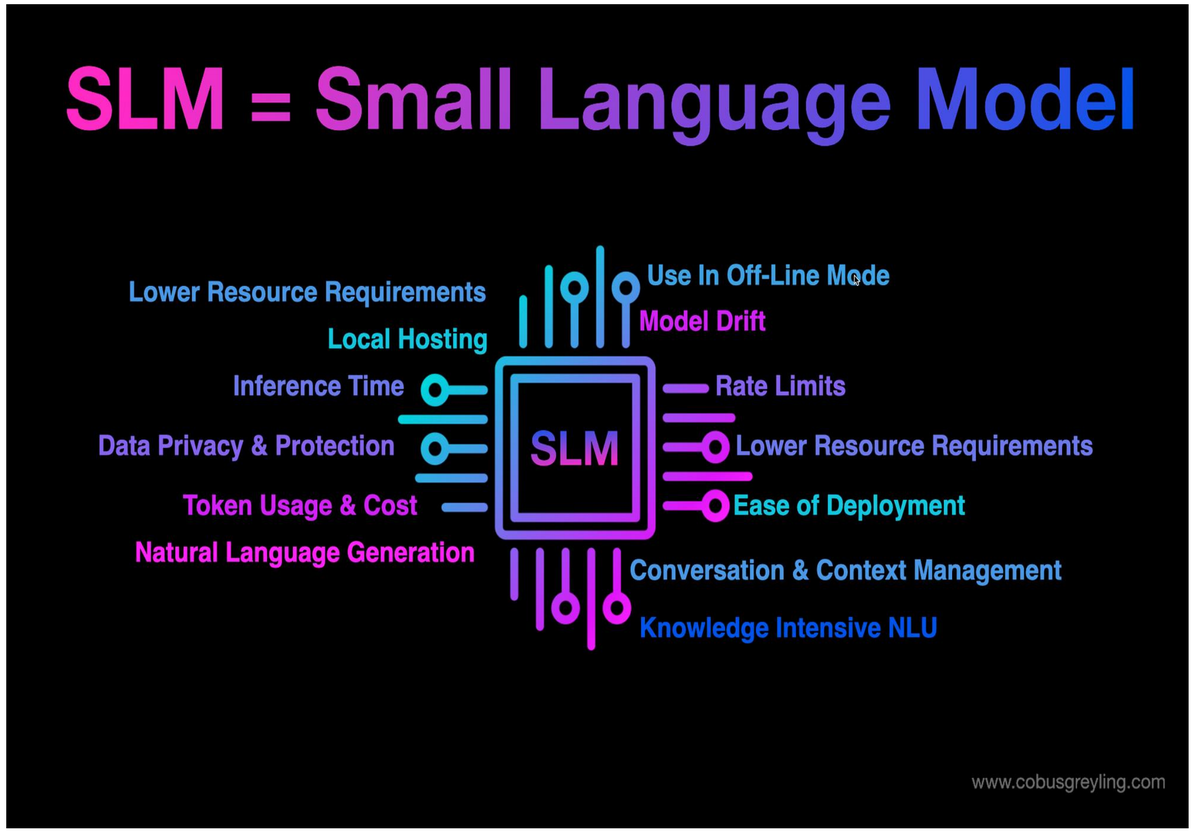


Source : Unesco

Gender bias, narrow demographic actors

CentraleSupélec | université PARIS-SACLAY

Small Language Models (SLMs) at the rescue



Small Languages Models (SLMs) :

- Less than 5 billions parameters
- Retain accuracy and/or adaptability of LLMs while being subject to constraints:
  - Training or inference hardware
  - Data availability
  - Bandwidth
  - Generation time

SLM example : CROISSANT LLM

An Industrial and Academic Partnership for a or a sovereign LLM

CroissantLLM: A Truly Bilingual French-English Language Model

Manuel Faysse[1,5] **Patrick Fernandes**[6,8,11] **Nuno M. Guerreiro**[2,5,6,8]
António Loison[1] Duarte M. Alves[6,8] Caio Corro[9] Nicolas Boizard[4,5]
João Alves[2] Ricardo Rei[2,7,8] Pedro H. Martins[2] Antoni Bigata Casademunt[10]
François Yvon[9] André F.T. Martins[2,6,8] Gautier Viaud[1] Céline Hudelot[5]
Pierre Colombo[3,5]

CentraleSupélec ILLUIN TECHNOLOGY IST TÉCNICO LISBOA Unbabel | cnrs INSTITUT DU DÉVELOPPEMENT ET DES RESSOURCES EN INFORMATIQUE SCIENTIFIQUE

Manuel Faysse
CentraleSupélec, MICS
PhD student

Pierre Colombo
CentraleSupélec, MICS
Assistant Professor

Céline Hudelot
CentraleSupélec, MICS
Professor

Nicolas Boizard
CentraleSupélec, MICS
Diabolocom
PhD Student

CentraleSupélec | université PARIS-SACLAY

A research projet with industrial aims

Research

CroissantLLM is a research project aiming to study how **bilingualism** impacts language model pretraining and performance.

Industry

The final model is designed to be **small enough** to run on local hardware, but **good enough** to run generative tasks that are often reserved to larger models (**inference-optimal** training). It is trained on **permissively licensed data** only.

Open-Source

This project is rooted in open-source, with openly released models, data, code bases and evaluation benchmarks, enabling **researchers and practitioners** to benefit from it.

CentraleSupélec | université PARIS-SACLAY

The Model (Chinchilla tradeoffs)

Generative models are often transformer *decoders* (GPT, Mistral, LLaMa) which performance is closely related to **(1) the # of model parameters**, and **(2) the # of training tokens**.

Training the best model given a fixed compute budget

For a given compute budget, there is an optimal ratio between parameter count and training data size **(~20 for Chinchilla Scaling laws)**

OR

Training the best model of a fixed size

By training longer than the Chinchilla ratio, we continue to improve the model but performance gains are increasingly costly.

We decide to overtrain a small model (1,3B) → 2307 token:param ratio vs 20 Chinchilla-optimal.

Lighter

Faster

Capable

Training Cost

CentraleSupélec | université PARIS-SACLAY

The data

Training language models requires huge amounts of data... In French and under permissive licenses, gathering sufficiently is even more challenging !

Corpus

Culture

Permissively licensed data source

- Public domain books
- Podcasts
- Poetry
- Song lyrics
- Movie subtitles

Business

Industrial and Administrative data

- Legal corpora
- Parliamentary debates
- Administrative Decisions
- Public business documents

Knowledge

Scientific and Factual data

- Encyclopedia
- Textbooks
- Theses abstracts
- Scientific publications

Translations

English and French Parallel data

- Huge quantity of translation pairs sourced from different domains
- Filtered with SOTA quality estimation methods

Internet

Filtered web data

- Web-scale data filtered to obtain high quality French and English texts
- Github Code under open licenses

The data

Training language models requires huge amounts of data... In French and under permissive licenses, gathering sufficiently is even more challenging !

Corpus

Culture
Permissively licensed data source

Business
Industrial and Administrative data

Knowledge
Scientific and Factual data

Translations
English and French Parallel data

Internet
Filtered web data

Identification and scrapping → **Deduplication** → **Filtering** → **Upsampling** → 

- License checks
- Large scale download scripts
- Data quality and diversity checks

CentraleSupélec | université PARIS-SACLAY

The data

Training language models requires huge amounts of data... In French and under permissive licenses, gathering sufficiently is even more challenging !

Corpus

Culture
Permissively licensed data source

Business
Industrial and Administrative data

Knowledge
Scientific and Factual data

Translations
English and French Parallel data

Internet
Filtered web data

→ **Identification and scrapping** → **Deduplication** → **Filtering** → **Upsampling** →

URL deduplication
Exact deduplication
Fuzzy deduplication (MinHash LSH)

CentraleSupélec | université PARIS-SACLAY

The data

Training language models requires huge amounts of data... In French and under permissive licenses, gathering sufficiently is even more challenging !

Corpus

Culture
Permissively licensed data source

Business
Industrial and Administrative data

Knowledge
Scientific and Factual data

Translations
English and French Parallel data

Internet
Filtered web data

→ **Identification and scrapping** → **Deduplication** → **Filtering** → **Upsampling** →

Rule-based filtering
Filtering of Toxic, Violent or Political content
Perplexity Filtering
(Data Quality)

CentraleSupélec | université PARIS-SACLAY

The data

Training language models requires huge amounts of data... In French and under permissive licenses, gathering sufficiently is even more challenging !

Corpus

Culture
Permissively licensed data source

Business
Industrial and Administrative data

Knowledge
Scientific and Factual data

Translations
English and French Parallel data

Internet
Filtered web data

Identification and scrapping → Deduplication → Filtering → Upsampling → 

Final data mix

Upsampling FR to obtain a balanced corpus FR / GB

CentraleSupélec | université PARIS-SACLAY

Transparency & Open-Source

Project rooted in transparency, to serve as a useful resource for industrial practitioners and researchers !

Documented training process from beginning to end

Openly available:

- Training corpus
- Model checkpoints
- Evaluation Benchmarks
- Code bases

No usage restrictions (MIT)

Foundation Model Transparency Index Total Scores

Stanford University

Company		Score
CroissantLLM		80%
Meta	Llama 2	54%
BigScience	BLOOMZ	53%
OpenAI	GPT-4	48%
stability.ai	Stable Diffusion 2	47%
Google	PaLM 2	40%
ANTHROPIC	Claude 2	36%
cohere	Command	34%
AI21labs	Jurassic-2	25%
Inflection	Inflection-1	21%
amazon	Titan Text	12%

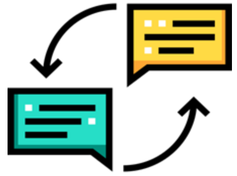
CentraleSupélec | université PARIS-SACLAY

Croissant LLM applications

Specific tasks

- Writing assistance
- Summarization
- Orthographic correction
- Prompt Compression (RAG)
- Retrieval

Translation



CroissantLLM is the best decoder of its size in translation FR / GB matching the performance of Mistral and Llama models of 10 x the size.

Phone & CPU

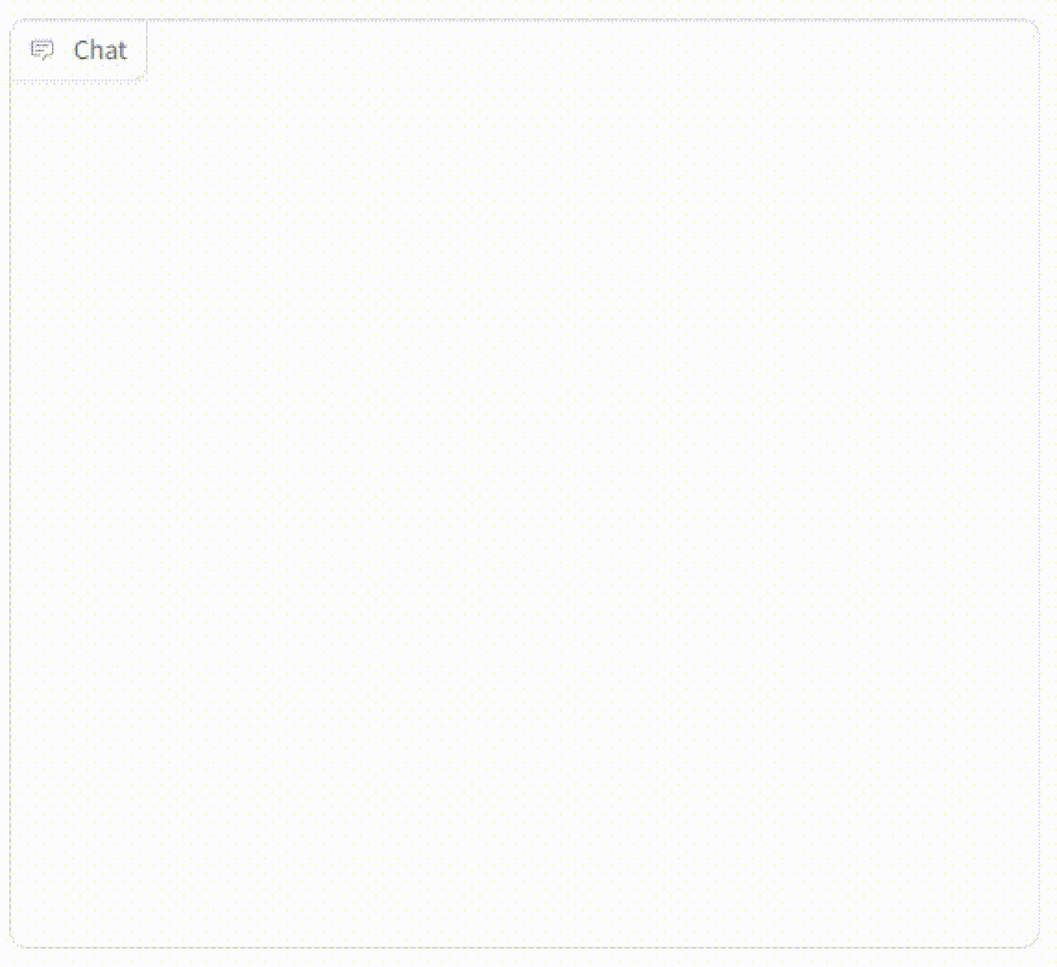
Unmatched performance in FR amongst models lightweight enough to run on phones and local hardware.

Frugality

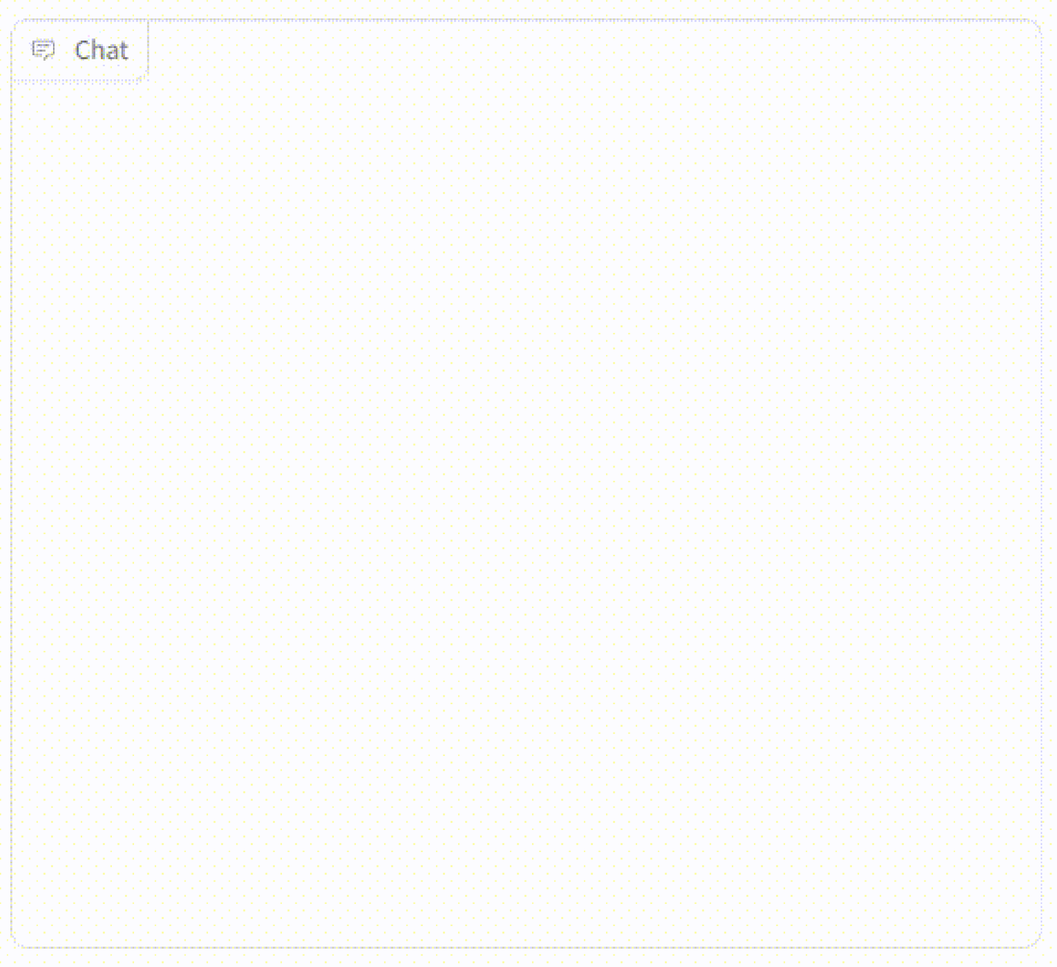
The lightweight model reduces costs and energy requirements.

CentraleSupélec | université PARIS-SACLAY

Generation speed



More than **30 tokens per second on CPU**



More than 120 **tokens per second on lower end GPU (T4)**

A small recap

- ✓ Generative AI, foundations models and Agent-based AI are important enablers of the democratization of AI.
- ✓ They can be used, with high performance, in a wide range of applications and by a wide range of users (**throught prompting or with small task-level adapters**)

BUT

Many obstacles need to be overcome to achieve true democratisation:

- **Trustworthy** Generative AI : explainable, fair, interpretable, robust, transparent, safe and secure.
- Encourage initiatives to avoid development being monopolised by a few large players (sovereignty at stake).
- LLMs and foundation models are not the only paths : SLMs, RAG, specialized models…
- Need of Governance and Education

The clusters

We presented projects involving a lot of compute

ADASTRA

JEANZAY

The Team

Pierre Colombo

Manuel Faysse

Nicolas Boizard

Hippolyte Gisserot-Boukhlef

Maria Vakalopoulou

Theo Moutakanni

Céline Hudelot

Duarte Alves

Caio Corro

Nuno Guerreiro

André Martins

And many others……