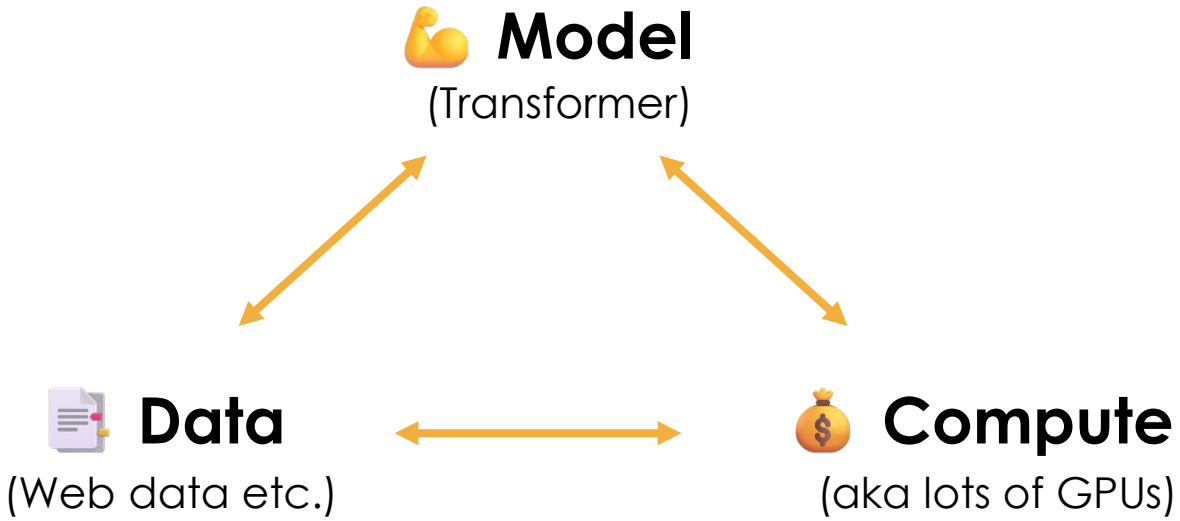
pen LLMs

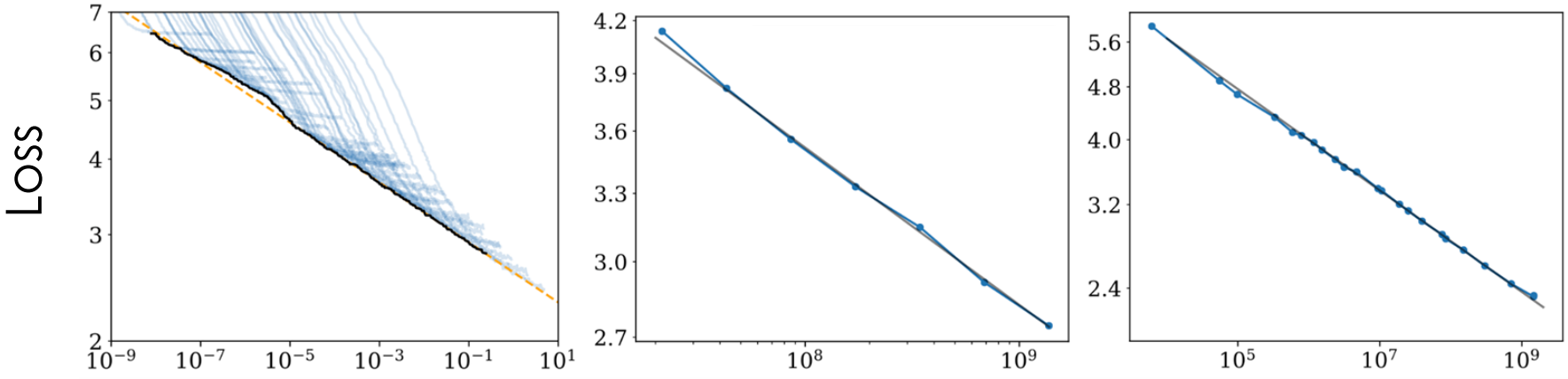
The comeback of open models and current trends in the LLM landscape

Leandro von Werra
Chief Loss Officer at Hugging Face

How did we get here?



Remarkable finding: **Scaling laws**



Compute Data Model size

https://arxiv.org/abs/2001.08361

Remarkable finding: **Scaling laws**

	Dataset (Billion Tokens)	Model size (Billion Parameter)	Compute:
GPT 1:	1-2	0.11	
			100x
GPT 2:	10-20	1.4	
			2000x
GPT 3:	300	175	
			300x
GPT 4:	10'000	1'800	

GPT-4 cost: ***~$100M Dollars***

Remarkable finding: **Chinchilla laws**

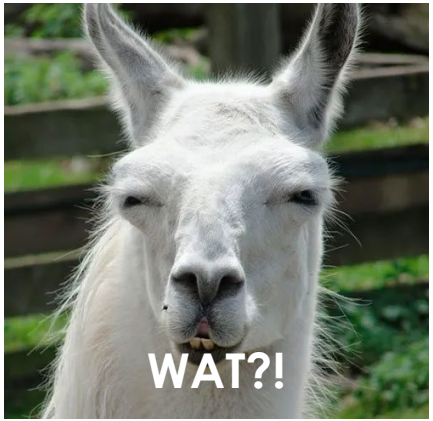
Parameters	FLOPs	Tokens
400 Million	1.92e+19	8.0 Billion
1 Billion	1.21e+20	20.2 Billion
10 Billion	1.23e+22	205.1 Billion
67 Billion	5.76e+23	1.5 Trillion
175 Billion	3.85e+24	3.7 Trillion
280 Billion	9.90e+24	5.9 Trillion

https://arxiv.org/abs/2203.15556

Remarkable finding: Chinchilla laws

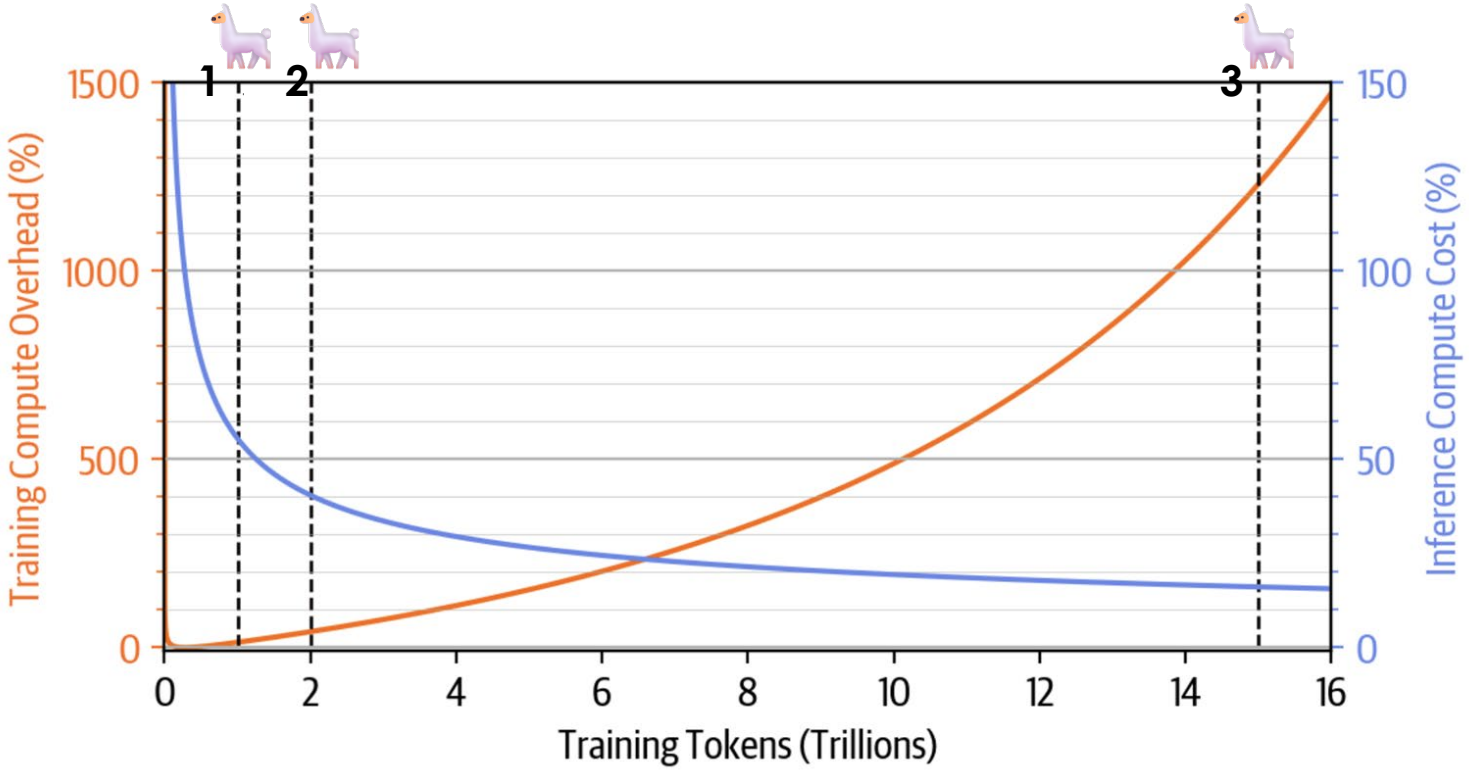
Parameters	FLOPs	Tokens
400 Million	1.92e+19	8.0 Billion
1 Billion	1.21e+20	20.2 Billion
10 Billion	1.23e+22	205.1 Billion
67 Billion	5.76e+23	1.5 Trillion
175 Billion	3.85e+24	3.7 Trillion
280 Billion	9.90e+24	5.9 Trillion

https://arxiv.org/abs/2203.15556



Llama-3 8B trained on 15T tokens

New trend: beyond Chinchilla optimal



Second spring for open models?

Data Center Dynamics

Meta to operate "600000 H100 GPU equivalents of compute" by year-end

Meta expects to field a fleet of 600,000 GPUs by the end of 2024. CEO Mark Zuckerberg told The Verge that the number includes some 340,000...

18 Jan 2024

***GPT-4 used 25'000 A100s for 3-4 months**

Synced

DeepMind's Gemma: Advancing AI Safety and Performance with Open Models

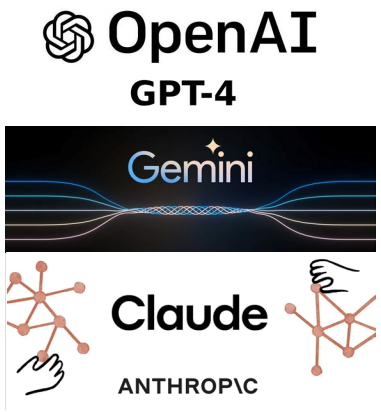
Large Language Models (LLMs) have proven their mettle across a spectrum of real-world applications, ranging from language modeling to visual...

Gemma

Mistral AI, a Paris-based OpenAI rival, closed its $415 million funding round

Romain Dillet @romaindillet / 12:47 PM GMT+1 • December 11, 2023 Comment

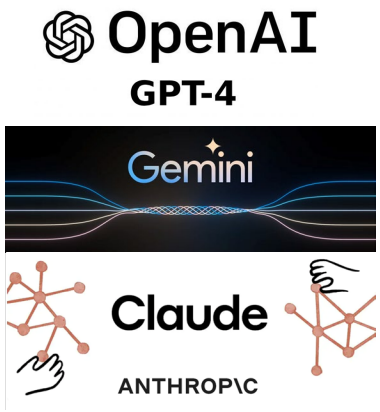
closed model APIs **open** model weights **fully open** model



model weights not available

- can't run the model locally
- no access to model's internals
- limits fine-tuning abilities

closed model APIs



model weights not available

- can't run the model locally
- no access to model's internals
- limits fine-tuning abilities

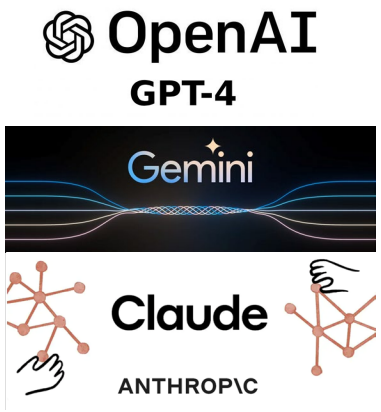
open model weights

no access to training data or code

- who's data is in the dataset?
- can't remove data on request
- can't inspect data for biases
- benchmark contamination
- limits scientific reproducibility

fully open model

closed model APIs



model weights not available

- can't run the model locally
- no access to model's internals
- limits fine-tuning abilities

open model weights

no access to training data or code

- who's data is in the dataset?
- can't remove data on request
- can't inspect data for biases
- benchmark contamination
- limits scientific reproducibility

fully open model

full access to model/code/data

- competitive edge
- liability issues
- maintenance

Open compute in Europe

LUMI (Finland): 11’912 GPUs (AMD MI250x)

JUWELS (Germany): 3’774 GPUs (NVIDIA A100)

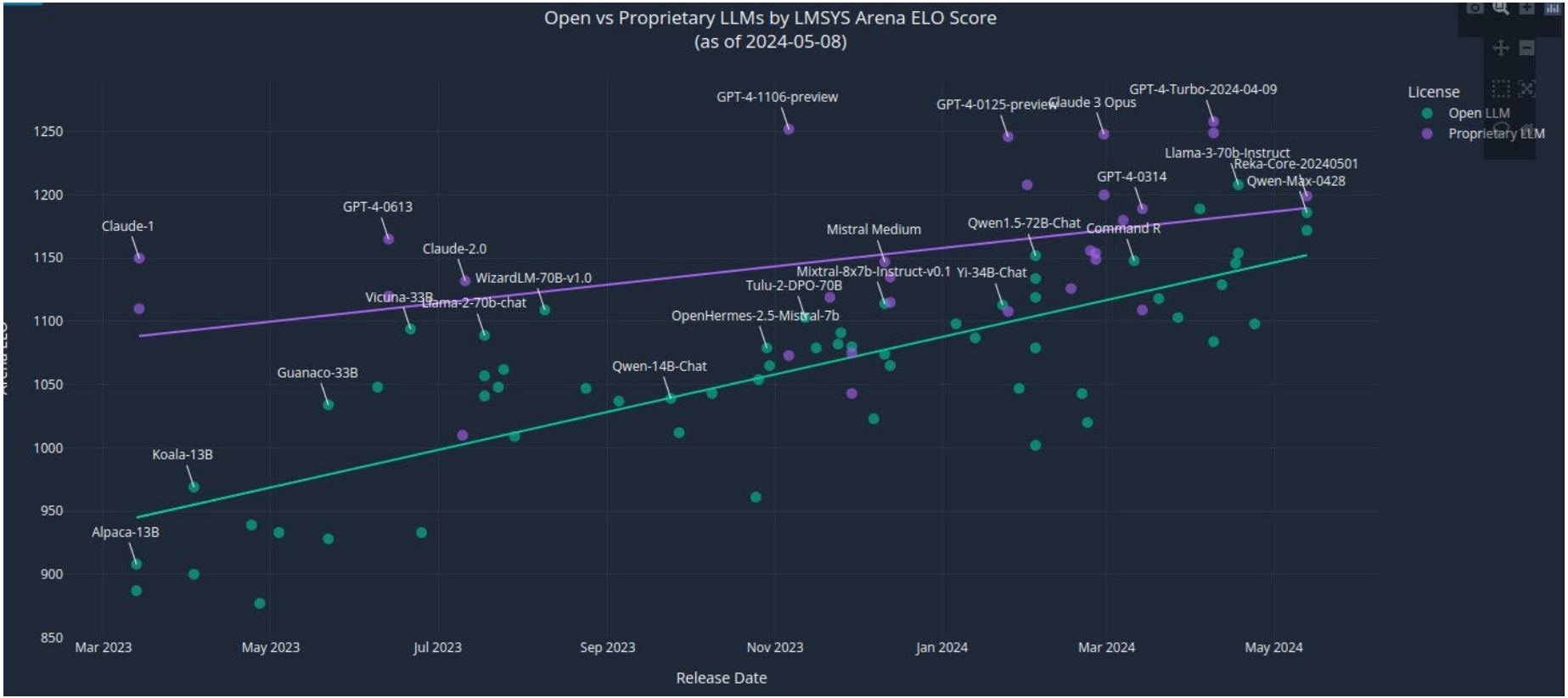
coming soon

JUPITER (Germany): 24’000 GPUs (NVIDIA GH200)

Leonardo (Italy): 13’824 GPUs (NVIDIA A100)

Alps (Switzerland): 10’000 GPUs (NVIDIA GH200)

Second spring for open models?

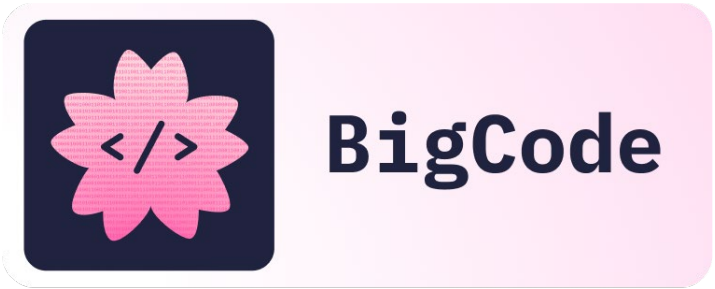


Closed vs **open** access models

“Software is eating the world.”

-Marc Andreessen, 2011


BigScience

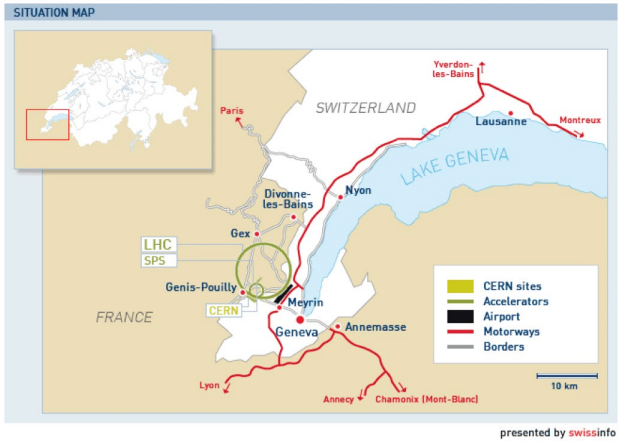

BigCode

BigScience

Inspiration

Large Hadron Collider

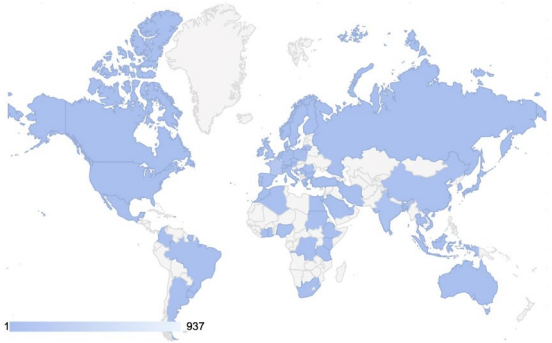
- involved **10.000 researchers**
- from **100 countries**
- discovery of **59 hadrons**
- more than **2.800 papers** (😱)



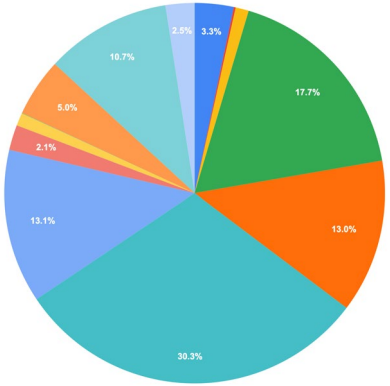
In many scientific fields such world-scale research collaborations create **research tools** which are essential for the research community: LHC, ITER, ISS, etc

Time for similar **large, diverse, open research collaboration** in AI?

BigScience *Project*



1000+ Collaborators
from 67 countries



Roots: Dataset with
300B tokens comprising
59 languages

Bloom (175B) was
trained on Jean Zay with
400 GPUs for 3 Months

BigCode

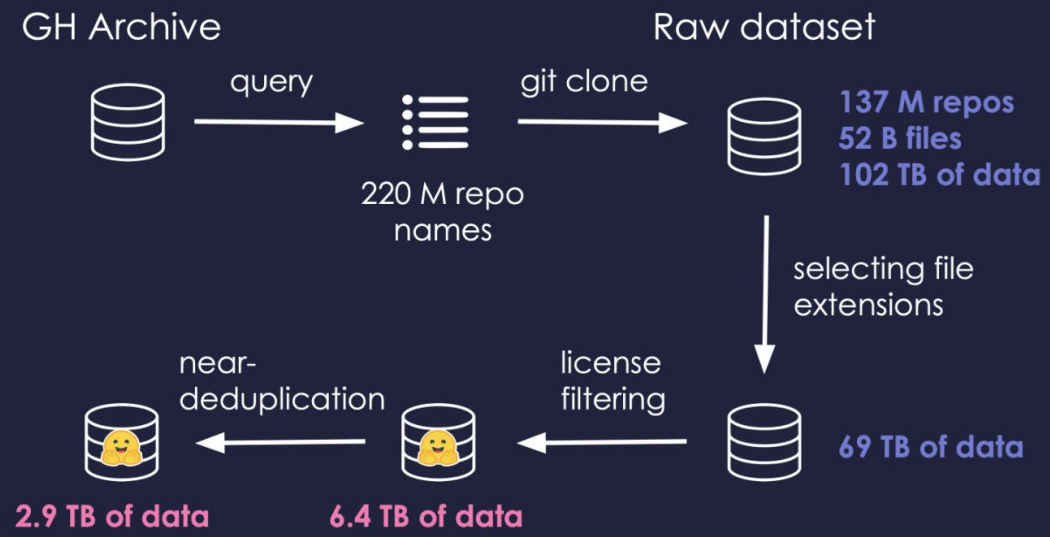
Building LLMs for code in a collaborative way:

- Full data transparency
- Open source processing and training code
- Model weights released with commercial friendly license

1100+ researchers, engineers, lawyers, and policy makers

Hugging Face servicenow

The Stack: data collection



Find the filtered and deduplicated datasets at: www.hf.co/bigcode

BigCode

Models

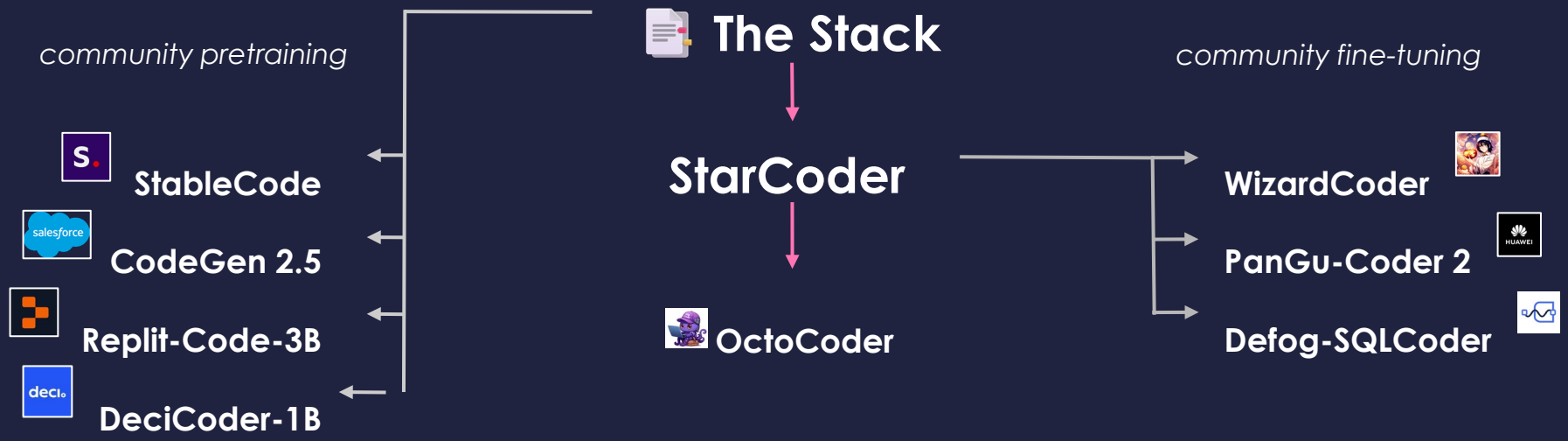
SantaCoder
Dec 2022

StarCoder
May 2023

StarCoder2
Feb 2024

	SantaCoder	StarCoder	StarCoder2
Cost:	$8 k	$480 k	$2.1 M
Languages:	3	~80	~600
Tokens:	0.24 T	1 T	4 T
Model size:	1.1 B	15 B	15 B

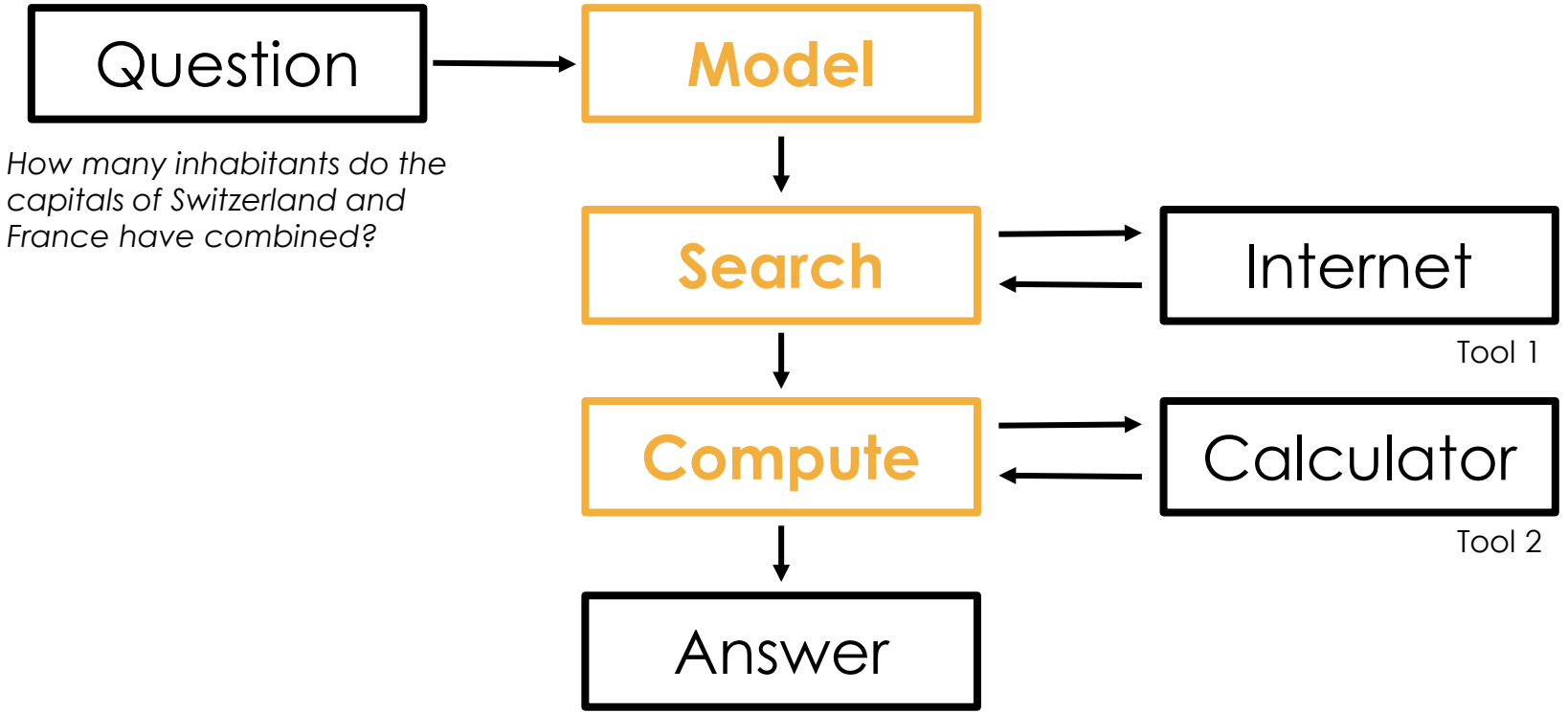
BigCode Ecosystem



Trend: **Agents & Tools**

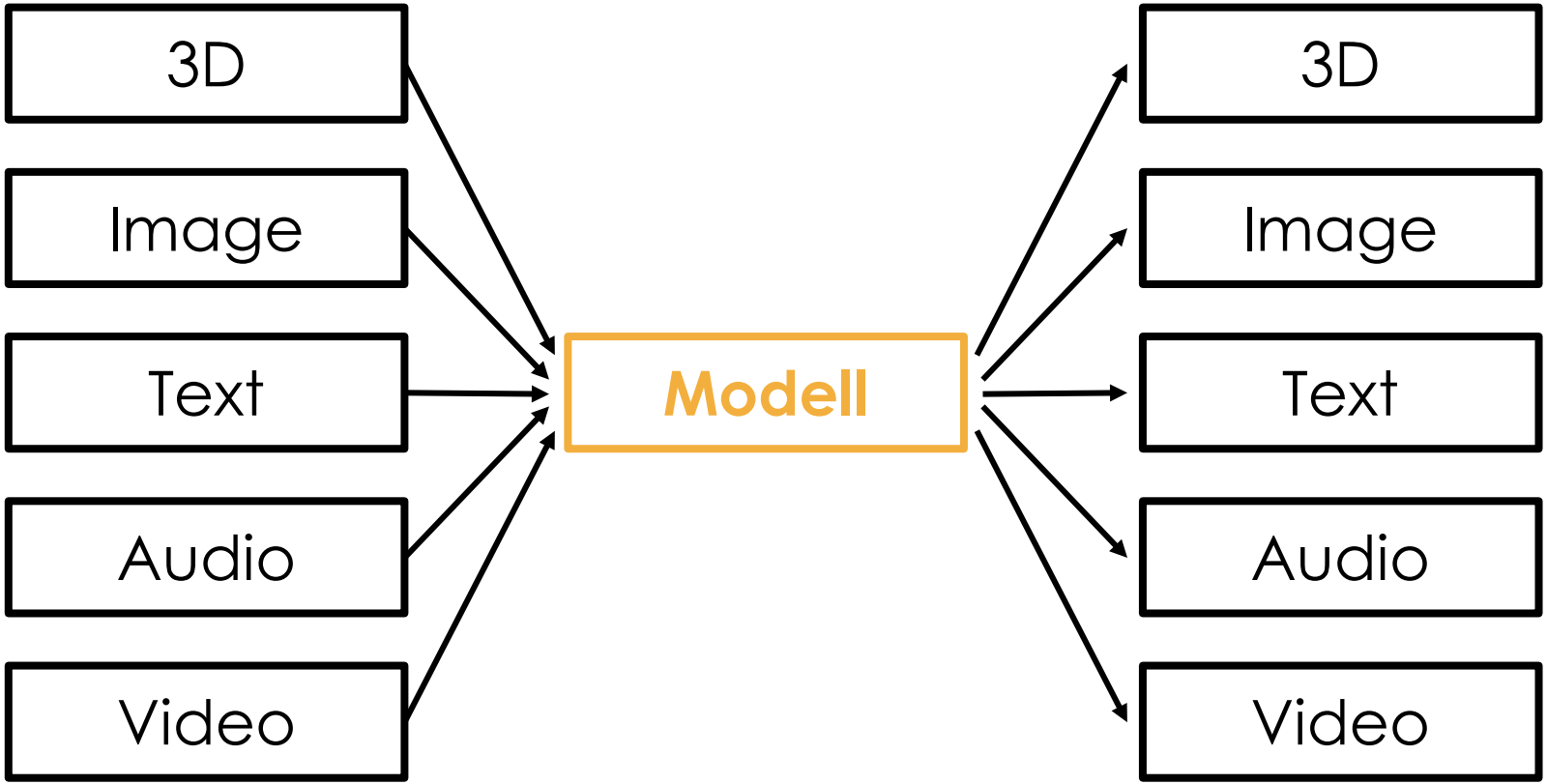
How many inhabitants do the capitals of Switzerland and France have combined?

Trend: Agents & Tools



Trend: **Multimodality**

Trend: **Multimodality**



Questions?

GitHub/HF Hub/X: **lvwerra**