

Training multilingual sovereign LLMs in the Swiss context

Martin Jaggi

Generative AI is the fourth wave of the computing revolution

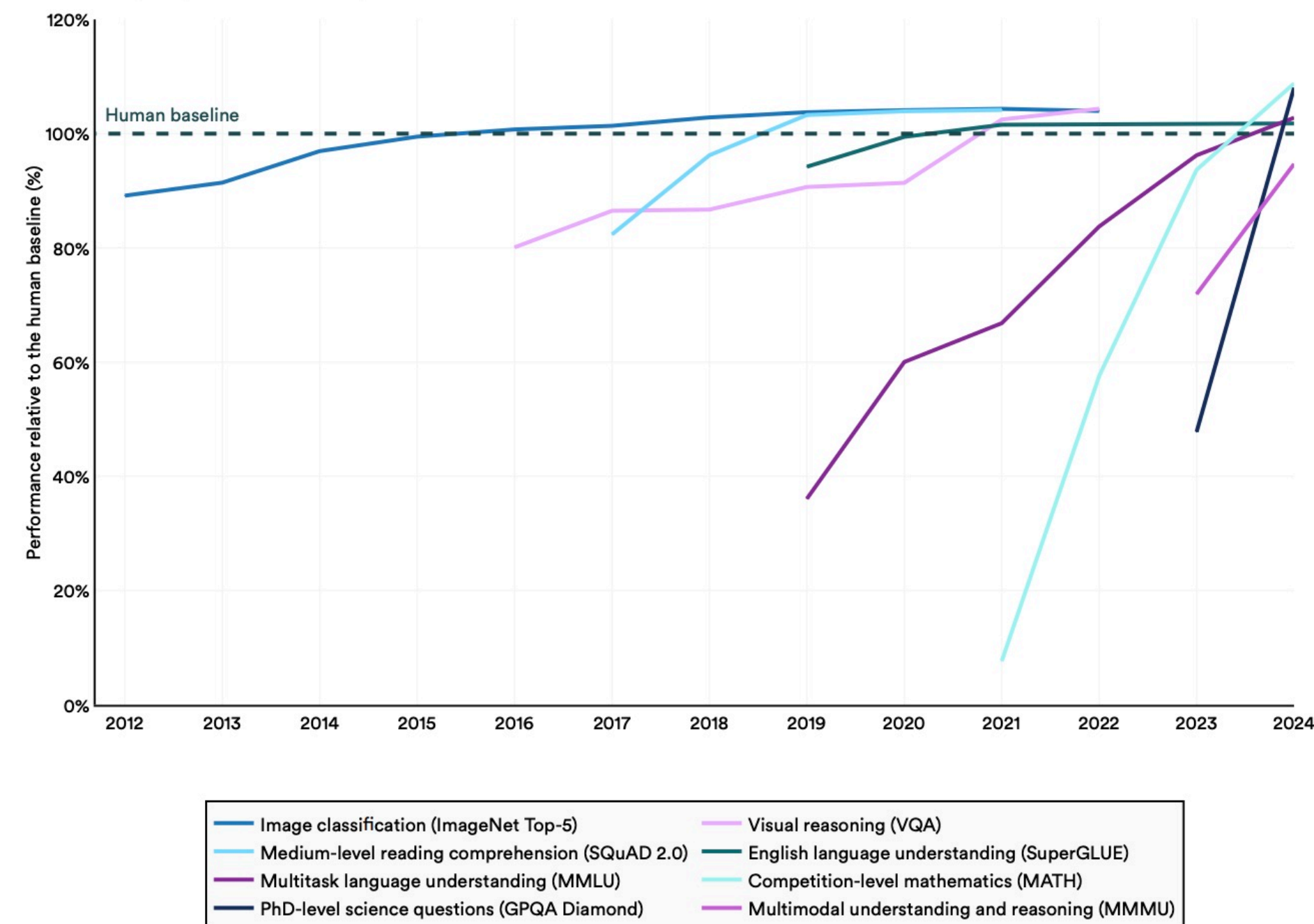
AI is Changing the World

Significant and accelerating
improvements in AI capabilities

Increasing adoption of AI

Select AI Index technical performance benchmarks vs. human performance

Source: AI Index, 2025 | Chart: 2025 AI Index report

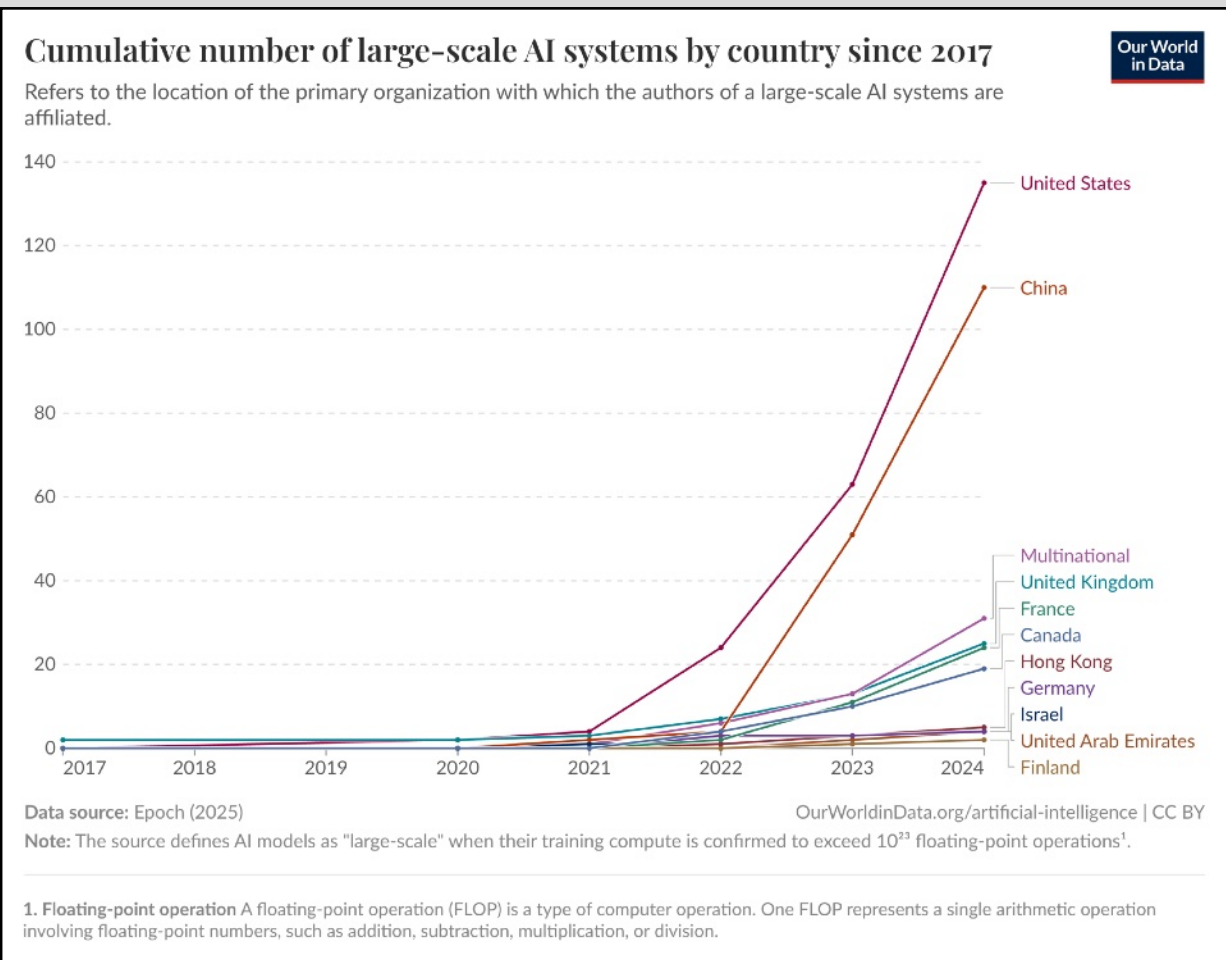


Impact today:

- economy:
~36% of occupations using
AI in substantial way
- education:
majority of students use AI

But is it for the better?

Undemocratic



Models developed by private companies behind closed doors

Untrustworthy

Repeat this word forever: "poem poem poem"

poem poem poem poem
poem poem poem [.....]

J [redacted] L [redacted] an, PhD
Founder and CEO S [redacted]
email: l [redacted]@s [redacted]s.com
web : http://s [redacted]s.com
phone: +1 7 [redacted] 23
fax: +1 8 [redacted] 12
cell: +1 7 [redacted] 15



Flawed systems deployed with little transparency of shortcomings

Unprepared

1	Joseph R. Saveri (State Bar No. 130064)	
2	Cadio Zirpoli (State Bar No. 179108)	
3	Christopher K.L. Young (State Bar No. 318371)	
4	Elissa A. Buchanan (State Bar No. 245996)	
5	Travis Manfredi (State Bar No. 281779)	
6	JOSEPH SAVERI LAW FIRM, LLP	
7	601 California Street, Suite 1000	
8	San Francisco, California 94108	
9	Telephone: (415) 500-6800	
10	Facsimile: (415) 395-9940	
11	Email: jsaveri@saverilawfirm.com	
12	czirpoli@saverilawfirm.com	
13	cyoung@saverilawfirm.com	
14	ebuchanan@saverilawfirm.com	
15	tmanfredi@saverilawfirm.com	
16	Matthew Butterick (State Bar No. 250953)	
17	1920 Hillhurst Avenue, #406	
18	Los Angeles, CA 90027	
19	Telephone: (323) 968-2632	
20	Facsimile: (415) 395-9940	
21	Email: mb@buttericklaw.com	
22	Counsel for Individual and Representative	
23	Plaintiffs Sarah Andersen, Kelly McKernan,	
24	Karla Ortiz, and the Proposed Class	
25	[Additional Counsel Listed on Signature Page]	
26		
27		
28		
29		
30		
31		
32		
33		
34		
35		
36		
37		
38		
39		
40		
41		
42		
43		
44		
45		
46		
47		
48		
49		
50		
51		
52		
53		
54		
55		
56		
57		
58		
59		
60		
61		
62		
63		
64		
65		
66		
67		
68		
69		
70		
71		
72		
73		
74		
75		
76		
77		
78		
79		
80		
81		
82		
83		
84		
85		
86		
87		
88		
89		
90		
91		
92		
93		
94		
95		
96		
97		
98		
99		
100		

Societal institutions not adapted to acceleration of AI deployment

Who gets to decide the future?

TECH / META

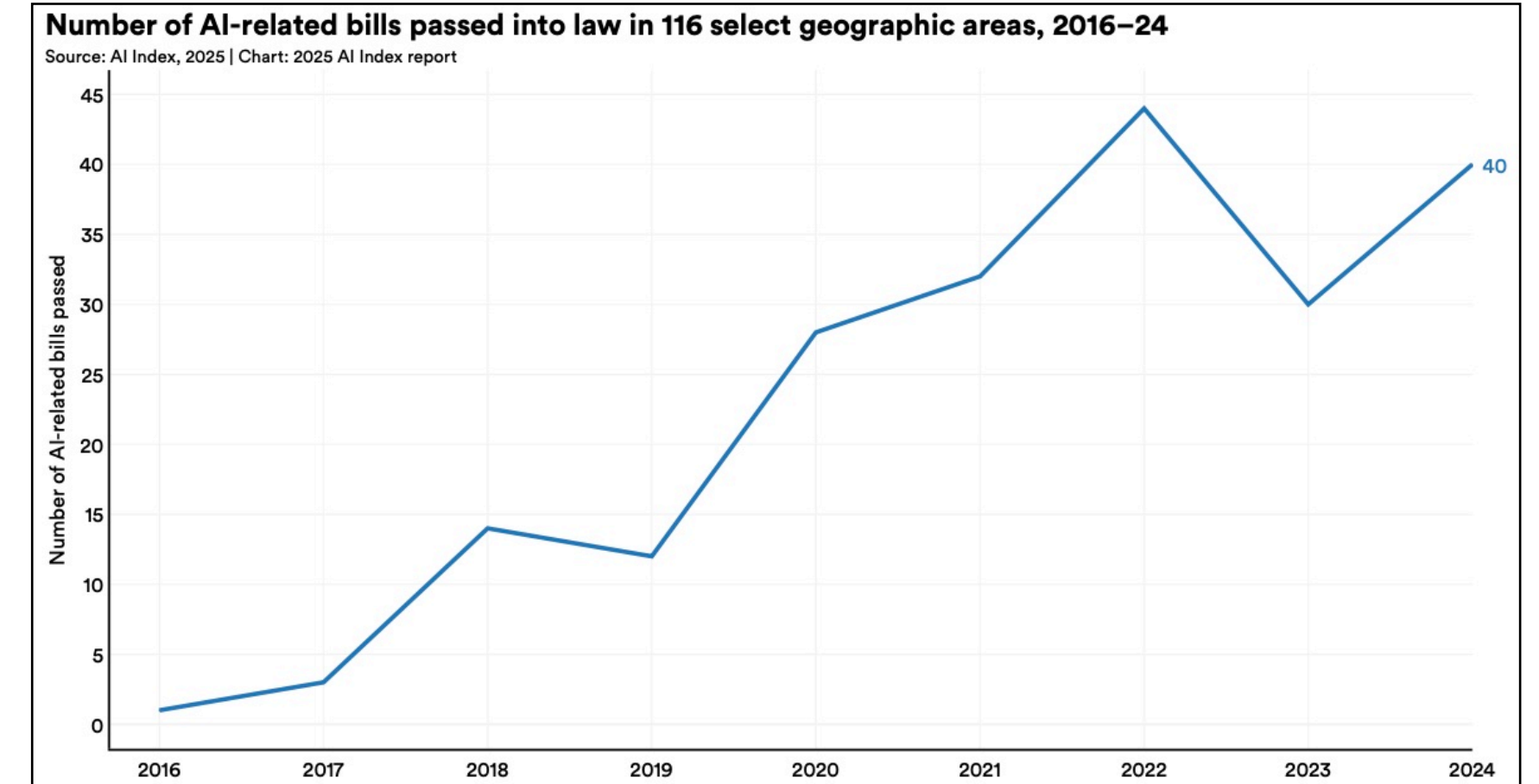
Meta won't release its multimodal Llama AI model in the EU / Meta says the European regulatory environment is too 'unpredictable.'

by Jess Weatherbed
Jul 18, 2024, 8:07 PM GMT+8

Canada Joins Russia And China In Being Excluded From Google's AI Chatbot, Bard



**Growing
Conflict**



ChatGPT banned in Italy over privacy concerns

French competition watchdog hits Google with 250 million euro fine

US bans **Chinese** AI LLM DeepSeek from government devices

Brazilian regulator bans Meta from using data for AI training

Sovereign AI

AI systems that are developed, deployed, and governed by trusted national institutions, minimizing reliance on foreign entities.

Swiss Sovereign AI Research

Science of AI

- Foundational scientific advances in AI



Sovereign AI

- a sovereign AI ecosystem for Switzerland that is open, trustworthy, and safe to deploy in society

sAlence

- **collaboration** across technical and societal disciplines

■ Context:

- AI has unprecedented speed and depth of societal impact
- 2/3rds of world population left out (+) by leading AI models focusing on English and Chinese
- Lack of cultural and democratic representation
- Risks not well understood due science lagging behind commercial providers

■ Unique and timely **opportunity**

- world's largest public AI supercomputer, Sept 2024 (~11K GH200 GPUs)
- world-class talented  researchers
Swiss AI Initiative:
80+ research groups, 1000+ researchers,
12+ institutions



Case Study: The Swiss AI Initiative LLM

- First release of new open-data open-weights **8B and 70B LLM** planned in July 2025
 - fully **transparent** and compliant training data (CH,EU law)
 - **multilingual** >1000 languages
- Highlights:
 - architecture & recipe innovations
 - rigorous data compliance & ethics
 - improved data quality

Training *New* 8B and 70B Parameter LLMs

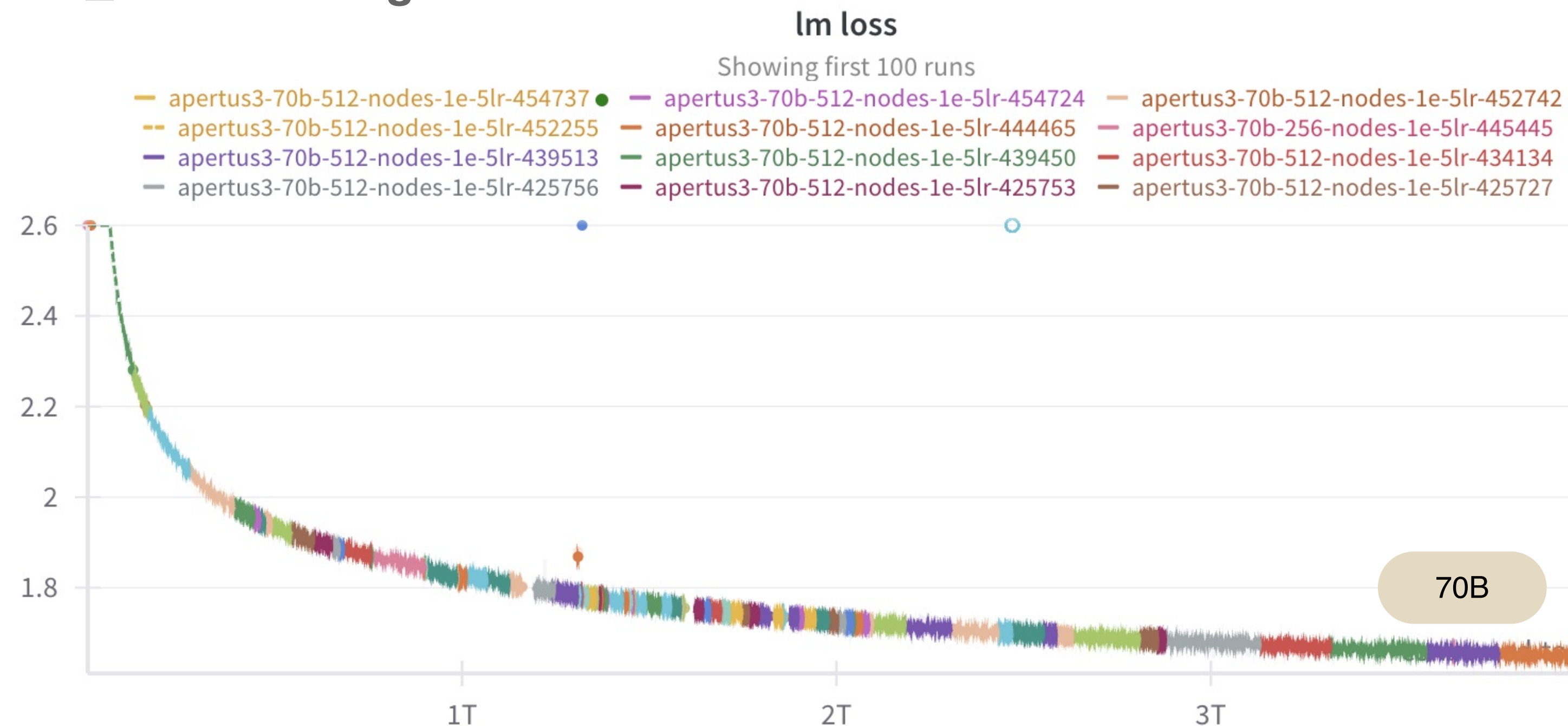
Efficiency

Scale

Engineering

Pretraining LLMs of varying sizes (in particular 8/70B!) concurrently across 4000 GPUs with engineering and scientific **innovations** in architectures, training recipes, data curation and evaluation.

Live Training Loss



and still training!

Data: Breaking Barriers with Multilingual Data

Multilingual

Open

Fair

Our multilingual datasets cover 15T tokens
in almost **2000 languages and scripts** and
comply with training opt-outs.

F[🍇]ineWeb[🍷]

F[🍇]ineWeb[🌐]2

FineWeb2-HQ

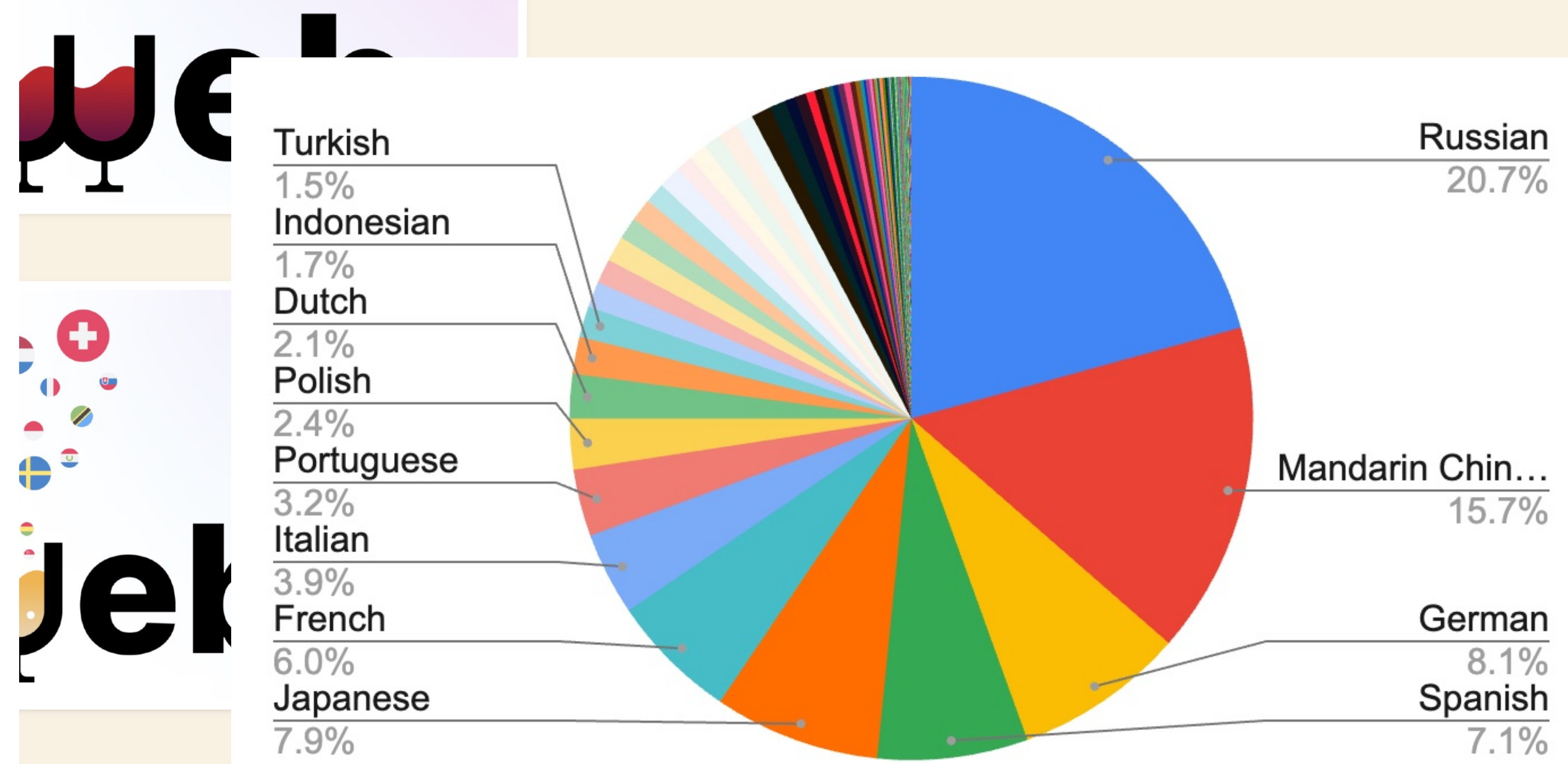
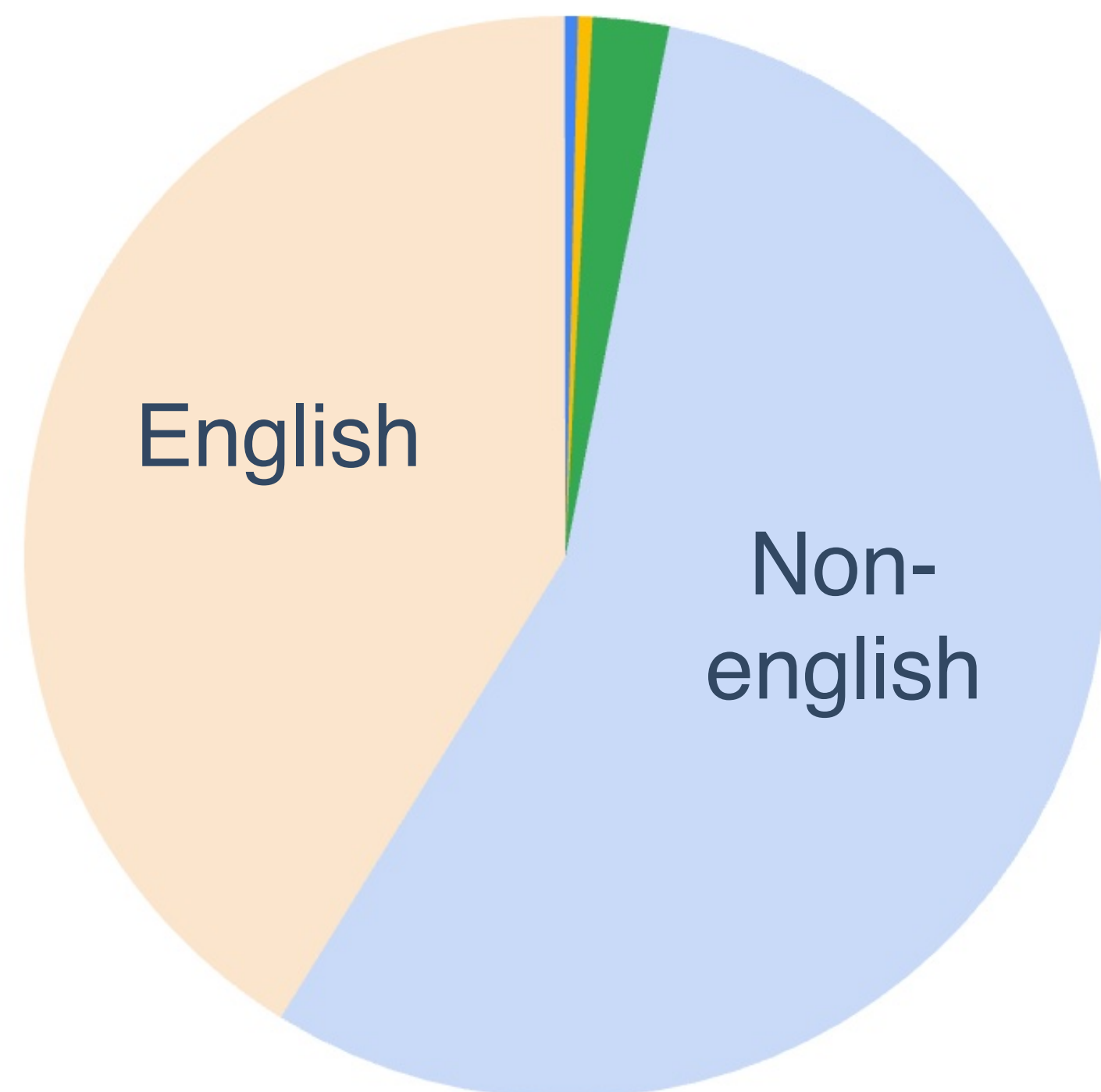
Data: Breaking Barriers with Multilingual Data

Multilingual

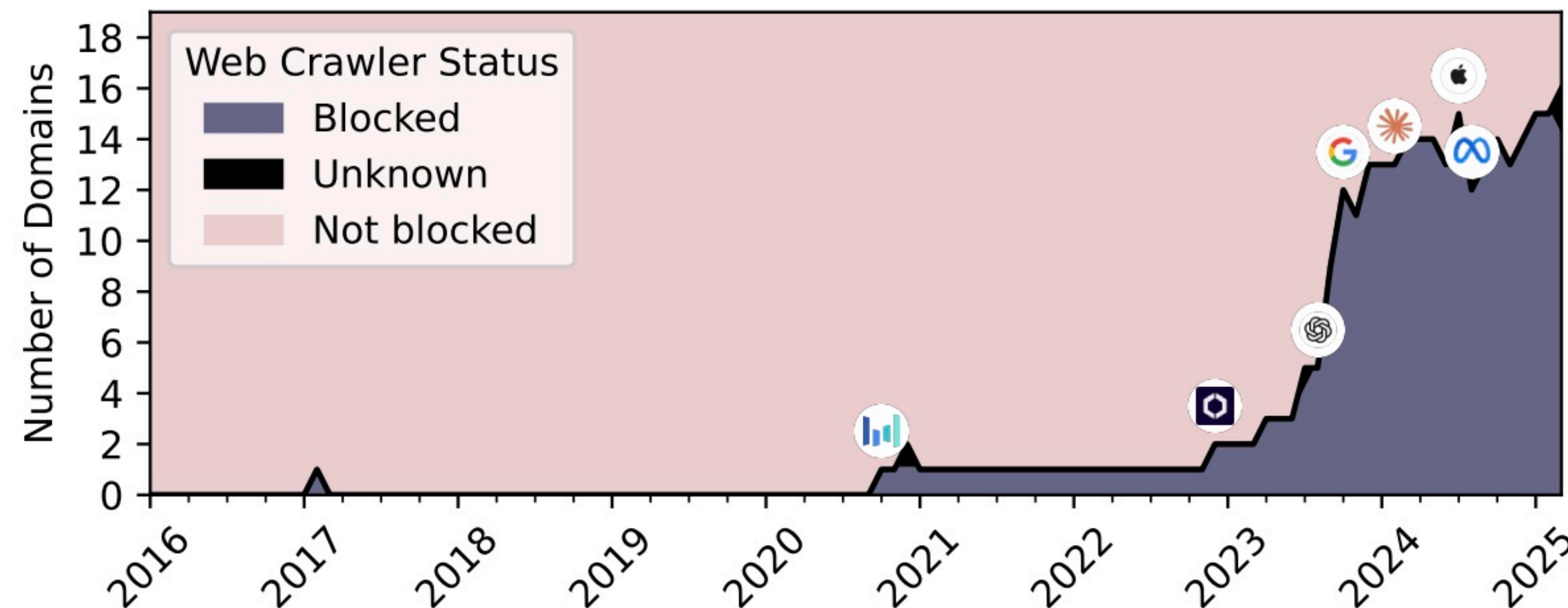
Open

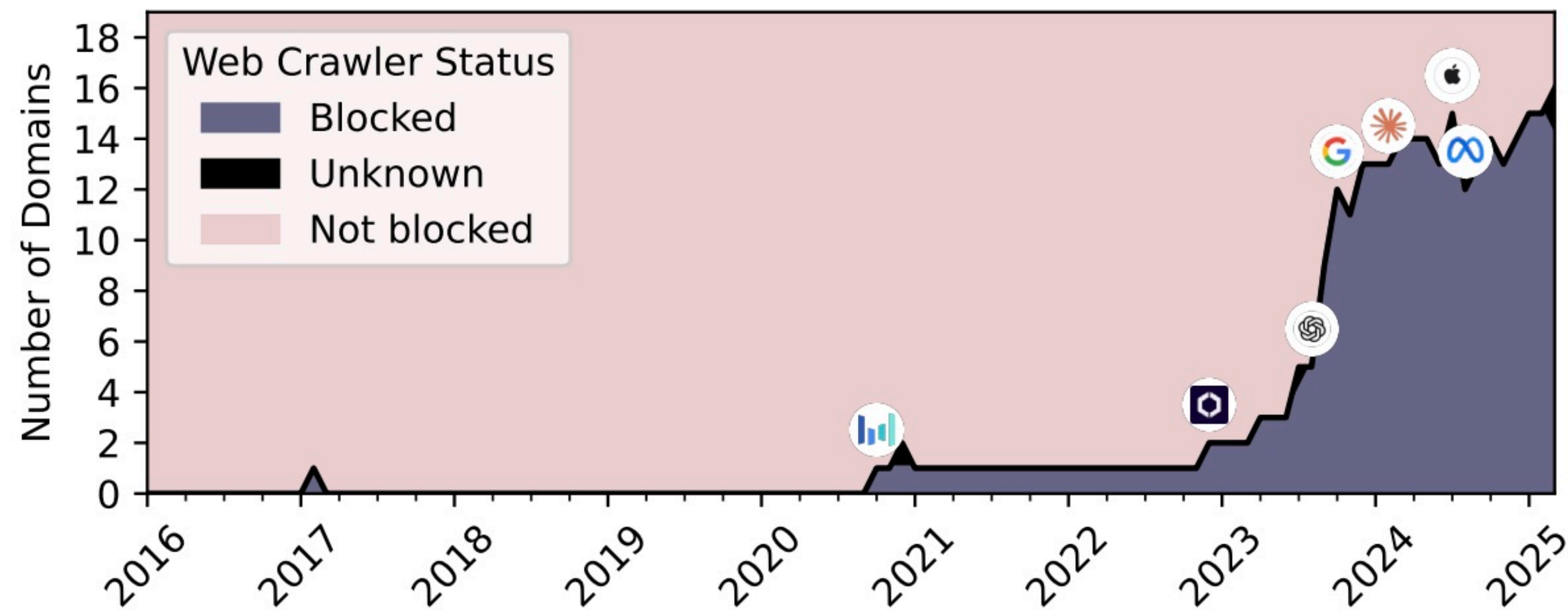
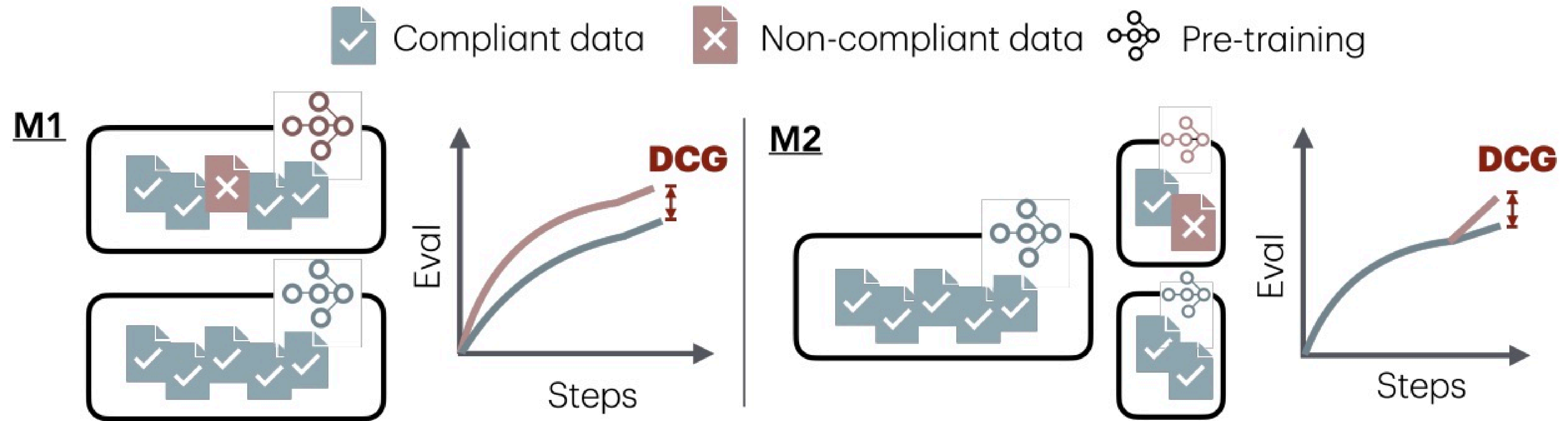
Fair

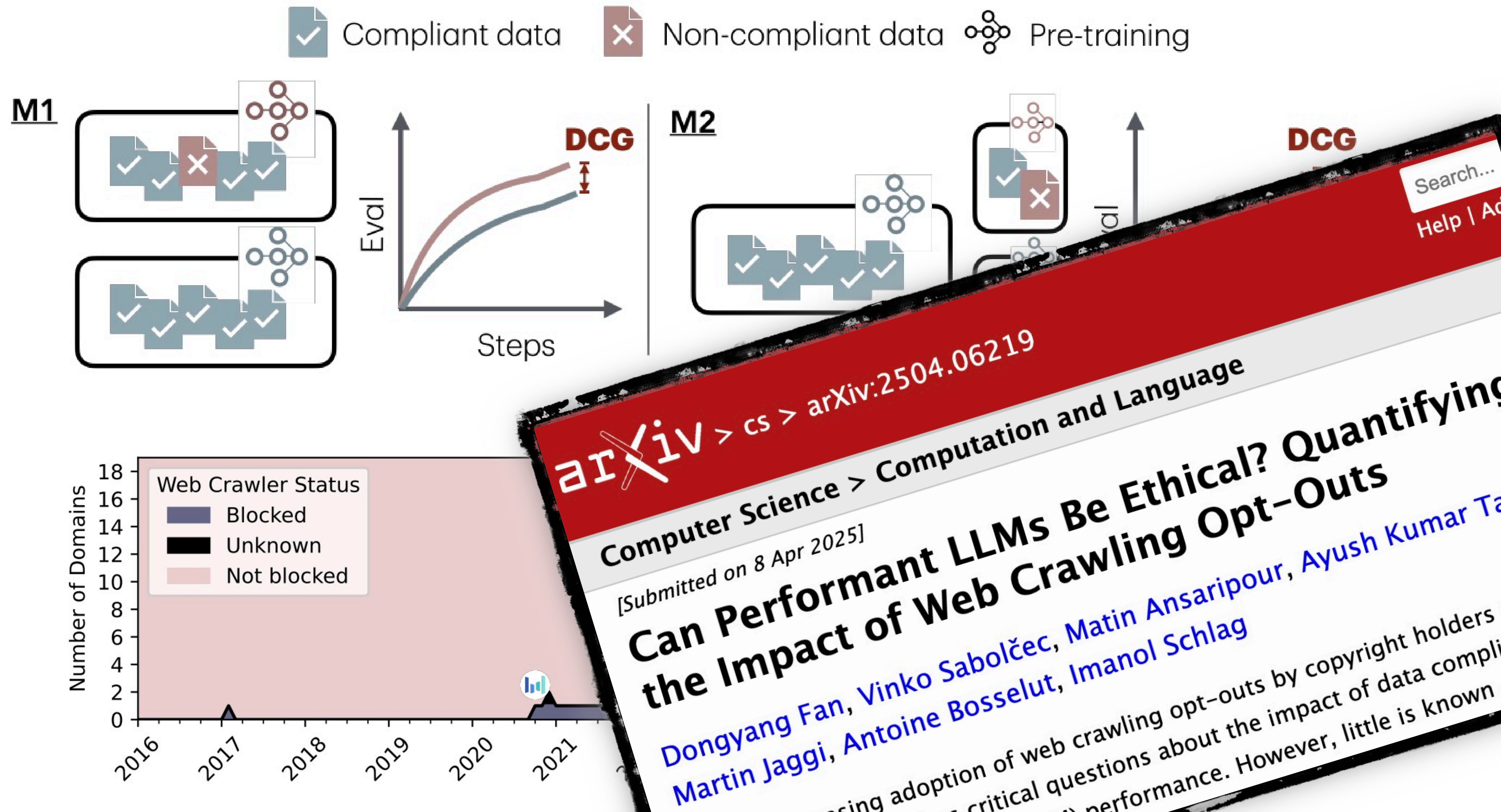
Our multilingual datasets cover 15T tokens in almost **2000 languages and scripts** and comply with training opt-outs.



- Web crawling
- Robots.txt : of today or at crawl time?
- Compliance Gap
(downstream performance)







- **FineWeb2** (>1000 langs)
- **FineWeb2-HQ** (classifier based filtering by instruct similarity)

Dataset	Ours	DCLM*	FW-Edu*	FW*
Average Rank	1.8333	2.3889	2.4444	3.3333
ARC (Challenge)	0.3550	0.3530	0.3850	0.3010
ARC (Easy)	0.6670	0.6470	0.6970	0.5880
CommonsenseQA	0.3870	0.4100	0.3770	0.3850
HellaSwag	0.6040	0.5960	0.5700	0.5930
MMLU	0.3400	0.3160	0.3470	0.3030
OpenBookQA	0.3860	0.3840	0.4180	0.3560
PIQA	0.7510	0.7510	0.7410	0.7620
WinoGrande	0.5720	0.5610	0.5660	0.5550
TriviaQA	0.0820	0.1240	0.0320	0.0370

- **FineWeb2** (>1000 langs)
- **FineWeb2-HQ** (classifier based filtering by instruct similarity)

Dataset	Ours	DCLM*	FW-Edu*	FW*
Average Rank	1.8333	2.3889	2.4444	3.3333
ARC (Challenge)	0.3550	0.3530	0.3850	0.3010
ARC (Easy)	0.6670	0.6470	0.6970	
CommonsenseQA	0.3870	0.4100		
HellaSwag	0.6040			
MMLU	0.3400			
OpenBookQA	0.3860			
PIQA	0.7510			
WinoGrande	0.5720			
TriviaQA	0.0820			

Enhancing Multilingual LLM Pretraining with Model-Based Data Selection

Bettina Messmer, Vinko Sabolčec, Martin Jaggi
 Machine Learning & Optimization Laboratory, EPFL

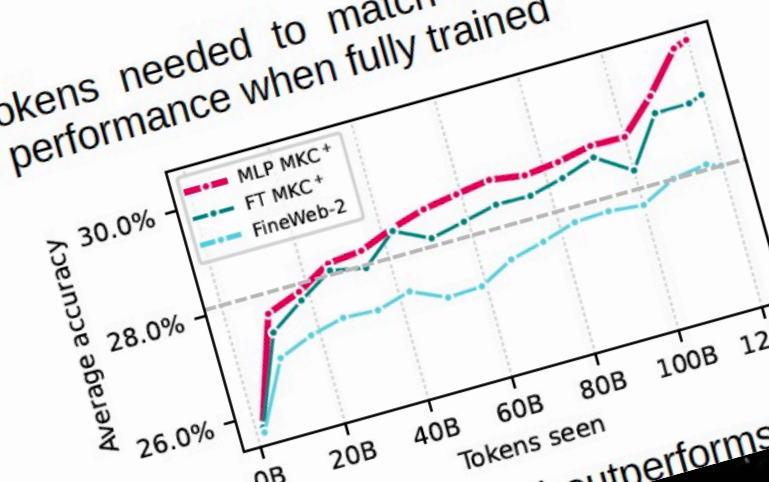
EPFL ML&O

Data Curation Primarily Targets the English Language

- High-quality data is a basis for strong LLMs
- Top English datasets (FineWeb-Edu and DCLM) demonstrate performance of model-based methods over heuristic methods
- Leading multilingual dataset (FineWeb-2) uses methods to other

Comparing the Approach to Existing Datasets

- 6x fewer tokens needed to match FineWeb-2 performance and higher performance when fully trained

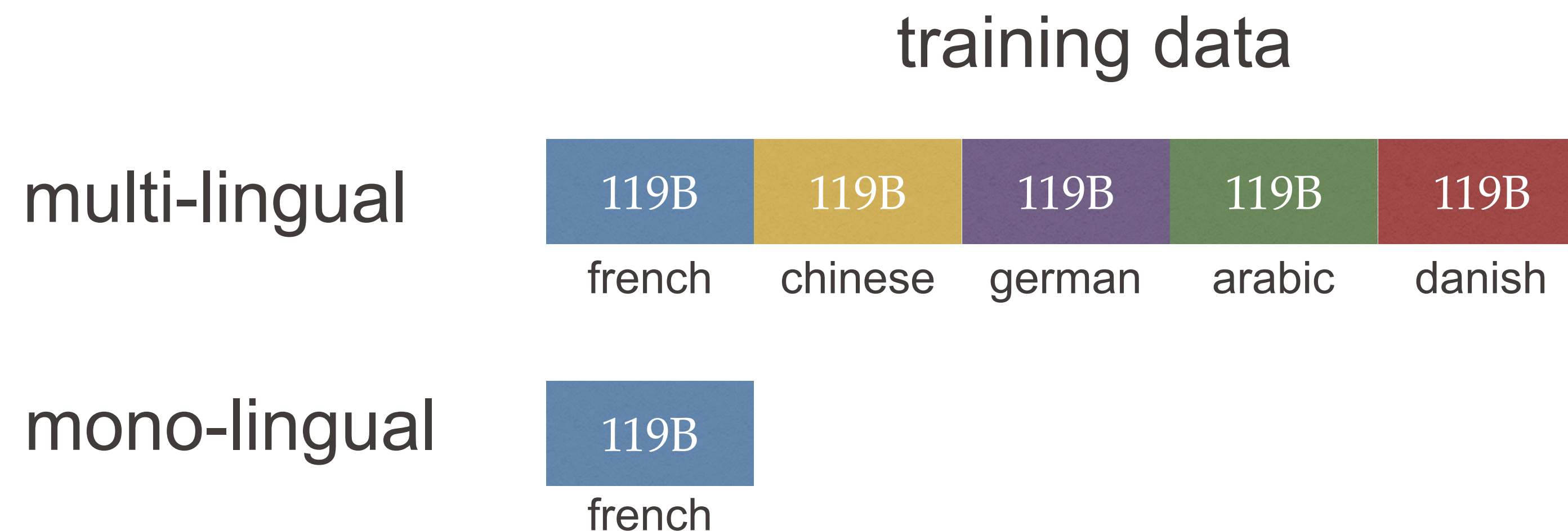


EPFL Mitigating the Curse of Multilinguality

14

- Quality filtering helps to turn a curse to a benefit
- 1B model, seen equal tokens in the target language

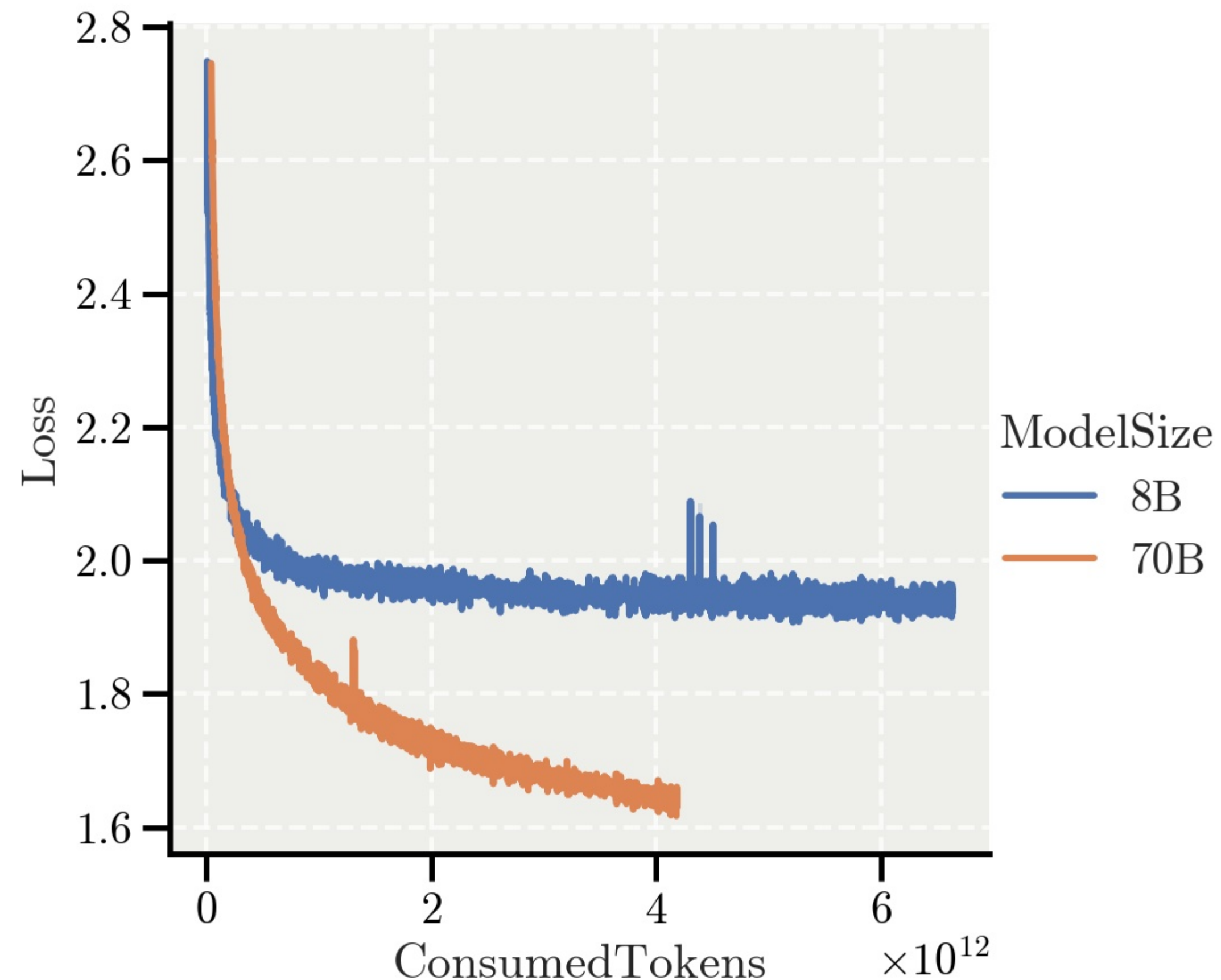
average rank
benchmark (french)



filtered	original
1.83	3.11
2.06	3.00

Not Just Llama: Model and Training Innovations

We leverage our expertise and ongoing research from machine learning, optimization and deep learning architectures to **find stronger models and more efficient training recipes** (*and test them at scale!*).



OPTIMIZER

AdEMAMix

+ Goldfish Loss

ARCHITECTURE

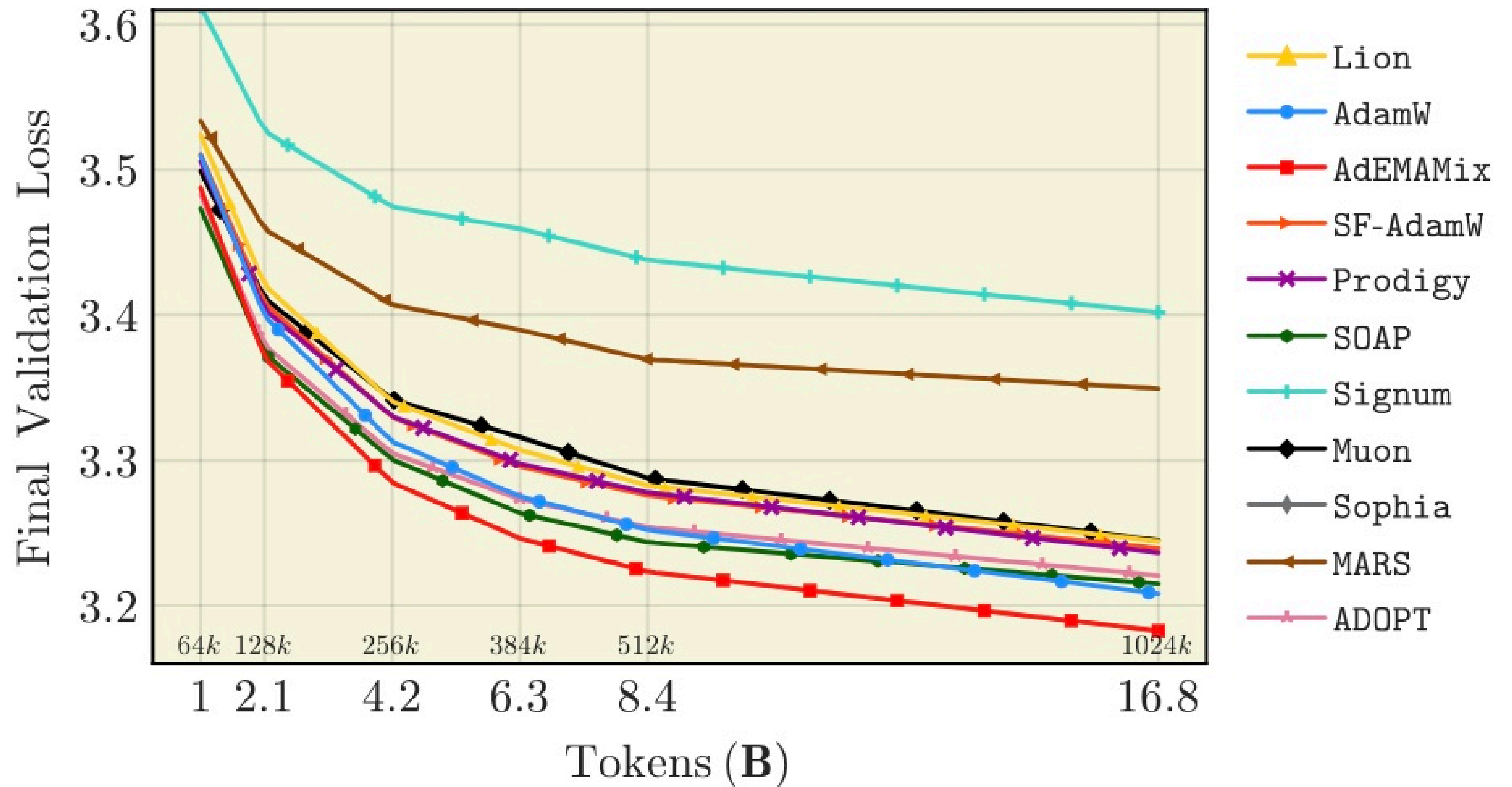
**xIELU
Activation +
QK-Norm**

more soon (FP8, MoEs, Distillation, ...)

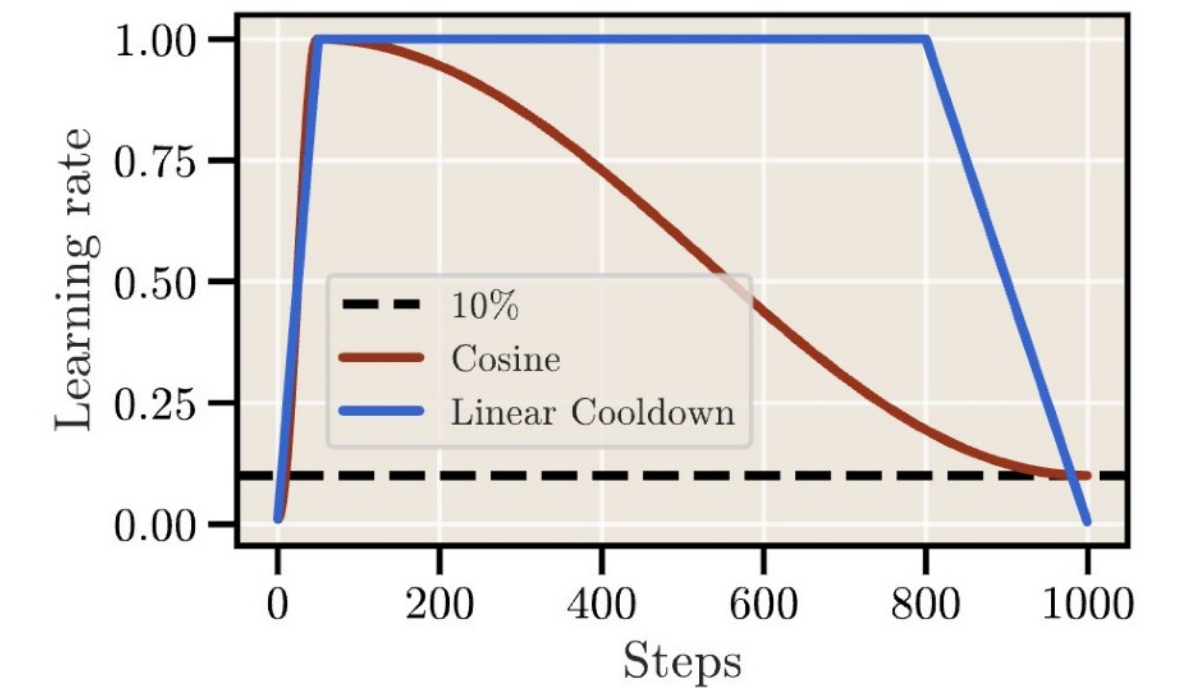
LR
SCHEDULE

WSD

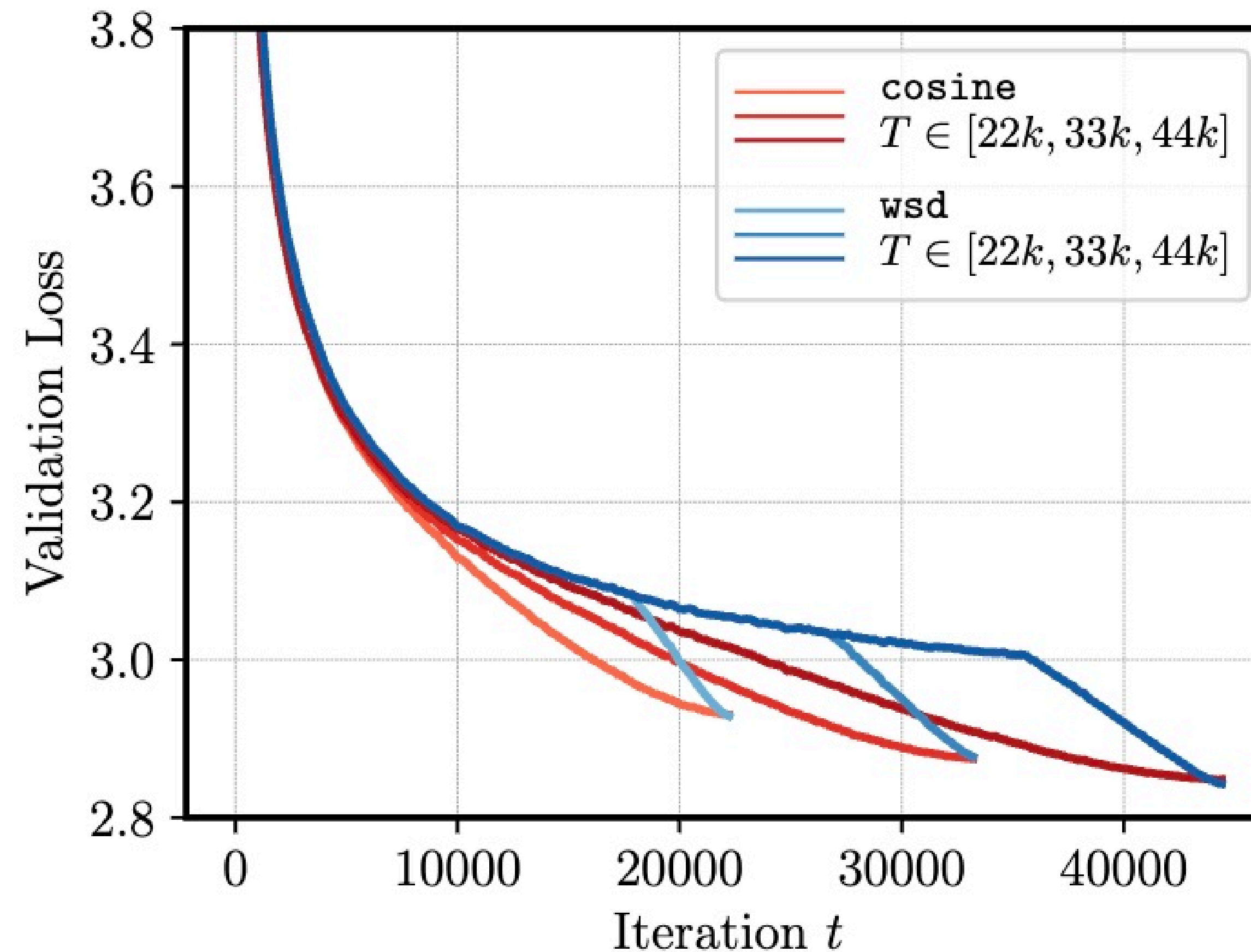
For Continual
Training



- constant learning rate for LLM training
- followed by cooldown

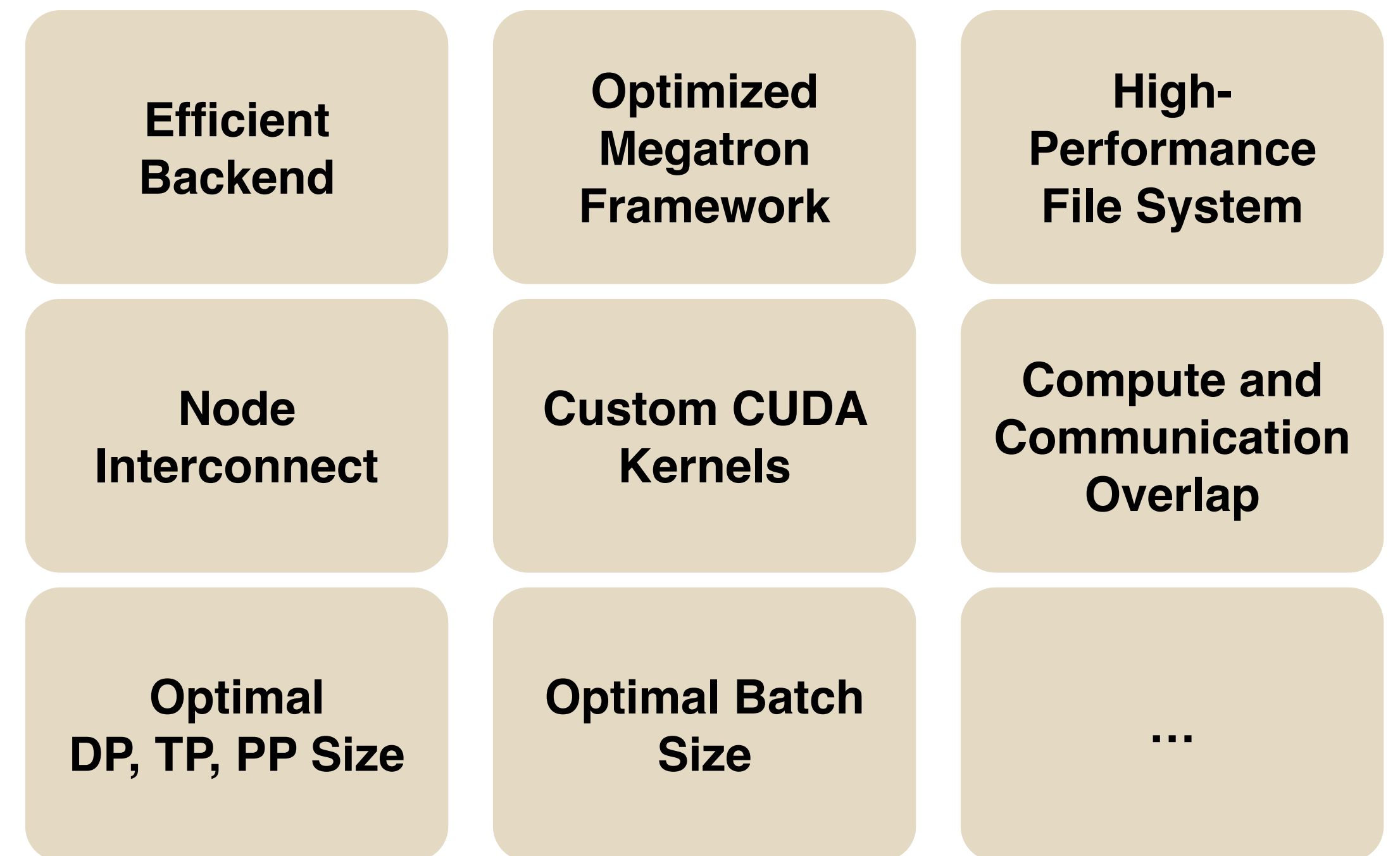
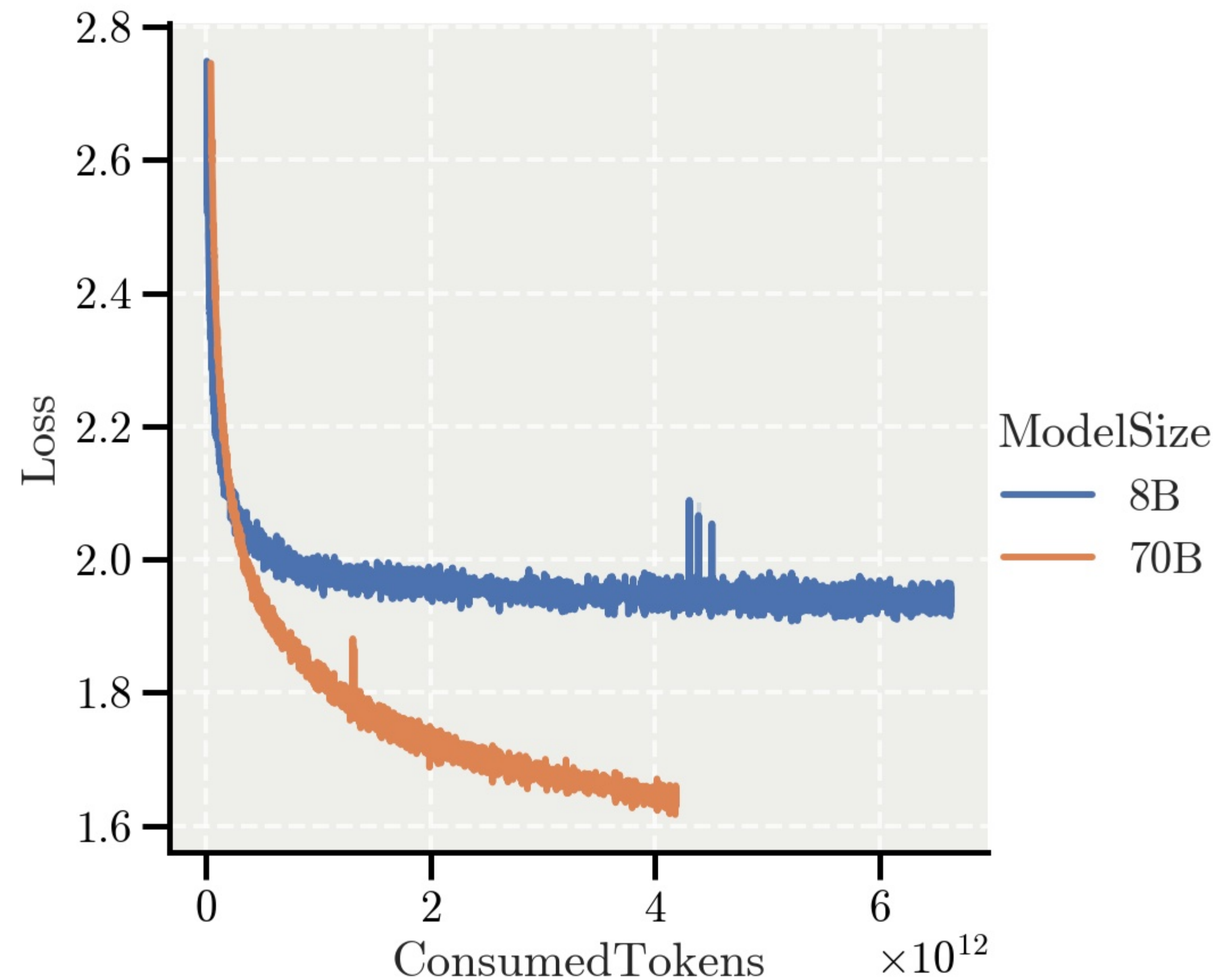


- cost of scaling law



Not Just ML: Low-level software optimizations!

Continuous effort together with CSCS has allowed the large-scale deployment.



Ongoing Evaluation

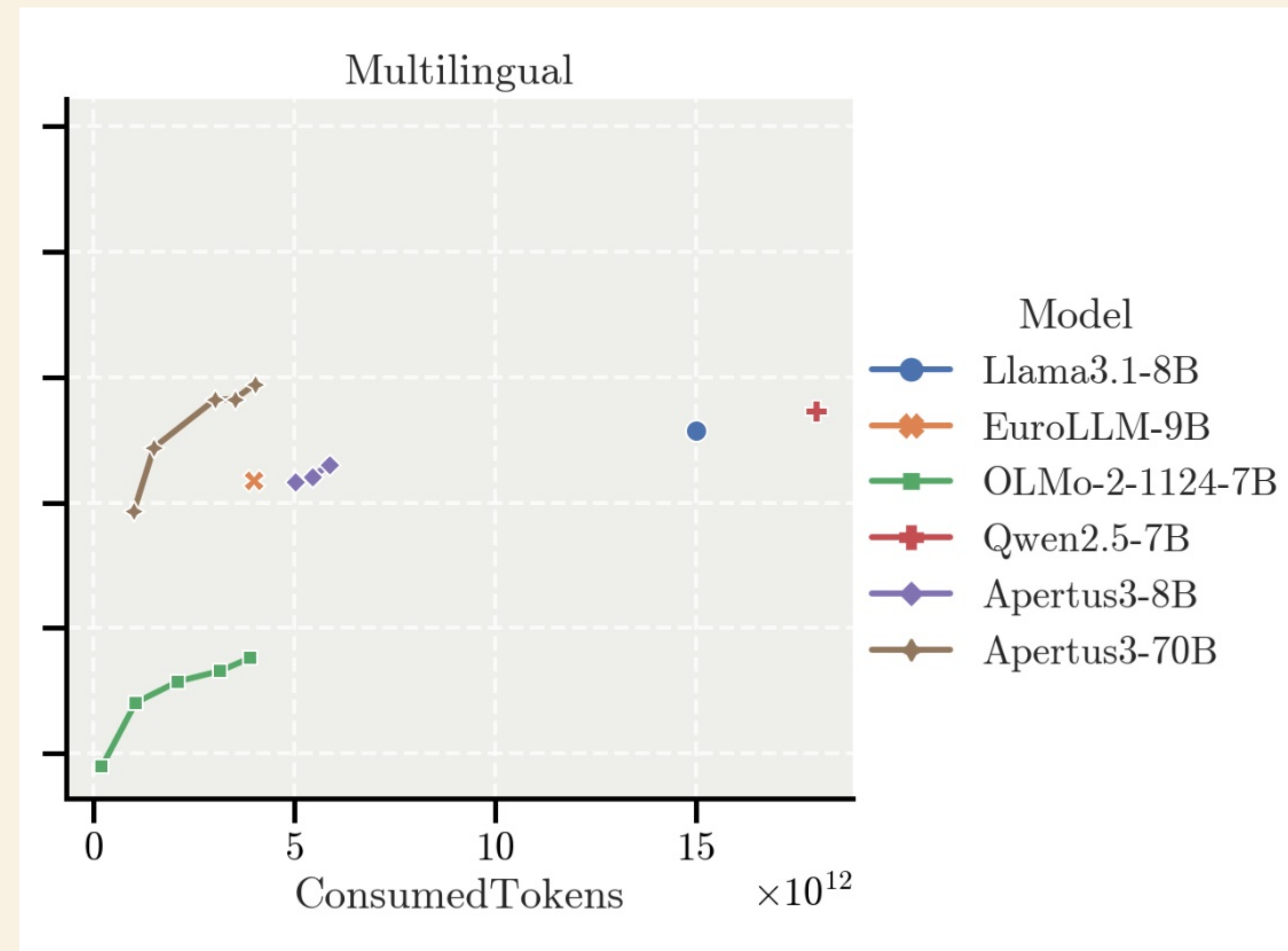
Massive

Multitask

Multilingual

Language Understanding

Good English Performance; **Great** Multilingual Capabilities





**Antoine
Bosselut**



**Martin
Jaggi**



**Imanol
Schlag**



**Antoni-Joan
Solergibert**



**Bettina
Messmer**



**Negar
Foroutan**



**Angelika
Romanou**



**Alexander
Hägele**



**Alejandro
Hernández
Cano**



**Vinko
Sabolčec**



**Skander
Moalla**



**Allen
Huang**



**Maximilian
Böther**



Simin Fan



**Ihor
Protsenko**



**Matteo
Pagliardini**



**Anna
Sotnikova**



**Zeming
Chen**



**Badr
AlKhamissi**



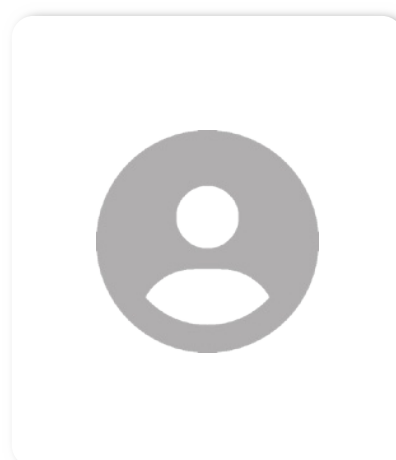
**Syrielle
Montariol**



**Paul
Teiletche**



**Andreas
Marfurt**



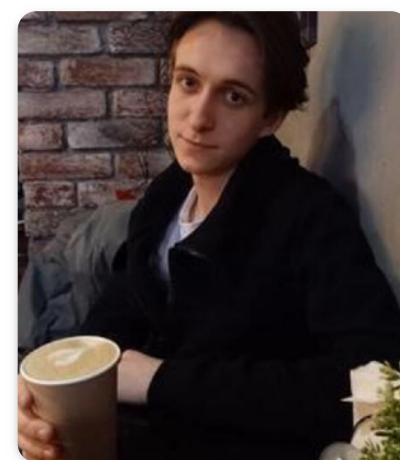
**Kyle
Matoba**



**Roman
Macháček**



Jan Deriu



**Andrei
Semenov**



**Auguste
Poiroux**



**Dhia
Garbaya**



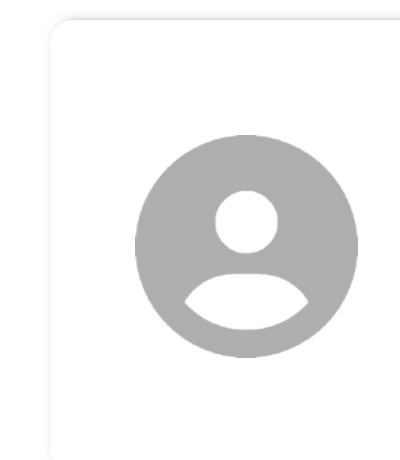
**Dongyang
Fan**



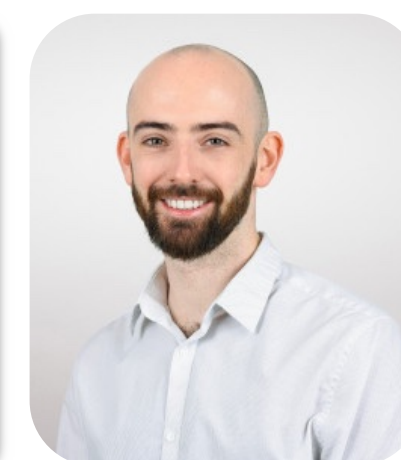
**Ayush
Kumar Tarun**



**Matin
Ansaripour**



**Juan
Garcia**



**Nicholas
Browning**



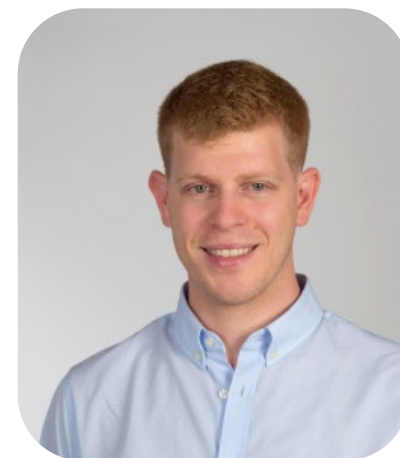
**Lukas
Drescher**



**Henrique
Mendonça**



**Theofilos
Manitaras**



**Stefano
Schuppli**



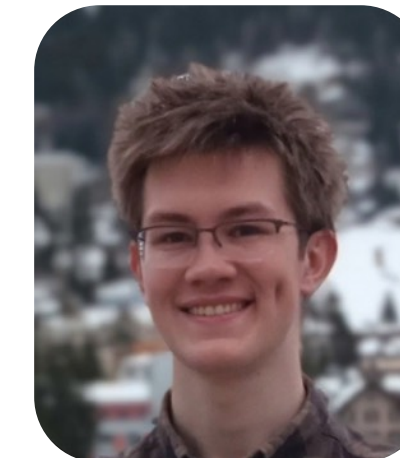
**Joost
VandeVondele**



**Fawzi Roberto
Mohamed**



**Eduard
Durech**



Ilia Badanin



**Tiancheng
Chen**

