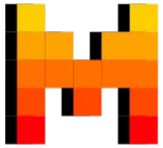
MISTRAL AI_

Frontier AI in your hands

March 2024

A brief history of “large” language models



A LLM that you can tame into a useful assistant (i.e, with higher than 60% MMLU)

2019: GPT-3 (175B)

2020: Gopher (280B)

2021: Megatron-Turing NLG (530B)

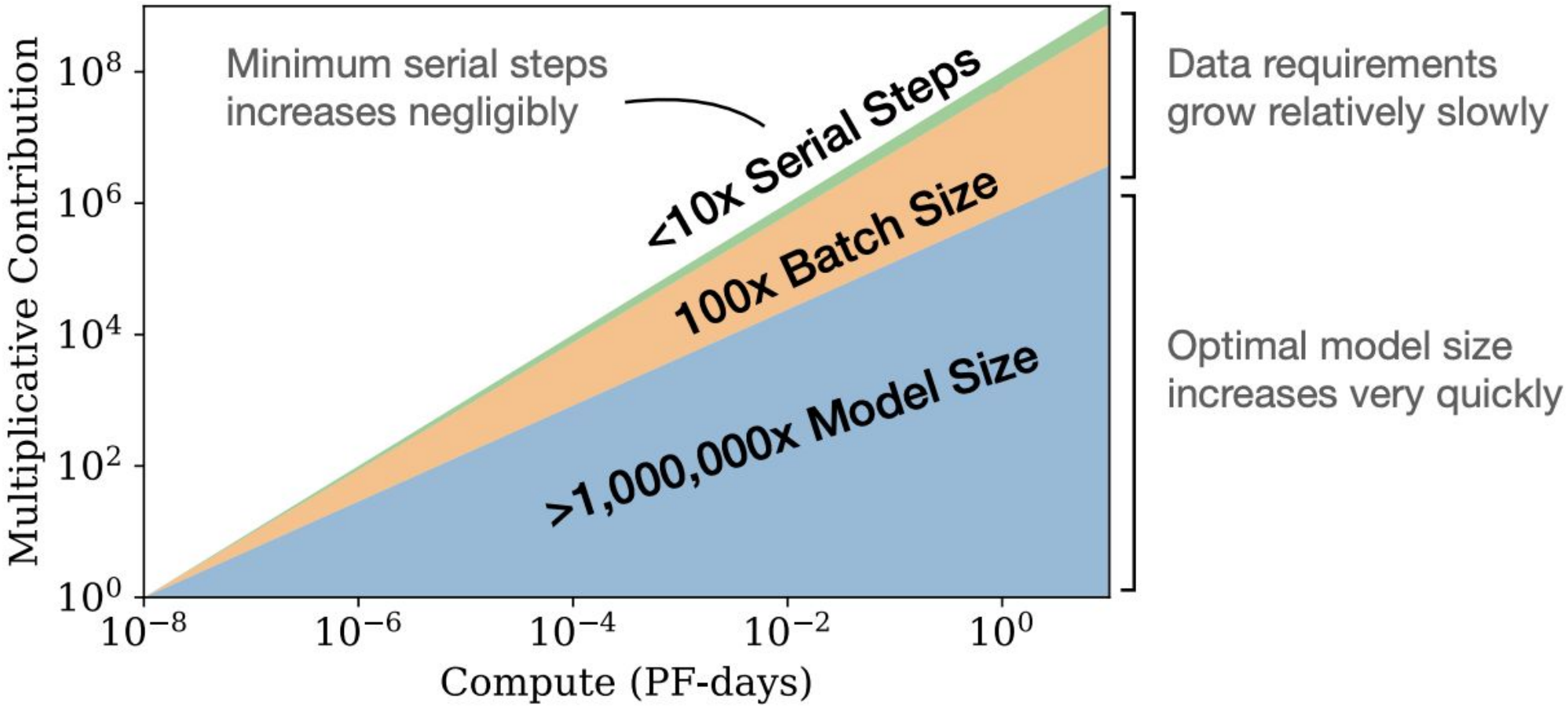
2022: Chinchilla (67B)

2023: Llama 2 (13B), GPT-4 (?), Claude 1-2 (?), Gemini (?)

2023: Mistral 7B (7B), Mistral 8x7B (12B active)

This is a definitely a biased slide, but there is a trend

Make those models infinitely big



Given 10x training budget, x5 on model size, x2 on data

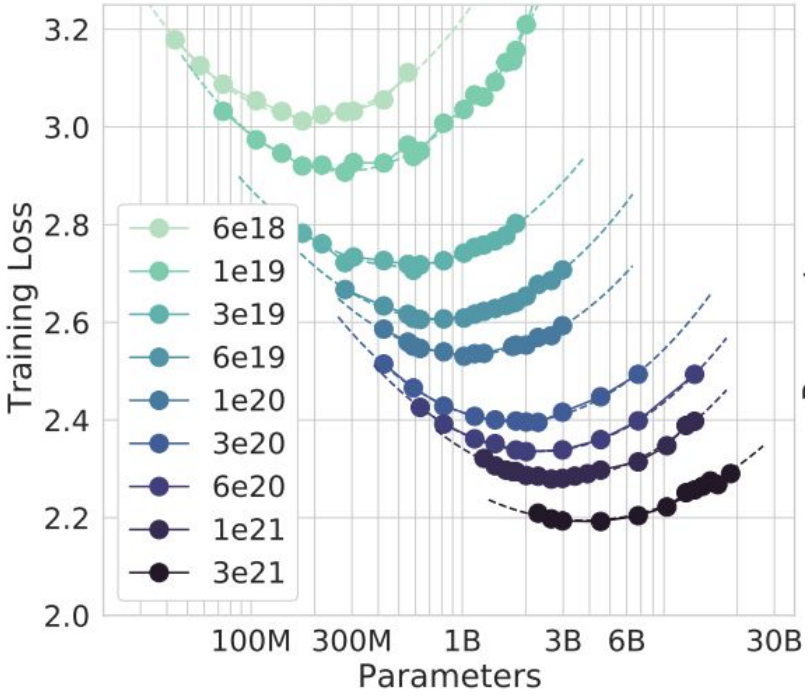
Kaplan et al., 2020

A little rodent noticed a problem

Hoffman et al., 2022

Increase the dataset size ! ~~300B tokens~~

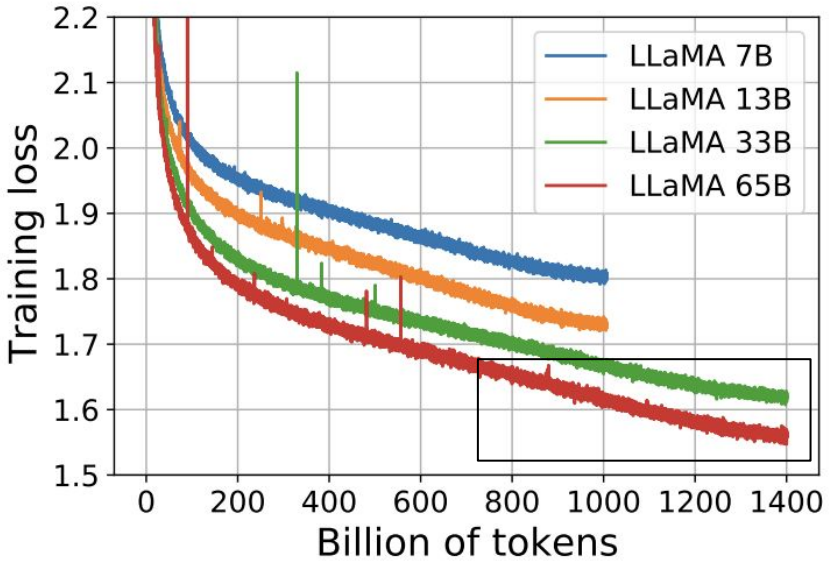
Find the best model size for a given compute budget



Given 10x training budget, x3.1 on model size, x3.1 on data

Wait, we can actually train on more tokens ?

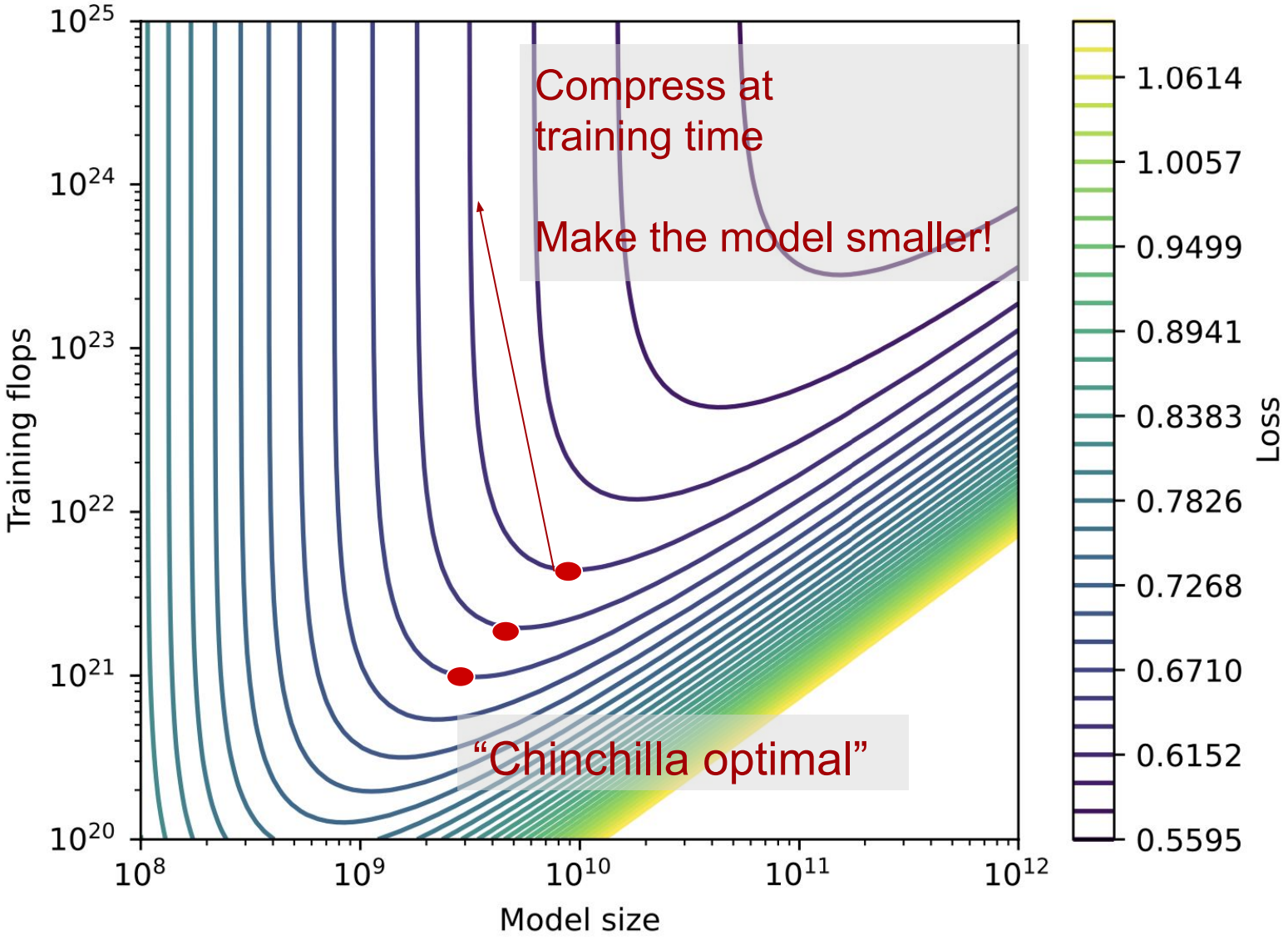
Touvron et al., 2023



“Overtraining”

Given 10x training budget, consider your inference budget

The cost of compression

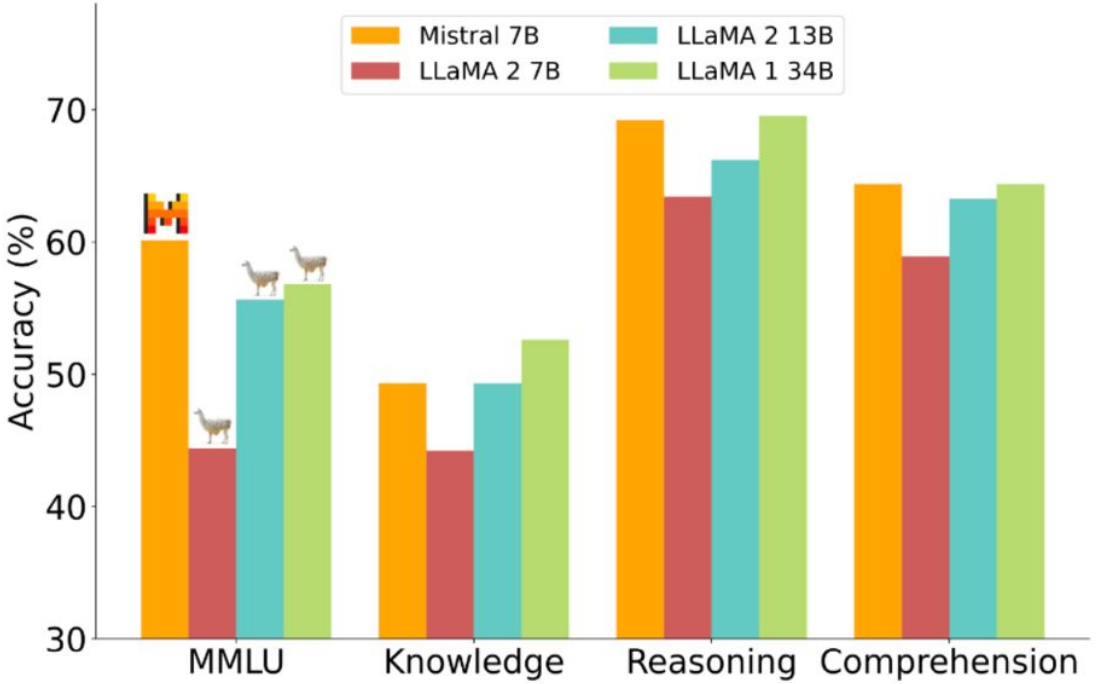


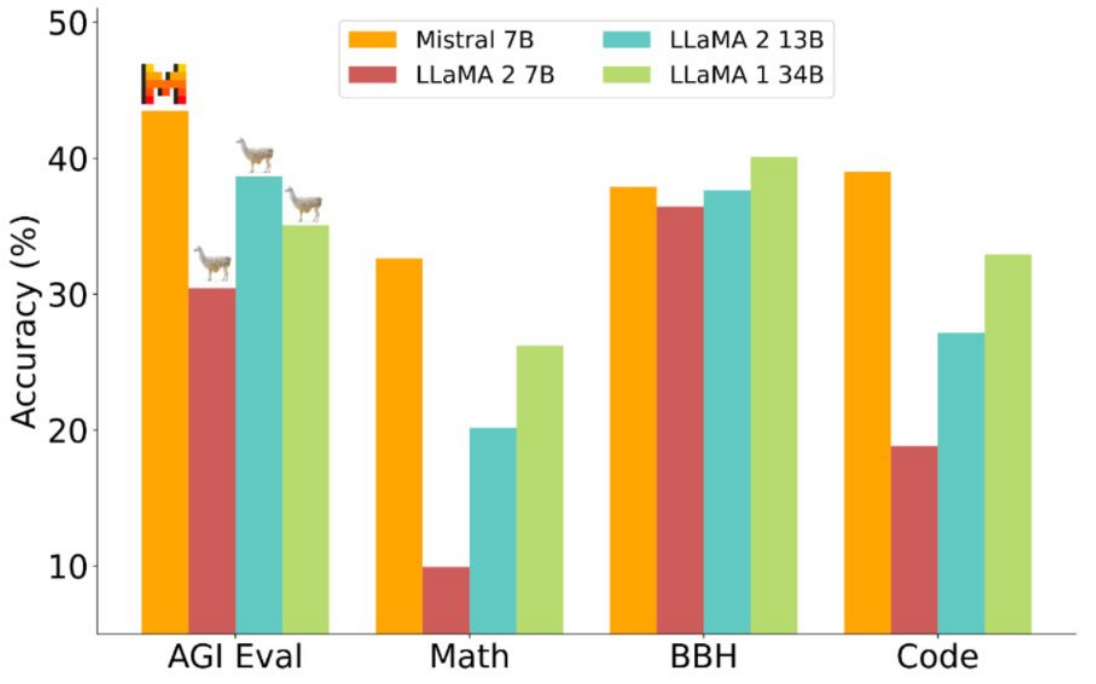
A functional approximation and a stochastic approximation

$$\hat{L}(N, D) \triangleq E + \frac{A}{N^{\alpha}} + \frac{B}{D^{\beta}}.$$

More weights! More tokens!

Let’s put that into practice





Meet Mistral AI – We bring AI into everybody's hands

Product releases
Strategic milestones

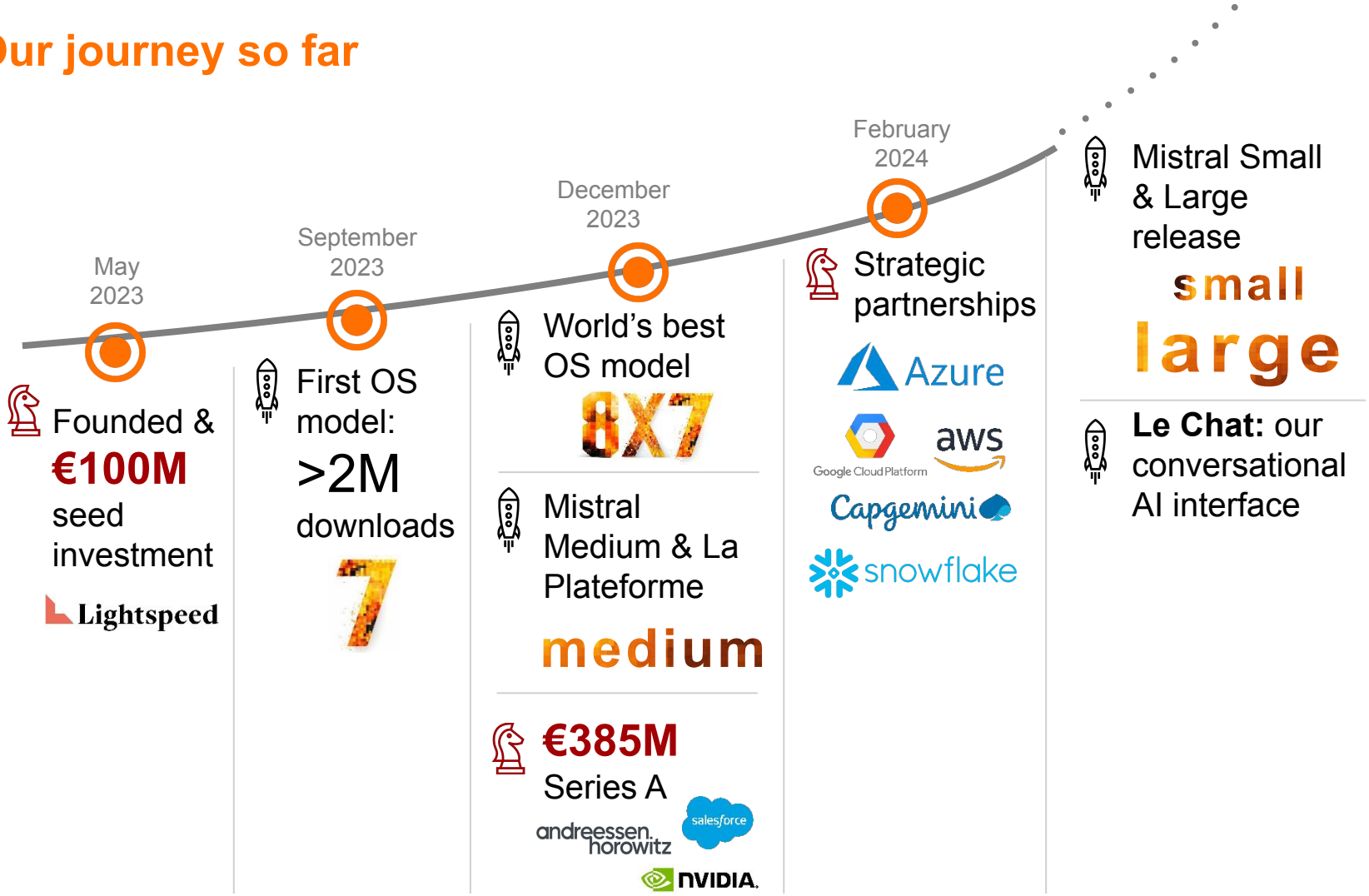
Who we are

Paris-based team of **>35 scientists and entrepreneurs**

Build the world's **most efficient LLMs and tools**

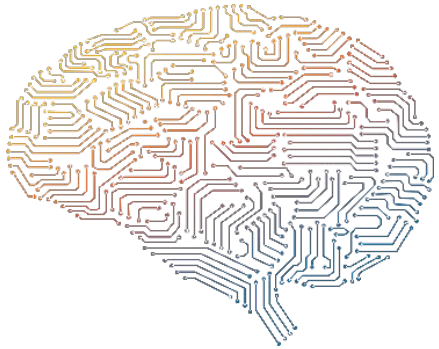
For **developers** and **businesses** all around

Our journey so far



We provide state-of-the-art portable generative AI models

Highest performance & efficiency



Scale without compromising on performance

Portable technology

Deploy where you want, matching the requirements of each use case

Client-side customization

Get the most value from your enterprise assets and gain a competitive edge

We offer 5 models for all use cases and business needs

Our open-source models

Our enterprise-grade models

PERFORMANCE

7

Our **first model,** fits on one GPU

8X7

Our SMoE model: only 14B inference

small

Best for **low latency** use cases

large

Our flagship model, for your **most sophisticated needs**

Adoption: >6M downloads and numerous derivatives

Apache 2.0 license

Capabilities: Multi-lingual, Function calling, JSON mode, performant on RAG use cases, concise, 32k context window, etc.

+ embed

Experiment and alter

Deploy and scale

Our models can be deployed wherever you want

La Plateforme: our API endpoint

API

La Plateforme

Public cloud

Model as a service
(managed API)

AI Studio

Bedrock

Model Garden

Your infrastructure

Self-deployment

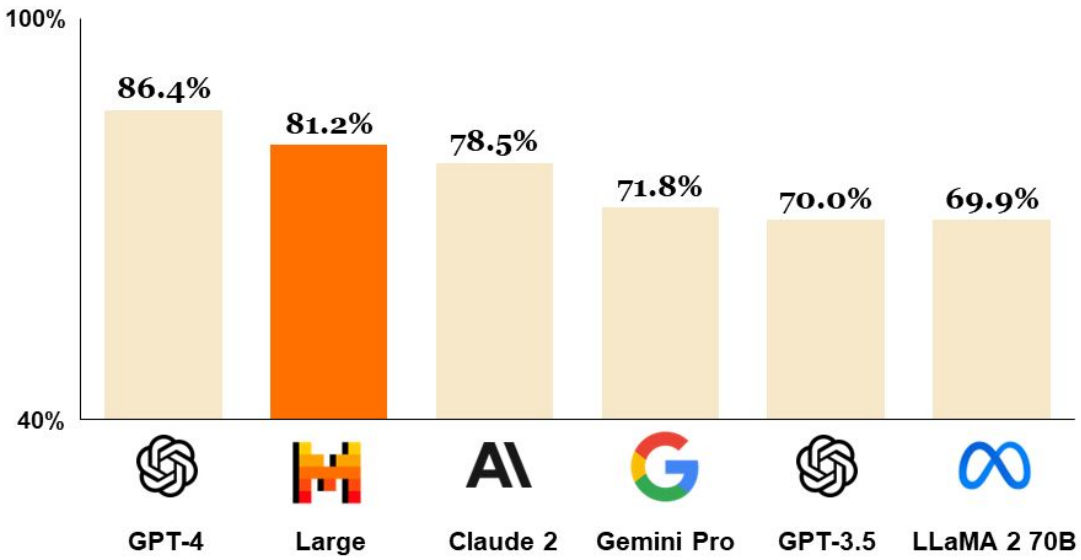
On-premise

Private cloud

Ease of deployment → **Increased control**

Mistral Large, our latest flagship model, is among the top tier models – and much more customisable

MMLU (Multi-Task Language Understanding)



MMLU (Massive Multitask Language Understanding) is a multi-task evaluation framework. It covers 57 subjects such as humanities, social sciences and history

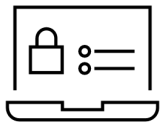
For your most sophisticated use cases

- Advanced reasoning capability
- Precise instruction following
- Language transformation

Advanced capabilities

- **Multi-lingual** by design - excellent in EN, FR, IT, DE, ES
- **Function calling** capabilities
- **RAG-enabled** - very performant on RAG use cases

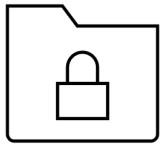
Our models are safe & trusted, putting controls in your hands



Privacy

Our APIs **do not track your inputs** or leverage outputs for training

Separate APIs in US and EU regions for stronger compliance



Security

Your **own security policies** on managed cloud platforms

Fully opaque screen on your proprietary assets when deploying on-premise (or VPC)

Bias-control

Our models demonstrate **top level performance** when it comes to gender, religion, politics and ethnicities biases

- *See benchmarks (e.g., BBQ or BOLD)*

Guardrails

Your **own guardrail policy:** define what is appropriate for your **specific use-case**

Our customers **make their own editorial choices** for optimal performance and adapted control


MISTRAL
AI_
Thank you!