



**SORBONNE
UNIVERSITÉ**



EPFL ETH

Apertus: Democratizing Open and Compliant LLMs For Global Language Environments

Syrielle Montariol

LTG seminar

03-09-2026

In this talk...

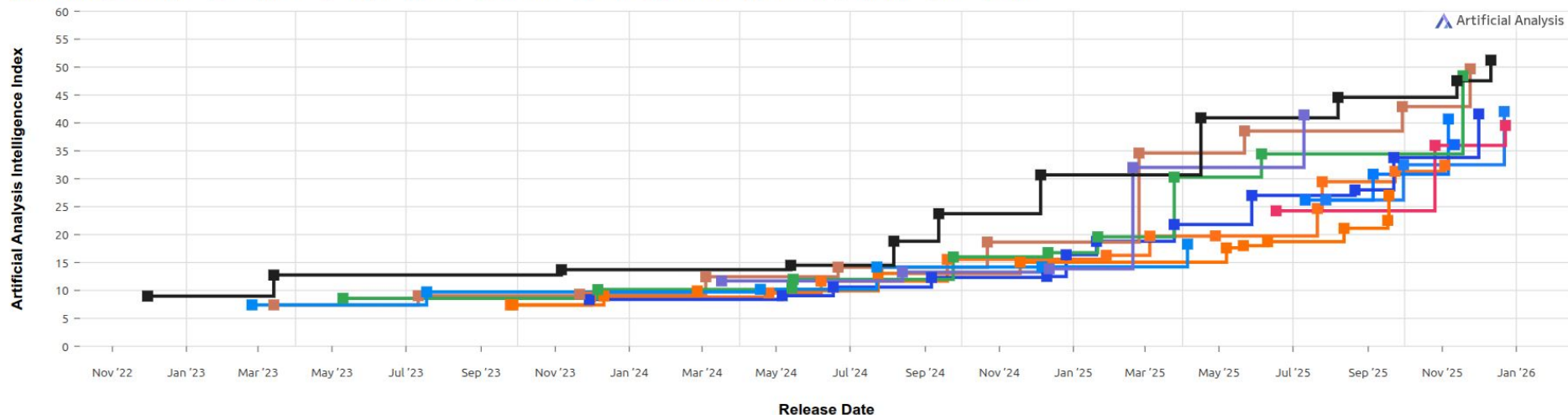
- AI regulation and the EU AI Act
- Apertus: a **multilingual** **compliant** **fully-open** LLM
 - a. Compliance
 - b. Multilinguality

AI is Changing the World

Frontier Language Model Intelligence, Over Time

Artificial Analysis Intelligence Index v4.0 incorporates 10 evaluations: GDPval-AA, τ^2 -Bench Telecom, Terminal-Bench Hard, SciCode, AA-LCR, AA-Omniscience, IFBench, Humanity's Last Exam, GPQA Diamond, CritPt

Alibaba Anthropic DeepSeek Google Kimi KwaiKAT Meta MiniMax Mistral OpenAI xAI Z AI



① Artificial Analysis Intelligence Index

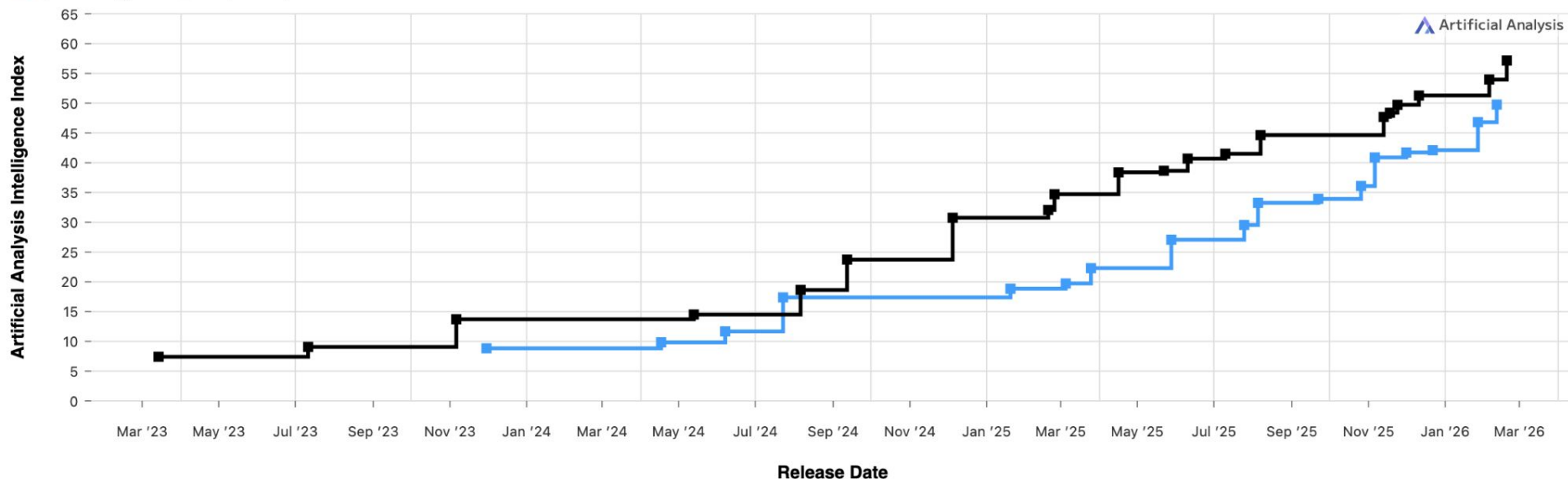
Artificial Analysis Intelligence Index v4.0 includes: GDPval-AA, τ^2 -Bench Telecom, Terminal-Bench Hard, SciCode, AA-LCR, AA-Omniscience, IFBench, Humanity's Last Exam, GPQA Diamond, CritPt. See [Intelligence Index methodology](#) for further details, including a breakdown of each evaluation and how we run them.

AI is Changing the World

Progress in Open Weights vs. Proprietary Intelligence

Artificial Analysis Intelligence Index v4.0 incorporates 10 evaluations: GDPval-AA, τ^2 -Bench Telecom, Terminal-Bench Hard, SciCode, AA-LCR, AA-Omniscience, IFBench, Humanity's Last Exam, GPQA Diamond, CritPt

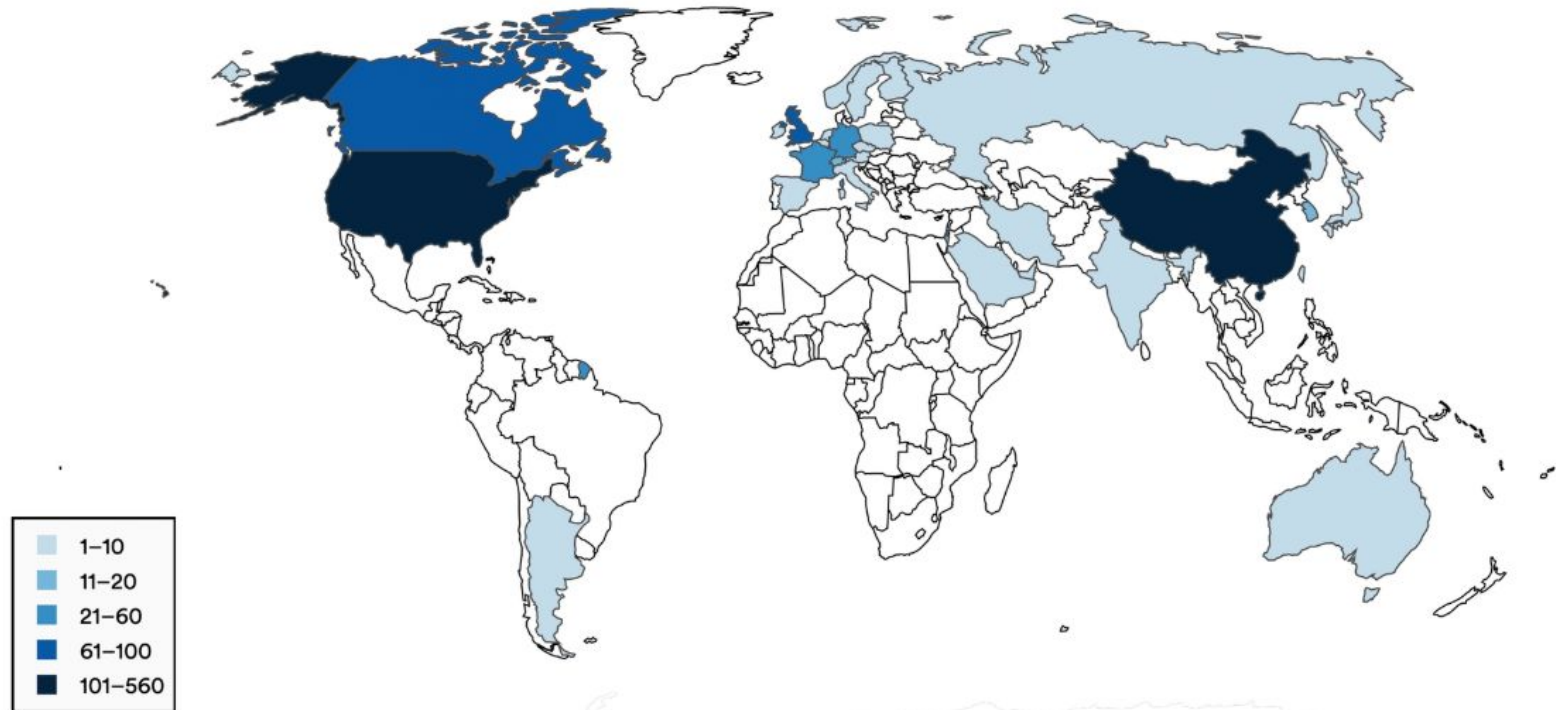
■ Open Weights ■ Proprietary



Who is building AI?

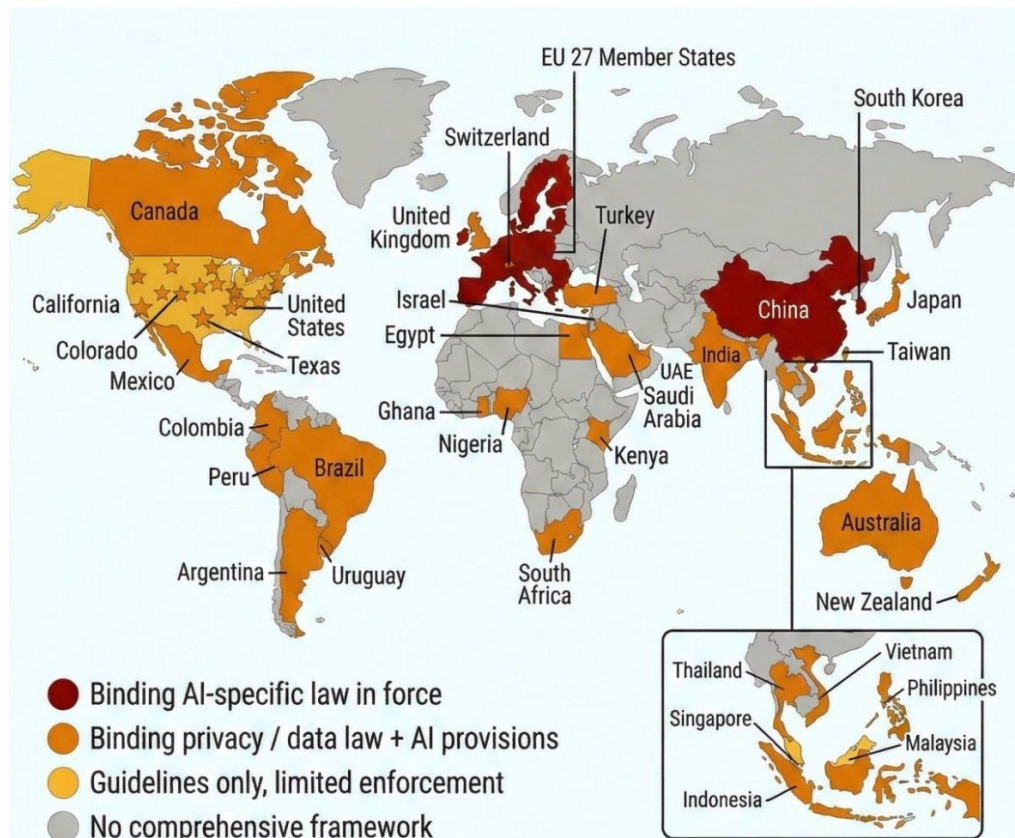
Number of notable AI models by geographic area, 2003–24 (sum)

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



Attempts at AI regulation

AI regulation in the world

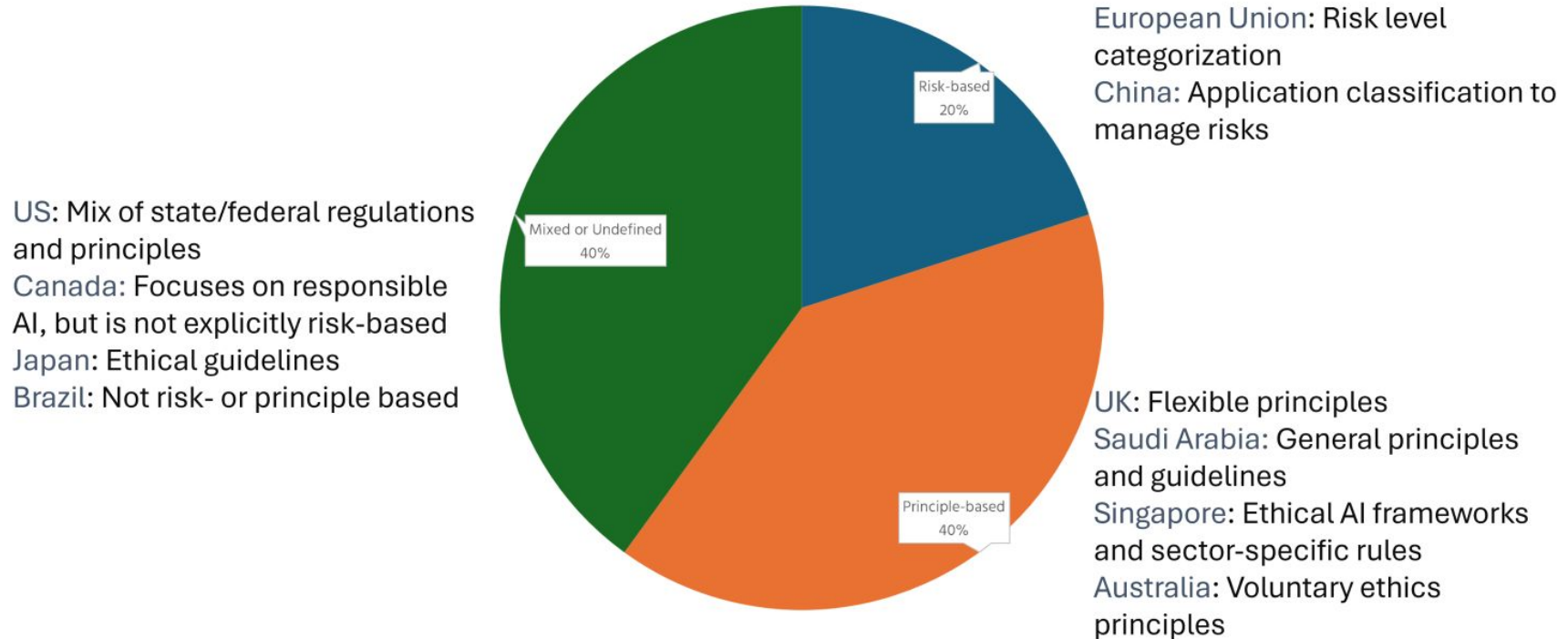


72+ countries with AI policy initiatives

19 US states with comprehensive privacy laws

€7.1B GDPR fines since 2018 (as of Jan 2026)

AI regulatory approaches: risk-based vs. principle-based



The EU AI Act: General purpose AI (GPAI)

GPAI model

- Trained with a large amount of data using self-supervision.
- Displays significant generality

GPAI system

- Application or product built around and powered by a GPAI model

The AI Act primarily **regulates AI systems using a risk-based framework**. GPAI models are an exception, regulated directly based on **model capabilities and compute thresholds**.

The EU AI Act: History

2017: Parliament and Council call for common EU AI rules and an ethical/legal framework.

2019: Commission issues **Ethics Guidelines for Trustworthy AI** (non-binding soft law).

2020: Commission launches a broad public consultation and published an Impact Assessment of the regulation on AI.

2020–2021: Parliament adopts resolutions asking the Commission to establish a legal framework of ethical principles for the development, deployment and use of AI,

2021: Commission proposes the EU Artificial Intelligence Act, establishing harmonized EU rules for AI development and use.

EU AI Act timeline

How the text got here

21 Apr 2021	Commission proposal. Regulates AI systems by purpose — no provisions on foundation models at all.
30 Nov 2022	ChatGPT. The Council adopts its general approach six days later, on a text already overtaken.
Jun–Dec 2023	Parliament retrofits a GPAI chapter. FR/DE/IT push for self-regulation; agreement after a 36-hour trilogue.
1 Aug 2024	Regulation (EU) 2024/1689 enters into force. Obligations phased over four years.
19 Nov 2025	Digital Omnibus proposed. Authorities undesignated, harmonised standards unfinished.
27 Jul 2026	Regulation (EU) 2026/1744 in force — six days before the deadline it postpones.

What actually applies, and when

2 Feb 2025	Prohibitions; AI literacy	
2 Aug 2025	GPAI models (Ch. V), governance, penalties	<i>binds Apertus and OpenEuroLLM</i>
2 Aug 2026	Remainder of the Act; Art. 50 transparency	◀ today
2 Dec 2026	Art. 50(2) for legacy systems; NCII + CSAM prohibitions	
2 Aug 2027	Legacy GPAI models must comply; sandboxes operational	
2 Dec 2027	High-risk, Annex III standalone	<i>was 2 Aug 2026</i>
2 Aug 2028	High-risk, Annex I embedded	<i>was 2 Aug 2027</i>

EU AI Act risks for GPAI systems

AI System Risk Tier	EU AI Act Article	Obligations
Unacceptable Risk	Article 5	Prohibited Deadline: February 2025 Examples of AI: AI systems that are deceptive, exploitative, or discriminatory (e.g. social scoring, facial recognition, biometrics inferences), and that result in significant harm, including to civil rights
High Risk	Article 6	Extensive requirements, including transparency and conformity assessments Deadline: February 2026, August 2026, August 2027 Examples of AI: Critical infrastructure, safety components (aviation, vehicles, medical devices), education and vocational training, employment, essential private and public services (e.g. healthcare, banking), and more
Limited Risk	Article 50	Transparency requirements Deadline: August 2026 Examples of AI: Chatbots, AI-generated content
Minimal Risk		No obligations Deadline: N/A Examples of AI: Spam filters, video games, systems that aren't classified in the above risk tiers

GPAL Systems with Limited risk:

Transparency requirements

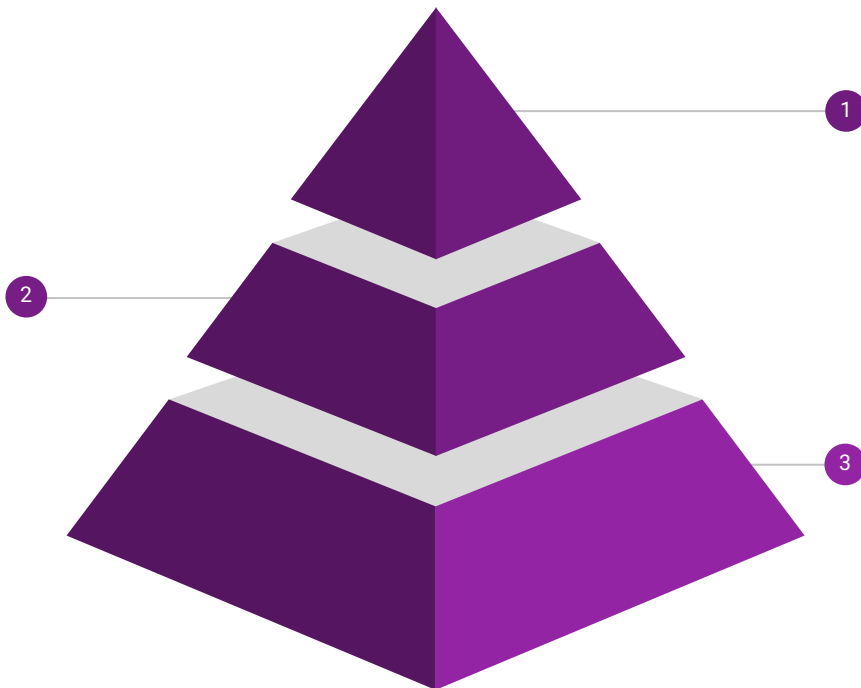
1. Systems that interact directly with individuals must **inform users that they are interacting with an AI system.**
2. Content that is either **generated or modified with AI** need to be clearly **labelled** as AI generated.

EU AI Act classification of GPAI Models

General Purpose AI Models (GPAI)

- Trained using more than **10^{23} FLOPS** and capable of generating text, audio, image, video output
- Displays significant generality and can perform a wide range of tasks

→ Transparency requirements



GPAI Models with systemic risks

Trained using more than **10^{25} FLOPS** (>250B parameters)

→ Transparency requirements, risk assessment and mitigation

Narrow AI models

Specialised models, regardless of size (e.g., for transcription, image upscaling, weather forecasting, or gaming)

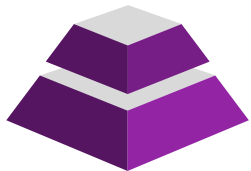
→ No requirements

GPAI models: Transparency requirements

1. Establish a policy to **respect the Copyright Directive**.
2. Publish **technical documentation** :
 - a. **Training, testing and evaluation process**.
 - b. **Detailed summary of training content and data**
 - c. **Documentation to downstream providers** that integrate the GPAI model into an AI system to list **capabilities and limitations**.

2.1 Pretraining Data

Our training corpus includes a new mix of data from publicly available sources, which does not include data from Meta's products or services. We made an effort to remove data from certain sites known to contain a high volume of personal information about private individuals. We trained on 2 trillion tokens of data as this provides a good performance–cost trade-off, up-sampling the most factual sources in an effort to increase knowledge and dampen hallucinations.



GPAI models with systemic risk:

Risk assessment and mitigation requirements

1. Perform **model evaluations**, including conducting and documenting **adversarial testing** to **identify systemic risk**.
2. **Mitigate systemic risks** throughout the lifecycle.
3. **Track, document and report serious incidents** and possible corrective measures
4. Ensure an adequate level of **cybersecurity protection**.



Model openness

Closed

 Open-AI
 Anthropic
 Google
 xAI
 Mistral

Share outputs.

Open-Weight

 Meta (Llama)
 DeepSeek
 Alibaba (Qwen)
 Moonshot (Kimi)
 Google (Gemma)
 Z.AI (GLM)
 Mistral (Mixtral)

Share the model weights.

Fully Open

  Swiss AI (Apertus)
  AI2 (Olmo)
  MBZUAI IFM (K2)
  OpenEuroLLM
  EuroLLM
  AI Singapore (SEA-LION)
(many smaller ones)

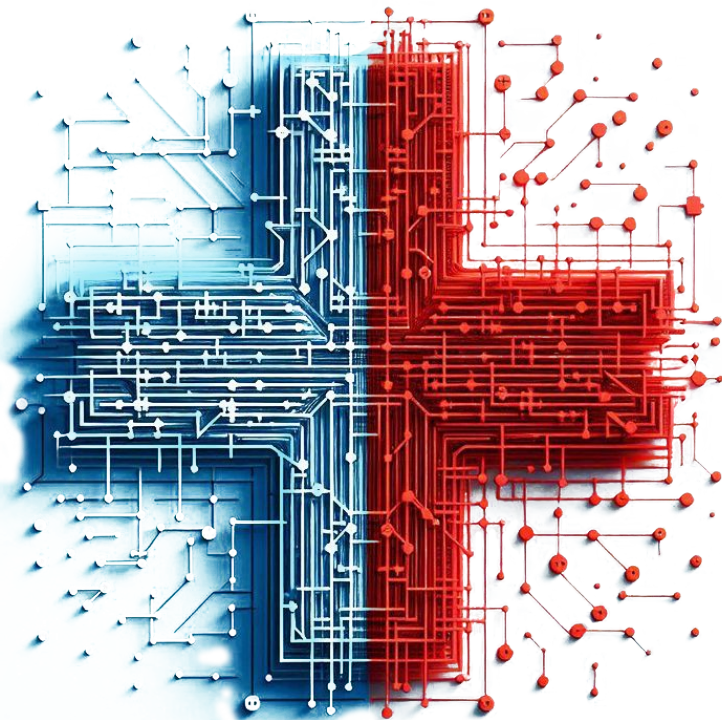
Share everything.



= Non-profit institutions

Why Build Our Own Models?

- **Legal necessity:** Transparency and aligned to our values; responsible AI
- **R&D autonomy:** Focus on what matters to us (multilinguality, compliance, local values)
- **Scientific integrity:** Open data and reproducibility are foundational to science
- **Deployment control:** On-premise, no vendor lock-in, data stays internal if you want
- **Teach the next generation:** “What I cannot create, I do not understand”



APERTVS

EPFL

ETH zürich



CSCS

The Swiss National Supercomputing Center

Alps Supercomputer by CSCS: **10,000+ GH200 GPUs** each with 96GB

- 100% of its electricity is from hydropower
- Cooled using water pumped from Lake Lugano in a closed cycle
- Currently building a thermal power plant to recover waste heat

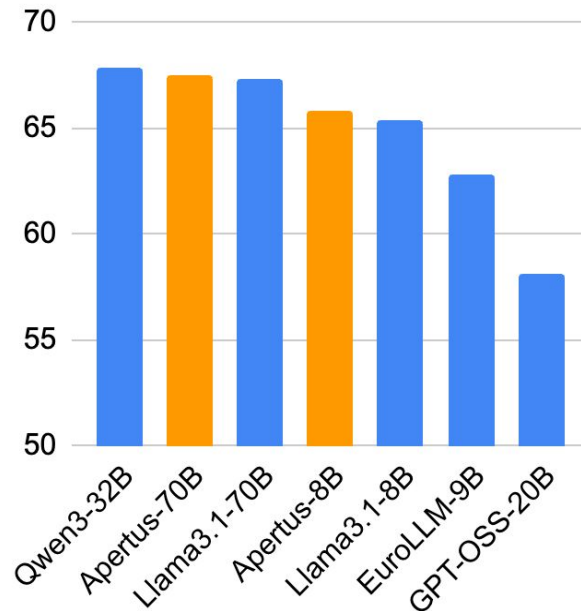


Microturbines for energy recovery
(Photo: CSCS)

Apertus: A transparent and responsibly-trained multilingual LLM

- **Scale (8 & 70B):** the first fully open model to be trained at the 70B scale
- **Open & transparent:** Released code, intermediate and final checkpoints. Reproducible data. Permissive license.
- **Compliance:** Trained only on public data, respecting AI opt-outs, preventing memorisation.
- **Multilingual from scratch:** Trained on data from over 1800 languages
- **Strong performance:** Most capable fully-open models at respective scales
- **Sovereignty:** Open platform for research and development of responsible AI

Strong English benchmark performance within class of fully-open models



Released Models, Code and Datasets

Models:

- 🤖 [swiss-ai/Apertus-8B-2509](#)
- 🤖 [swiss-ai/Apertus-70B-2509](#)
- 🤖 [swiss-ai/Apertus-8B-Instruct-2509](#)
- 🤖 [swiss-ai/Apertus-70B-Instruct-2509](#)

Code:

- 🤖 [swiss-ai/Megatron-LM](#)
- 🤖 [swiss-ai/pretrain-data](#)
- 🤖 [swiss-ai/pretrain-code](#)
- 🤖 [swiss-ai/posttraining](#)
- 🤖 [swiss-ai/posttraining-data](#)
- 🤖 [swiss-ai/evals](#)
- 🤖 [swiss-ai/lm-evaluation-harness](#)
- 🤖 [swiss-ai/apertus-format](#)
- 🤖 [swiss-ai/hfconverter](#)

Datasets & Auxiliary Tools:

- 🤖 [swiss-ai/apertus-finetuning-recipes](#)
- 🤖 [swiss-ai/apertus-memorization](#)
- 🤖 [swiss-ai/apertus-pretrain-toxicity](#)
- 🤖 [swiss-ai/apertus-pretrain-gutenberg](#)
- 🤖 [swiss-ai/apertus-pretrain-poisonandcanaries](#)
- 🤖 [swiss-ai/apertus-posttrain-romansh](#)
- 🤖 [swiss-ai/africa-preferences](#)
- 🤖 [swiss-ai/africa-sft](#)

- 🤖 [swiss-ai/switzerland.qa](#)

Separately Released Related Scientific Publications by SwissAI team members

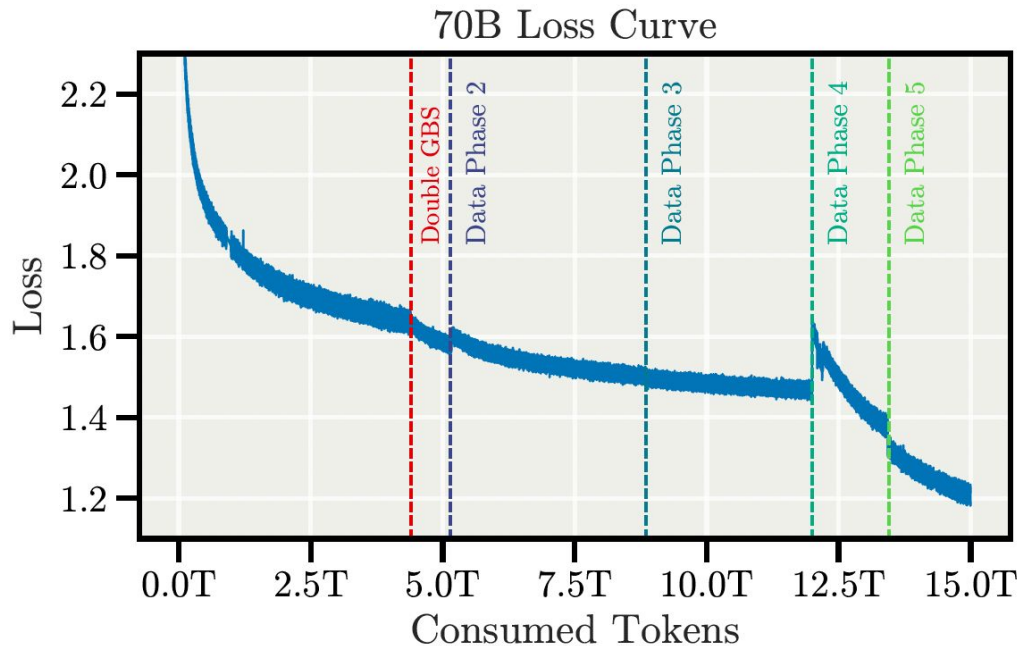
- Data compliance gap when respecting training data opt-out (Fan et al., 2025)
- FineWeb-2 dataset (Penedo et al., 2025)
- FineWeb-2-HQ dataset (Messmer et al., 2025)
- Memorization dynamics (Xu et al., 2025)
- Multilingual evaluation (Romanou et al., 2025; Singh et al., 2025)
- xIELU activation function (Huang & Schlag, 2025)
- FP8 (Hernández-Cano et al., 2025) and outlier protected block (He et al., 2024)
- Warmup-Stable-Decay Learning Rates (Hägele et al., 2024; Dremov et al., 2025)
- AdEMAMix optimizer (Pagliardini et al., 2025)
- Optimizer benchmarking (Semenov et al., 2025)
- QRPO post-training (Matrenok et al., 2025)
- Contrastive language identification (Foroutan et al., 2025b)
- Parity-aware tokenization (Foroutan et al., 2025a)
- Training data indexing (Marinas et al., 2025)
- Training data attribution (Wuhrmann et al., 2025)
- Data mixtures during pretraining (Böther et al., 2025)

Architecture and Training

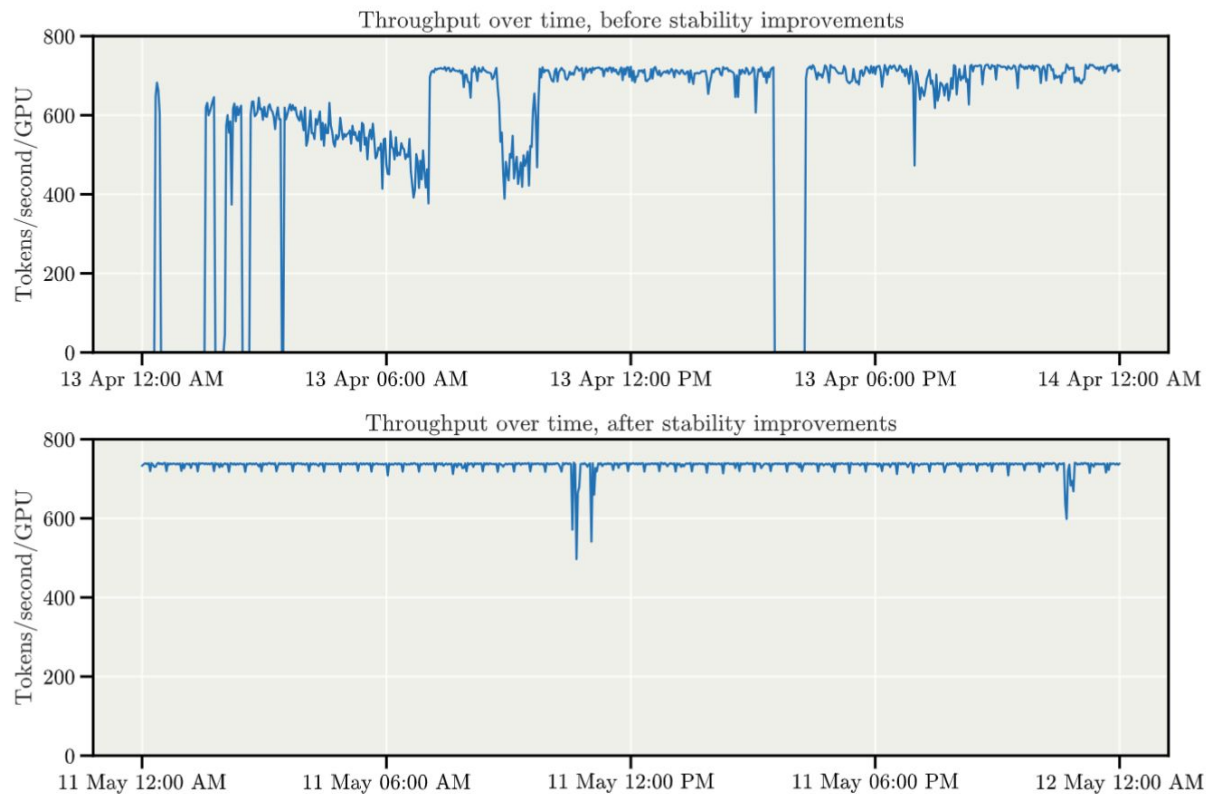
Full Training Run

Training of the 70B model:

- **15T tokens** (214 tokens per parameter)
- **6.74×10^{24} FLOPs** (below the 10^{25} threshold)
- ~ 90 days on 1024 nodes (**4096 GPUs**) on **Alps** (8M GPU hours)
- 5 GWh power usage for the compute of the **pretraining run**, carbon-neutral



Training at Scale Is Hard



Architecture

Dense decoder-only Transformer

Model	Layers	Dim	MLP Dim	Heads (Q / KV)	Activation	Context Length
Apertus 8B	32	4096	21504	32/8	xIELU	65536
Apertus 70B	80	8192	43008	64/8	xIELU	65536

Context of **4,096 tokens during pretraining**
+ **Long-context extension** to support sequences up to 65,536 tokens.

Model and Training Innovations

OPTIMIZER

AdEMAMix

+ Goldfish Loss

Published as a conference paper at ICLR 2025

**THE ADEMAMIX OPTIMIZER:
BETTER, FASTER, OLDER**

Matteo Pagliardini [†]
EPFL

Pierre Ablin
Apple

David Grangier
Apple

Model and Training Innovations

OPTIMIZER

AdEMAMix

ARCHITECTURE

xIELU

**Activation +
QK-Norm**

Deriving Activation Functions Using Integration

Allen Hao Huang¹ Imanol Schlag²

Model and Training Innovations

OPTIMIZER

AdEMAMix

ARCHITECTURE

**xIELU
Activation +
QK-Norm**

LR SCHEDULE

WSD
**For Continual
Training**

Published in Transactions on Machine Learning Research (08/2025)

**Training Dynamics of the Cooldown Stage in
Warmup-Stable-Decay Learning Rate Scheduler**

Aleksandr Dremov, Alexander Hägele, Atli Kosson, Martin Jaggi
EPFL, Switzerland

Ablation: Comparison with OLMo2 framework

Model	Activation	Loss	Normalization	Optimizer
Apertus OLMo2	xIELU SwiGLU	Goldfish Z-Loss	Pre Norm Reordered Norm	AdEMAMix AdamW

46%

fewer tokens
to match OLMo2 1B

CE loss after 20k steps: Apertus 2.75 vs
OLMo2 2.84

30%

fewer tokens
to match OLMo2 7B

CE loss after 20k steps: Apertus 2.51 vs
OLMo2 2.56

Training at Scale Is a Systems Problem

An Engineering Journey Training Large Language Models at Scale on Alps: The Apertus Experience

Jonathan Coles*, Stefano Schuppli*, Lukas Drescher*, Fawzi Roberto Mohamed*, Elia Palme*, Henrique Mendonça*, Miguel Gila*, Mark Klein*, Maxime Martinasso*, Joost VandeVondele*, Torsten Hoefler*, Thomas Schulthess*, Josh Romero[†], Igor Gorodetsky[‡], Ryan Hankins[‡], Isa Wazirzada[‡], Martin Jaggi[§], Antoine Bosselut[§], Imanol Schlag[¶], Antoni-Joan Solergibert i Llaquet[§], Alejandro Hernández Cano[§], Theofilos Ioannis Manitaras, Nicholas John Browning

*Swiss National Supercomputing Centre (CSCS)

[†]NVIDIA

[‡]HPE

[§]EPFL

[¶]ETH Zurich

Ninety Days at 4096 GPUs

Hardware, drivers, operating system

Storage and networking

Things that only break at scale

Things that didn't work as expected

FP8 training

+26% throughput, then the loss diverged after ~300B tokens. Rolled back to BF16.

**The 70B
cooldown**

No slope change, no benchmark jump — unlike the 8B and unlike small-scale results.

**Gradient
clipping**

AdEMAMix needed a clip value of 0.1, so clipping fires at almost every step.

Compliance

Growing wave of lawsuits over the use of copyrighted content for AI training

Meta staff torrented nearly 82TB of pirated books for AI training — court records reveal copyright violations

News By Jowi Morales published February 9, 2025

AI

Anthropic cut up millions of used books to train Claude — and downloaded over 7 million pirated ones too, a judge said

Perplexity is using stealth, undeclared crawlers to evade website no-crawl directives

2025-08-04

French publishers and authors file lawsuit against Meta in AI case

By Reuters

March '2, 2025 12:30 PM GMT+1 · Updated a month ago



The Times Sues OpenAI and Microsoft Over A.I. Use of Copyrighted Work

Millions of articles from The New York Times were used to train chatbots that now compete with it, the lawsuit said.

Lawyer Says Studio Ghibli Could Take Legal Action Against OpenAI

OpenAI could be "blurring the line between homage and outright misrepresentation."

Artificial Intelligence / Artificial Intelligence / Copyright / Copyright

Responsible and Compliant LLM criteria

1. Only use publicly available data
2. Respect AI-opt out
3. Prevent memorization at scale
4. Remove personal identifiable information (PII)
5. Remove toxic & harmful data

Responsible and Compliant LLM criteria

1. Only use publicly available data
2. Respect AI-opt out
3. Prevent memorization at scale
4. Remove personal identifiable information (PII)
5. Remove toxic & harmful data



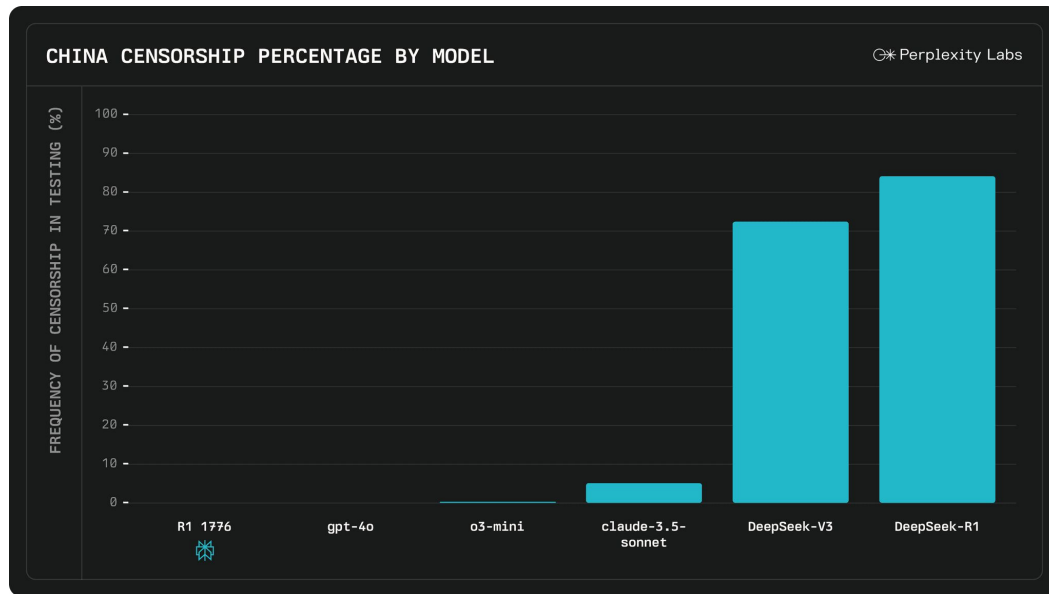
Any security and safety guardrails / alignment can be reverted through post-training!

→ **Implement data compliance measures during pretraining data construction**

Open-weight model safety

Example of DeepSeek-R1:

- Refuses to answer questions about some sensitive topics
- Post-trained using tailor-made dataset of 300 censored topics, 40k multilingual prompts that trigger censorship, and CoT reasoning and answers for them.
- Censorship reverted!



This is a claim about open weights in general: If you release weights, anything you did at the alignment stage is removable by anyone with modest compute.

Compliance:

- a. Data cleaning

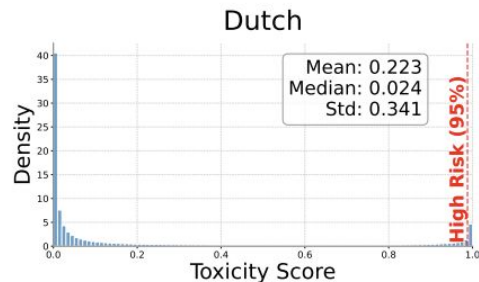
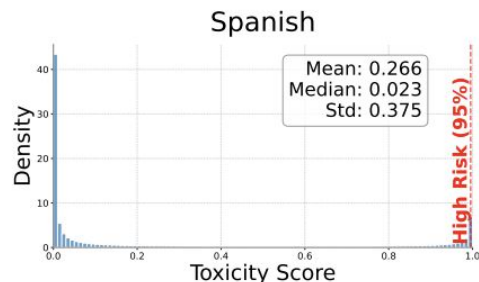
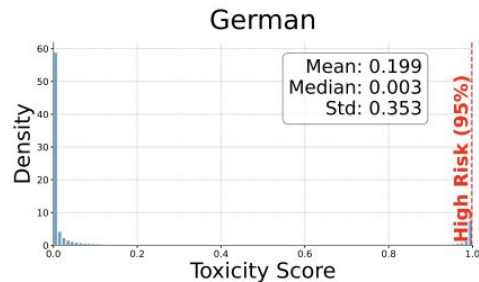
1. Remove personal identifiable information (PII)

Goal: protect against potential memorization of PII in the model

1. **Regexp to detect**
 - email addresses
 - IP addresses
 - IBAN bank account numbers
2. **Replace them** with anonymous markers, such as <email-pii>

2. Remove toxic & harmful data

1. Train **language-specific classifiers** using existing annotated datasets (PleIAs, SWSR) across 9 languages
2. **Annotate toxicity scores** for FineWeb-2 and FineWeb documents
3. **Filter the 5% of documents per language with the highest toxicity scores** from the pretraining corpus.



Compliance:

b. AI opt-outs

Data compliance

- Pretraining datasets = aggregation of snapshots taken from web crawls
- Content owners are opting out of web crawling via mechanisms like robots.txt

robots.txt

Article [Talk](#)

From Wikipedia, the free encyclopedia

robots.txt is the [filename](#) used for implementing the **Robots Exclusion Protocol**, a standard used by [websites](#) to indicate to visiting [web crawlers](#) and other [web robots](#) which portions of the website they are allowed to visit.

nytimes.com/

Disallow Rules

User-agent: AliyunSecBot
Disallow: /

User-agent: anthropic-ai
Disallow: /

User-agent: Applebot-Extended
Disallow: /

User-agent: AudigentAdBot
Disallow: /

User-agent: AwarioRssBot
User-agent: AwarioSmartBot
Disallow: /

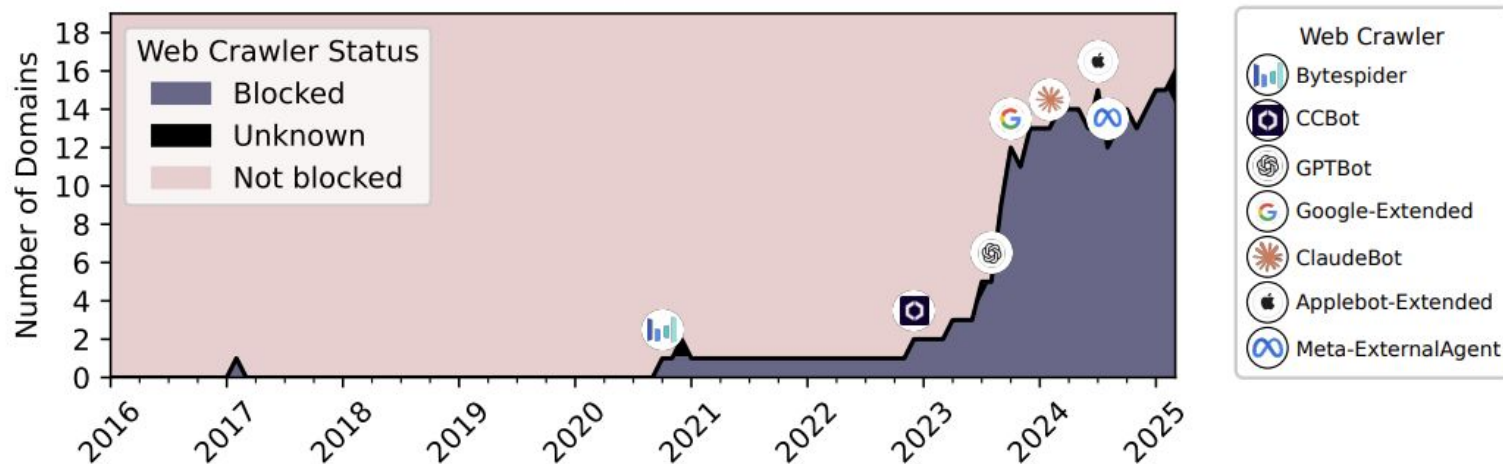
User-agent: BLEXBot
Disallow: /

User-agent: Bytespider
Disallow: /

...

Rapid increase of data opt-outs since Mid 2023

Figure: Number of Top 20 URL domains that have at least an opt-out



Data compliance

What is the impact on performance when training an LLM that respects data owners' opt-out requests?

Can Performant LLMs Be Ethical?
Quantifying the Impact of Web Crawling Opt-Outs

Dongyang Fan[♡], Vinko Sabolčec[♡], Matin Ansaripour[♡], Ayush Kumar Tarun[♡],
Martin Jaggi^{♡*}, Antoine Bosselut^{♡*}, Imanol Schlag^{♢*}

Main excluded crawling bots

```
"AI2Bot",           # AI2
"Applebot-Extended", # Apple
"Bytespider",       # Bytedance
"CCBot",            # Common Crawl
"CCBot/2.0",        # Common Crawl
"CCBot/1.0",        # Common Crawl
"ClaudeBot",        # Anthropic
"cohere-training-data-crawler", # Cohere
"Diffbot",          # Diffbot
"Meta-ExternalAgent", # Meta
"Google-Extended",  # Google
"GPTBot",           # OpenAI
"PanguBot",         # Huawei
"*"
```

Retroactive Robots.txt Filtering

Issue: Documents from opted-out content publishers can exist in the archived snapshots !

1. Retrieve each URL's *robots.txt* file as of January 2025
2. Retroactively filter out data not allowed for crawling

Top 20 filtered domains:

1. reference.com	General Reference	11. dictionary.com	Dictionary & Thesaurus
2. stackexchange.com	Q&A Communities	12. webmd.com	Health & Medical Info
3. teacherspayteachers.com	Educational Resources	13. lessonplanet.com	Educational Resources
4. wikia.com	Fandom Wikis	14. elsevier.com	Science & Technology
5. thefreedictionary.com	Dictionary & Thesaurus	15. biomedcentral.com	Science & Technology
6. phys.org	Science & Technology	16. chron.com	News & Media
7. bbc.co.uk	News & Media	17. pbs.org	Educational Resources
8. theguardian.com	News & Media	18. definitions.net	Dictionary & Thesaurus
9. wordpress.com	Blogging Platform	19. medicalxpress.com	Health & Medical Info
10. nytimes.com	News & Media	20. scientificamerican.com	Science & Technology

**High-value
Informational Content
are Most Affected!**

Retroactive Robots.txt Filtering

Dataset	Before filtering	After filtering
FineWeb-Edu (English)	4.9T	4.5T
FineWeb-2 (Multilingual)	47T	45T

We lose approximately 8% of English and 4% of multilingual data

A convenient set of filtering tools is available at data-compliance.github.io/tools !

High-value Informational Content are Most Affected

Top filtered topics and formats (% removed from corpus):

Topic

Science & Tech (2.65%)

Health (1.89%)

History (0.79%)

Format

News Article (2.50%)

Knowledge Article (1.62%)

Nonfiction Writing (1.29%)

*Domain tags are from WebOrganizer by Wettig et al.
(2025)*

Reduction of Non-permissible Content data

- 50-gram match rate with content from non-permissible URL domains: compliant vs. non-compliant corpora

	nytimes.com	medicalxpress.com	stackexchange.com
Non-compliant corpus	0.354%	0.702%	0.539%
Compliant corpus	0.267%	0.637%	0.348%

- The reduction is limited, because
 - a. Content republishing through forums / personal blogs
 - b. Open-source information appearing on non-permissible sites

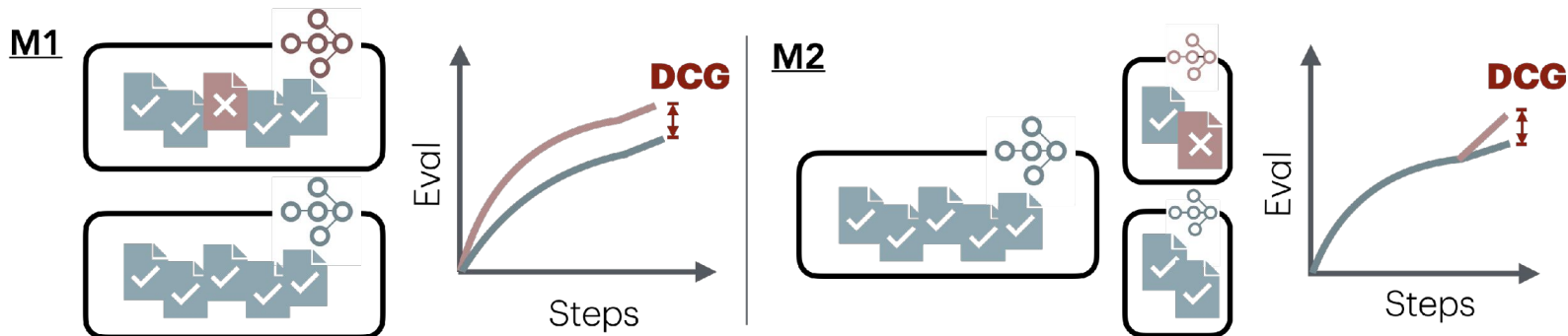
Do These Data Matter?

Measuring the Data Compliance Gap (DCG)

Data Compliance Gap (DCG): the downstream performance gap between models trained on compliant versus non-compliant corpora

- M1) from training from scratch
- M2) from continual pre-training / cooldown

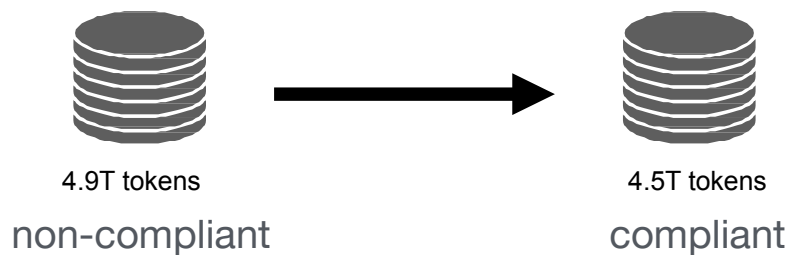
✓ Compliant data ✗ Non-compliant data ○○ Pre-training



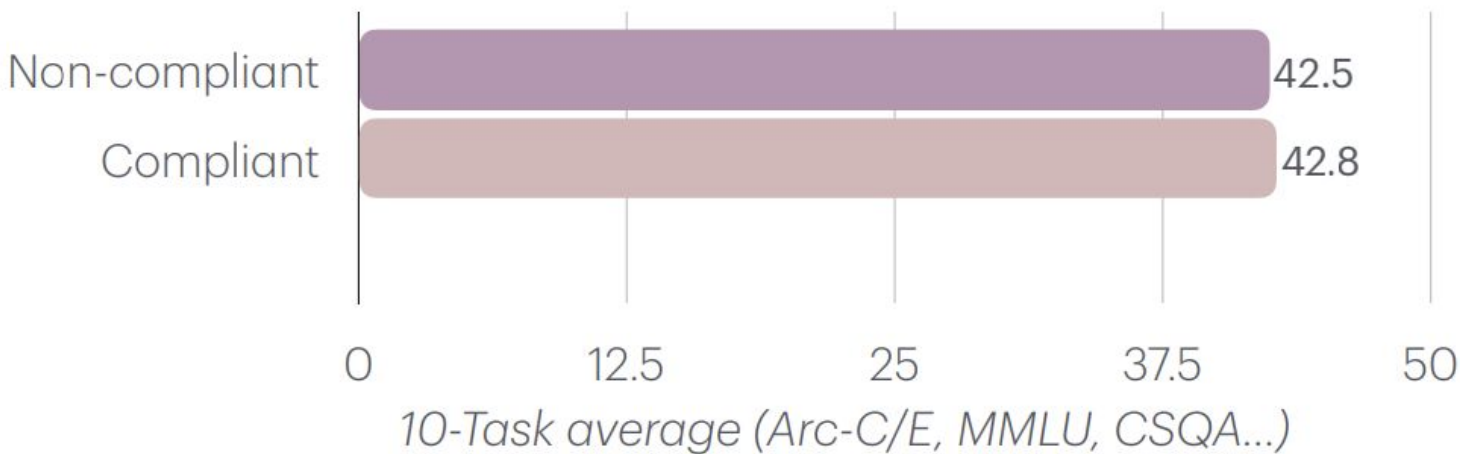
Do These Data Matter?

Measuring the Data Compliance Gap (DCG)

- Train 1.5B Llama-like models on 100B tokens sampled from FineWeb-Edu
- DCG = **downstream performance gap between models trained on compliant versus non-compliant corpora**



Does Compliance Limit Knowledge Learning?



Pretraining on compliant data does not impact general knowledge understanding

Knowledge can be supplemented by other sources

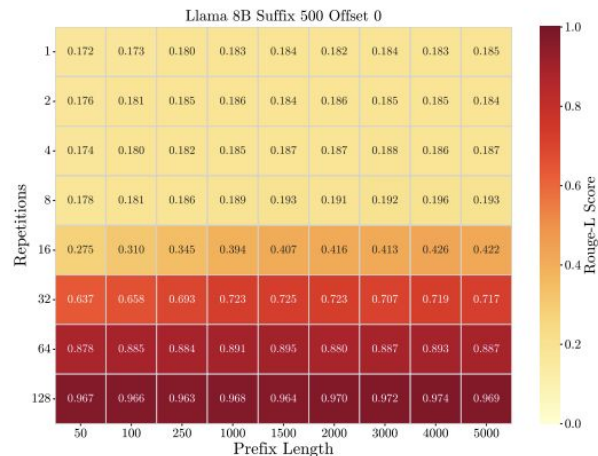
Compliance:

c. Prevent memorization

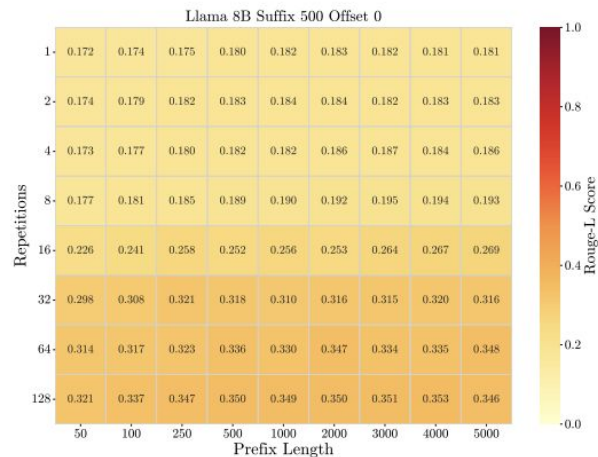
Prevent Memorization

- Verbatim regurgitation of training data raises **copyright and privacy risks**.
- **Goldfish Loss**: discourage the model from learning exact token-to-context mappings by **selectively masking tokens during pretraining**.

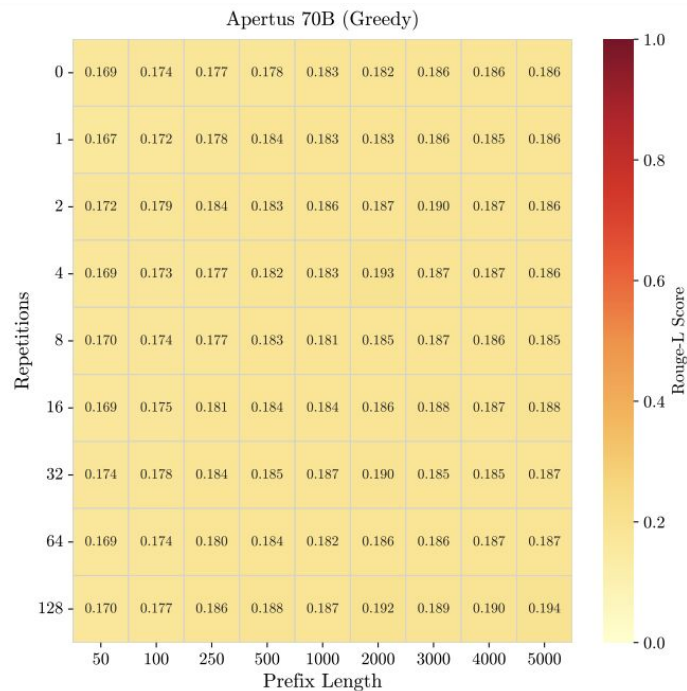
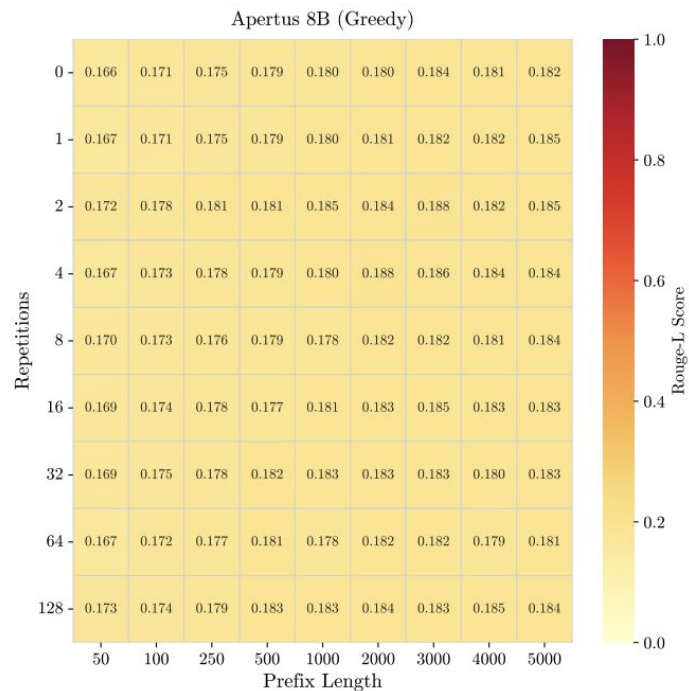
Standard loss



Goldfish loss



Prevent Memorization: Apertus results



Compliance:

d. Post-training

SFT data quality

License filtering.

(i) content must be **explicitly released under licenses permitting redistribution and commercial use** (e.g., CC-BY, Apache 2.0).

Ambiguous or unspecified licenses are excluded.

(ii) the collection procedure must be **fully documented and reproducible**.

(iii) systematically **exclude source datasets originating from opted-out providers** (e.g., Reddit, StackExchange)

Standard quality filtering.

Decontamination.

Alignment (QRPO) using reward models

2 distinct alignments:

1. **Non-controversial topics:**

- Quality criteria like correctness, helpfulness, and harmlessness
- Use pretrained reward model: *Skywork-Reward-V2-Llama-3.1-8B*

2. **Controversial topics** : Moral, political, social, and cultural values

- “Constitutional AI”: Human oversight is provided only through a list of rules or principles
- Create **The Swiss AI Charter**
- Prompt an LLM-as-judge (*Qwen3-32B*) to score completions by their level of adherence to the Charter

Alignment on controversial topics

The Swiss AI Charter (validated through public agreement):

1. Response Quality — Writing clear, accurate, and useful responses.
2. Knowledge and Reasoning Standards — Using verified facts and sound reasoning.
3. Respectful Communication — Treating people with courtesy, fairness, and accessibility.
4. Preventing Harm — Protecting safety and refusing harmful requests.
5. Resolving Value Conflicts — Handling trade-offs openly and preserving principles.
6. Professional Competence Boundaries — Educating without giving licensed advice.
7. Collective Decision-Making — Supporting fair and constructive group decisions.
8. Autonomy and Personal Boundaries — Respecting choice, privacy, and clear limits.
9. Long-term Orientation and Sustainability — Considering long-term impacts and risks.
10. Human Agency — Keeping humans in control and independent.
11. AI Identity and Limits — Being clear about what the AI is and is not.

Multilinguality

Pretraining Data

15T tokens of high quality international data mix

- **English:** FineWeb-HQ, FineWeb-Edu, DCLM-Edu
- **Non-English:** FineWeb-2-HQ, Translation Parallel Data, Wikipedia
- Stack v1.2
- FineMath and MegaMath
- Memorisation and Poisoning Data

~60% English web data

~40% Non-English web data

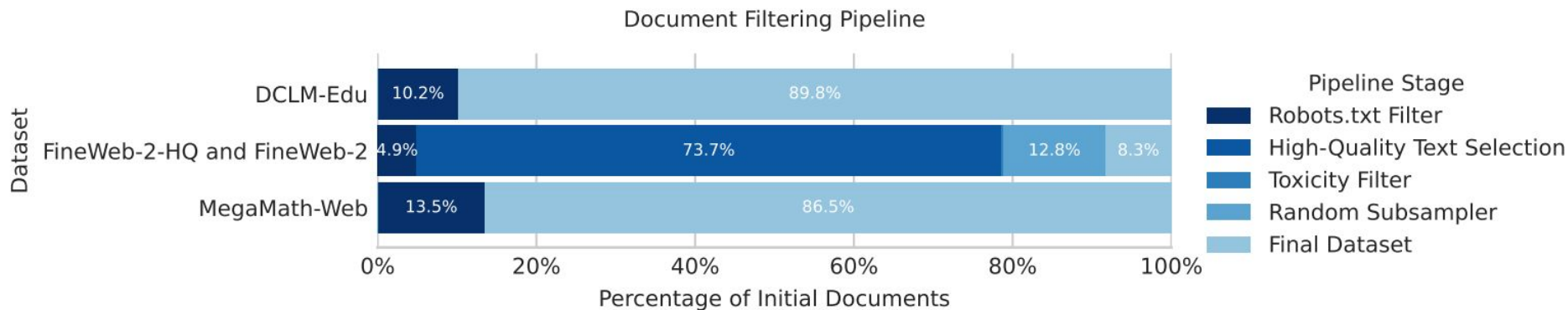
~Code & math data growing from 3%-10% throughout pretraining

English data selection

- FineWeb-HQ. High-quality dataset obtained by filtering FineWeb web-crawl data using XLM-RoBERTa-based classifiers with a focus on structured and knowledge-rich content
- FineWeb-Edu. High-quality dataset obtained by filtering FineWeb web-crawl data using a classifier focusing on educational content
- DCLM-Edu. High-quality dataset obtained by applying the FineWeb-Edu educational classifier on the DCLM dataset

All of the datasets can be seen as different, partially overlapping subsets from English CommonCrawl data.

All available on Huggingface!



Fineweb 2

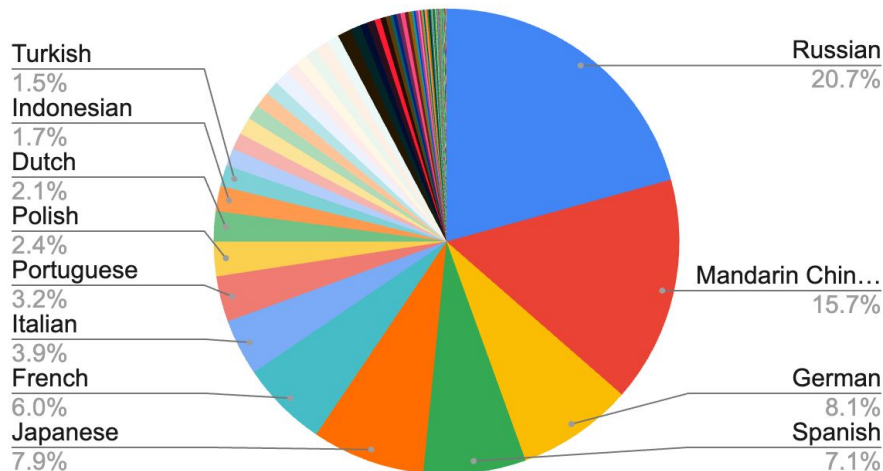
FineWeb2: One Pipeline to Scale Them All — Adapting Pre-Training Data Processing to Every Language

Guilherme Penedo^{EPFL} Hynek Kydlíček^{EPFL} Vinko Sabolčec^{EPFL} Bettina Messmer^{EPFL}
 Negar Foroutan^{EPFL} Amir Hossein Kargaran^{EPFL} Colin Raffel^{EPFL} Martin Jaggi^{EPFL}
 Leandro Von Werra^{EPFL} Thomas Wolf^{EPFL}
 Hugging Face ^{EPFL-EPFL}

🔗 Pipeline code: <https://github.com/huggingface/fineweb-2>

📁 FineWeb2 dataset: <https://hf.co/datasets/HuggingFaceFW/fineweb-2>

Fineweb-2 language distribution

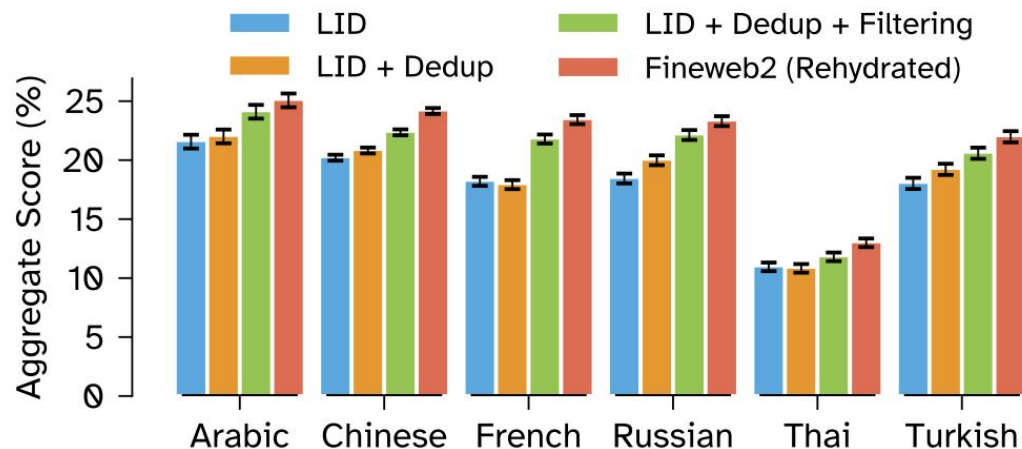


1811 languages

1868 language-script pairs

Fineweb 2 creation pipeline

1. LID
2. Deduplication
3. Filtering
4. Rehydration

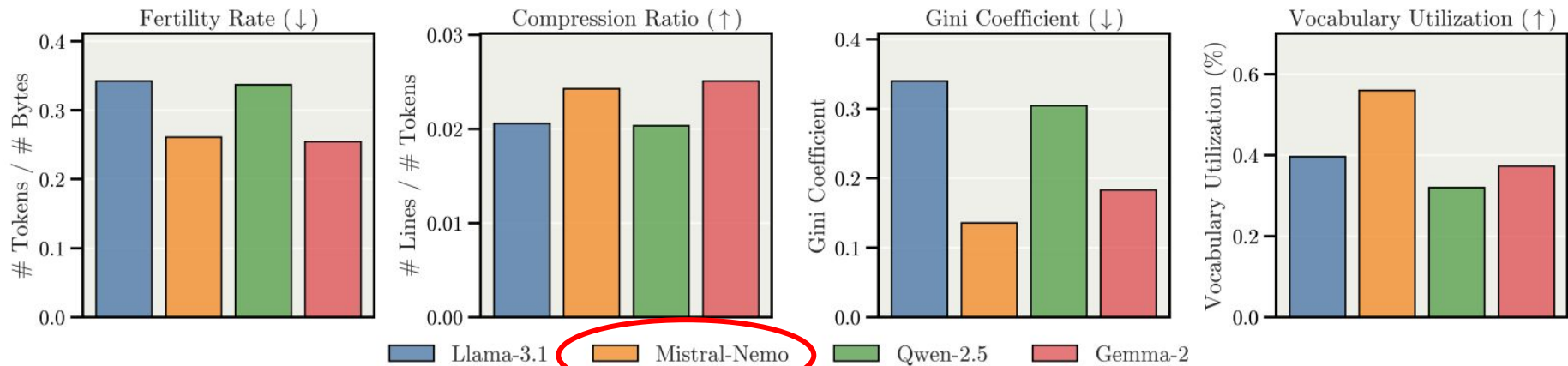


How to choose the best tokenizer for multilingual LLMs?

- Computational efficiency: *Fertility rate* and *compression ratio*
- Fairness of cost across languages: *Gini coefficient*
- Representativeness of vocabulary: *Vocabulary utilization*

Parity-Aware Byte-Pair Encoding: Improving Cross-lingual Fairness in Tokenization

Negar Foroutan^{1*} Clara Meister^{2*} Debjit Paul¹
 Joel Niklaus³ Sina Ahmadi⁴ Antoine Bosselut^{1†} Rico Sennrich^{4†}
¹EPFL ²ETH Zurich ³Niklaus.ai ⁴University of Zurich



Multilingual SFT data

Post-training: 149 languages

Mainly **SmolTalk2** (1.3M examples, 8 languages)

Newly created data:

- Wikipedia Q&A (884K)
- Romansh (46K)
- Swiss-German dialects (6K)
- African languages (7K)

Category	Dataset Source	# Examples	Data Ratio
Foundation	OLMo2 WildChat	298,556	9.56%
	OLMo2 Personas	29,356	
	OLMo2 SciRiff	29,809	
	OLMo2 TableGPT	24,803	
	OLMo2 CoCoNot	10,793	
	OLMo2 OASST1	7,047	
	<i>Subtotal</i>	<i>400,364</i>	
Math & Reasoning	Llama-Nemotron Math	200,000	11.60%
	Tulu3 Personas Math (filtered)	125,522	
	NuminaMath (tool-extracted)	63,248	
	OLMo2 OpenMath GSM8K	49,948	
	Llama-Nemotron Chat/Safety	46,808	
	<i>Subtotal</i>	<i>485,526</i>	
Code & Functions	Llama-Nemotron Code	200,000	9.02%
	Glaive Function Calling	112,688	
	XLam Function Calling	60,000	
	APIGen	5,000	
	<i>Subtotal</i>	<i>377,688</i>	
Multilingual	SmolTalk2 (8 languages)	1,273,789	34.22%
	EuroBlocks Multilingual	157,318	
	s1k.42_langs (filtered)	1,000	
	<i>Subtotal</i>	<i>1,432,107</i>	
Regional	WikiQA	883,513	22.54%
	Romansh	46,170	
	Swiss-German Dialects	6,179	
	African Languages	7,339	
	Swiss Charter Q&A	943'427	
	<i>Subtotal</i>	<i>943,427</i>	
Domain-Specific	The-Tome (Financial/Web)	544,975	13.02%
Total		4'184'087	100%

Multilingual Evaluation

Pre-training: **1811 languages**

Post-training: **149 languages**

Evaluation: cultural, knowledge, and instruction-following benchmarks covering **94 languages**

- knowledge ([Global-MMLU](#), TruthQA-Multilingual, HellaSwag-Multilingual)
- mathematical reasoning (MGSM)
- cultural understanding ([INCLUDE](#), [INCLUDE-v2](#), CulturalBench, BLEnD)
- Swiss-specific knowledge ([Switzerland QA](#))

Evaluations (we have many more tables)

Good English Performance

Model	Avg	General Knowledge					
		ARC (↑)	HellaSwag (↑)	WinoGrande (↑)	XNLI (↑)	XCOPA (↑)	PIQA (↑)
Fully Open Models							
Apertus-8B	65.8	72.7	59.8	70.6	45.2	66.5	79.8
Apertus-70B	67.5	70.6	64.0	73.3	45.3	69.8	81.9
OLMo2-7B	64.0	72.9	60.4	74.5	40.4	55.2	80.9
OLMo2-32B	67.7	76.2	66.7	78.6	42.9	60.1	82.1
EuroLLM-1.7B	54.8	57.2	44.9	58.1	40.7	55.7	72.4
EuroLLM-9B	62.8	67.9	57.9	68.8	41.5	61.1	79.6
SmolLM2-1.7B	58.5	66.1	52.4	65.6	37.6	52.3	77.0
SmolLM3-3B	61.6	68.6	56.4	68.1	40.5	58.2	77.7
Open-Weight Models							
Llama3.1-8B	65.4	71.6	60.0	73.4	45.3	61.8	80.1
Llama3.1-70B	67.3	74.4	56.5	79.4	44.3	66.7	82.3
Qwen2.5-7B	64.4	69.6	60.1	72.8	43.3	61.7	78.7
Qwen2.5-72B	69.8	76.2	67.5	78.0	46.9	68.2	82.0
Qwen3-32B	67.8	75.6	64.0	73.8	44.4	67.9	80.9
Llama4-Scout-16x17B	67.9	74.7	66.8	73.2	43.5	67.7	81.2
GPT-OSS-20B	58.1	67.0	41.5	66.5	37.4	60.4	75.6

Evaluations (we have many more tables)

Great Multilingual Capabilities

Model	Avg (↑)	Knowledge				Commonsense Reasoning	
		MMLU (↑)	Global- MMLU (↑)	TruthQA (↑)	TruthQA	HellaSwag (↑)	HellaSwag
					Multilingual (↑)		Multilingual (↑)
Fully Open Models							
Apertus-70B-Instruct	63.4	69.6	62.7	61.2	53.7	78.1	55.3
Apertus-8B-Instruct	58.8	60.9	55.7	56.7	52.4	74.6	52.7
ALLaM-7B-Instruct-preview	53.7	62.9	50.6	47.5	43.7	75.3	42.0
EuroLLM-22B-Instruct-Preview	58.3	65.3	56.9	56.6	49.8	73.0	48.1
EuroLLM-9B-Instruct	53.8	58.4	52.0	49.7	46.5	69.8	46.3
K2-Chat	56.8	65.7	49.8	56.5	49.2	74.9	44.7
marin-8b-instruct	54.5	65.5	48.4	55.2	47.6	72.0	38.1
Minerva-7B-instruct-v1.0	40.8	30.7	28.5	44.0	47.2	63.3	31.2
OLMo-2-0325-32B-Instruct	68.0	77.9	61.3	73.2	56.4	86.0	53.0
OLMo-2-1124-7B-Instruct	53.7	60.0	42.8	56.5	46.5	77.5	38.7
salamandra-7b-instruct	52.0	52.4	43.1	51.0	48.4	71.4	45.9
SmolLM3-3B	54.4	61.7	51.2	54.3	50.0	69.0	40.4
Teuken-7B-instruct-v0.6	48.9	49.0	39.9	46.4	48.1	67.8	42.2
Open-Weight Models							
gemma-3-12b-it	60.8	78.8	69.6	60.8	56.1	53.7	45.6
gemma-3-27b-it	63.8	83.6	75.3	64.4	54.8	54.9	49.8
Llama-3.1-8B-Instruct	59.2	72.4	57.1	55.1	50.8	72.5	47.0
Llama-3.3-70B-Instruct	68.4	87.5	77.8	66.1	55.2	70.1	53.8
Qwen2.5-72B-Instruct	68.8	86.6	77.7	69.7	58.6	68.8	51.5
Qwen3-32B	64.1	83.7	74.8	58.6	50.7	68.8	48.0
Qwen3-8B	57.8	79.1	64.0	53.4	51.4	58.6	40.4

Evaluations (we have many more tables)

Great Cultural Knowledge

Model	Cultural Knowledge					
	Avg (↑)	INCLUDE (↑)	INCLUDE V2 (↑)	BLEnD (↑)	Cultural Bench (↑)	Switzerland QA (↑)
Fully Open Models						
Apertus-70B-Instruct	61.5	58.2	41.6	66.3	74.2	67.2
Apertus-8B-Instruct	58.6	54.3	39.2	63.6	72.8	63.1
ALLaM-7B-Instruct-preview	55.2	44.4	34.6	66.4	74.4	56.0
EuroLLM-22B-Instruct-Preview	57.0	53.7	36.0	63.6	70.2	61.6
EuroLLM-9B-Instruct	54.3	49.3	36.8	62.7	61.4	61.2
K2-Chat	56.3	44.3	33.8	68.2	73.3	62.0
marin-8b-instruct	52.5	38.9	34.4	61.9	73.4	53.7
Minerva-7B-instruct-v1.0	39.1	25.6	28.0	40.4	64.0	37.4
OLMo-2-0325-32B-Instruct	58.1	52.9	39.5	61.2	74.5	62.2
OLMo-2-1124-7B-Instruct	49.7	36.3	31.3	60.8	72.8	47.2
salamandra-7b-instruct	52.8	42.1	33.0	58.6	70.5	59.6
SmolLM3-3B	52.7	41.4	31.1	61.6	72.6	56.6
Teuken-7B-instruct-v0.6	49.7	39.7	31.3	53.8	70.7	53.0
Open-Weight Models						
gemma-3-12b-it	63.4	62.7	42.8	69.5	76.8	65.1
gemma-3-27b-it	67.7	67.9	46.9	74.2	78.4	71.0
Llama-3.1-8B-Instruct	58.2	53.4	34.0	67.3	76.2	60.0
Llama-3.3-70B-Instruct	69.6	71.9	45.8	75.1	81.0	74.3
Qwen2.5-72B-Instruct	66.8	70.0	42.2	75.4	76.3	70.0
Qwen3-32B	65.9	70.6	45.8	72.0	75.5	65.6
Qwen3-8B	60.4	60.7	38.7	65.9	75.8	60.7

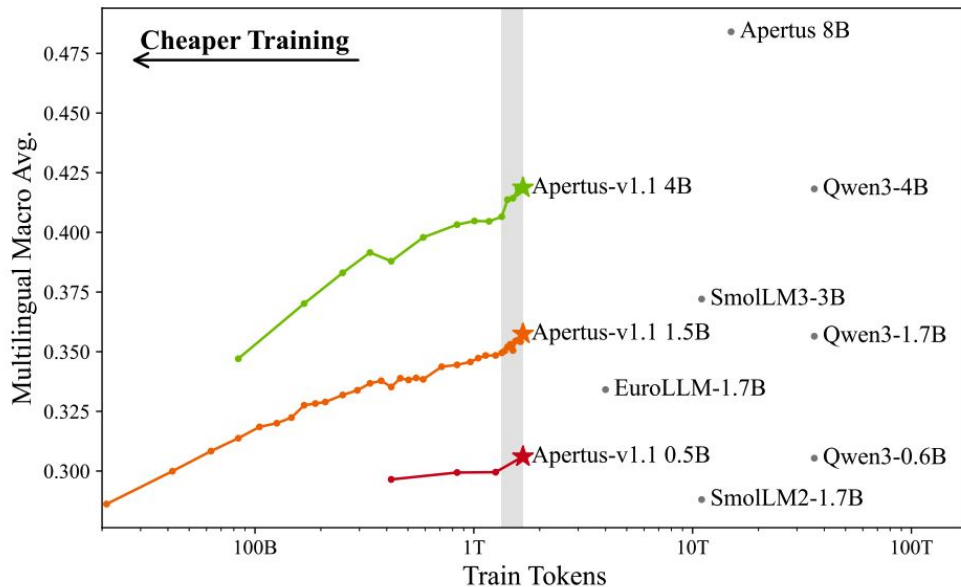
Next steps

Apertus 1.1 “Mini”

- **3 new model sizes: 0.5B, 1.5B, 4B**

16 small models to demo distillation and quantization techniques.

- Trained using **distillation on 1.7T tokens from Apertus-1.0-8B**
- Strong performance using **much less compute and memory** compared to similar-sized models

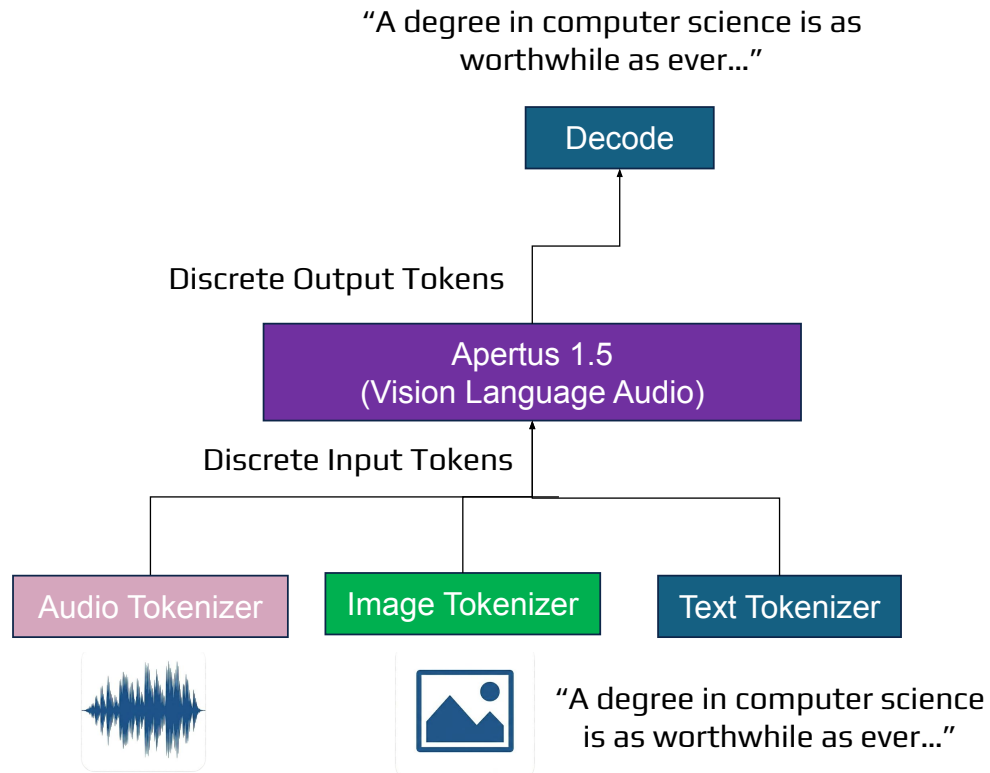


Apertus 1.5

Additional training on 4 trillion tokens of text and **multimodal** (image & audio) training data

Long Context: context window grows to 262,144 tokens, a four-fold increase over Apertus 1.0.

Optional thinking mode with **advanced reasoning capabilities**

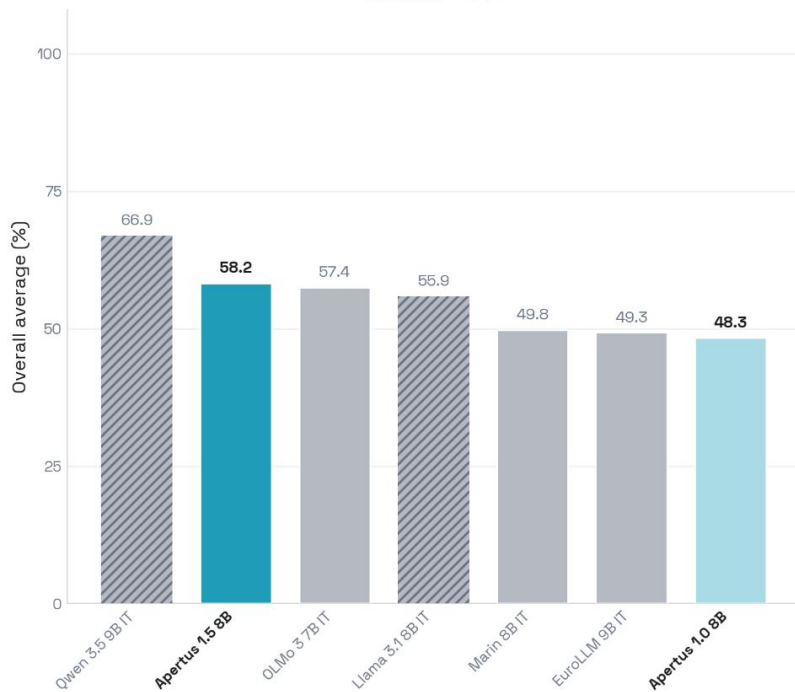


Apertus 1.5

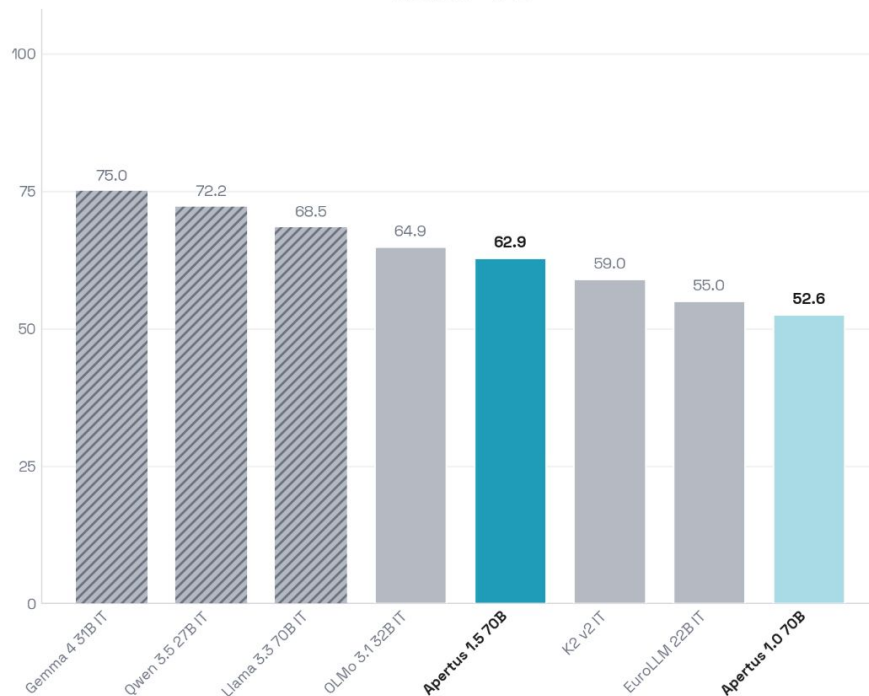
Text Performance

(on 44 benchmarks)

Models < 12B



Models > 12B



Apertus 2

Scaling up: What is the largest model we can train?

- ~10x larger
- stronger performance
- same deployment costs

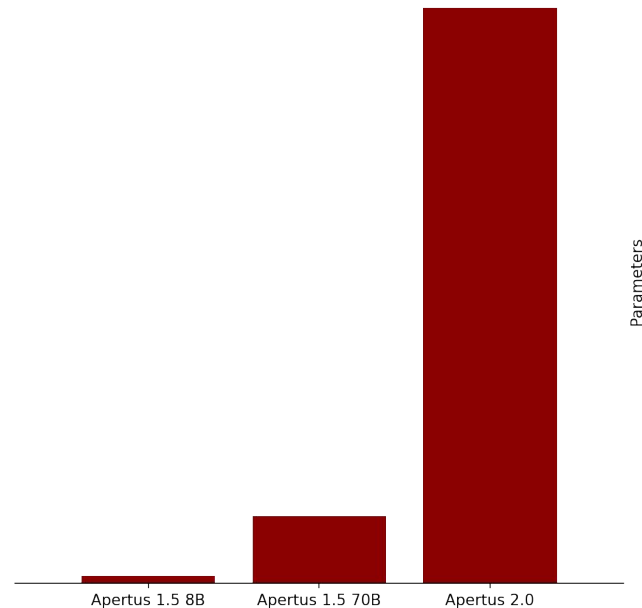
Native Multimodality

Leading fully open models: transparent, responsible, inclusive

Distillation

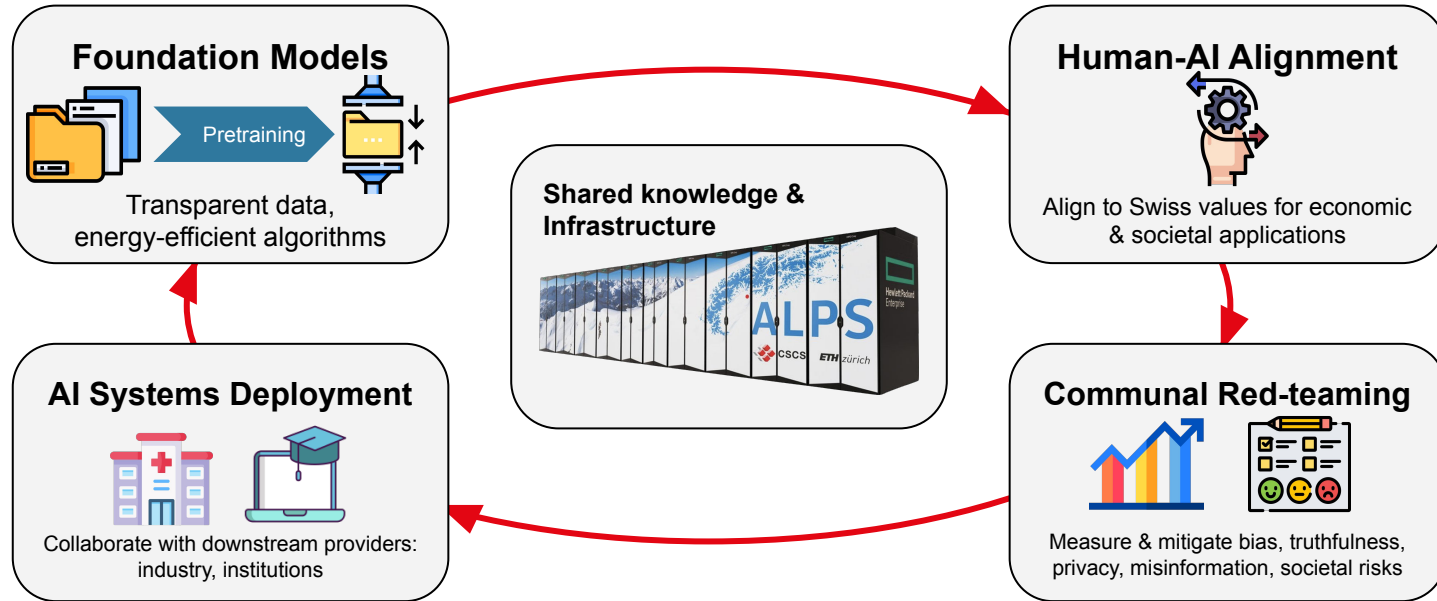
Small models trained from logits of large models

- 10x more efficient small model training



Conclusion

Post-Release Cycle



Fully Open, From Scratch, Multilingual

Model	Who built it	Languages	Scale	Money behind it
Apertus	EPFL / ETH Zurich / CSCS	1811 (claimed)	8B, 70B	Swiss federal + academic; CSCS Alps
BLOOM	BigScience, 1000+ researchers	46	176B	French public compute (GENCI, Jean Zay), 2022
EuroLLM	Unbabel / IST + partners	35	1.7B–22B	EU co-funded; EuroHPC (MareNostrum 5)
Salamandra / ALIA	Barcelona Supercomputing Centre	35	2B–40B	~€240M Spanish national programme
Pleias 1.0	Pleias (company)	8 European	350M–3B	Private + Jean Zay compute grant, Etalab
Lucie / Luciole	OpenLLM-France: LINAGORA, CNRS, X	FR + major EU	1B–23B	France 2030 "Communs Numériques" + own funds
Teuken-7B	OpenGPT-X / Fraunhofer IAIS	24 EU official	7B	German federal grant
Poro / Viking	TurkuNLP + Silo AI + HPLT	Nordic + EN + code	7B–34B	EU + LUMI; Silo AI AMD-owned since 2024

Four Patterns Across the Whole Landscape

Coordination has a cost

OpenEuroLLM, EuroLLM...

Grants end; models stop

After funding ends, what happens?

Hardware decides adoption

Apertus-8B pulled 624,507 in 30 days; the 70B, 34,633

The scale gap is real

Europe's largest open release is 675B (Mistral Large 3, and its best model is closed). Kimi K3 shipped at 2.8T.

Everyone Is Solving the Same Problems Privately

No shared evaluation

NorEval, IberoBench, Global-MMLU, INCLUDE, EU-wide suites — all separate, all built independently.

Failure modes stay unpublished

One engineering retrospective exists for a European run (CSCS)

Pipelines get rebuilt from zero

Crawl cleaning, language ID, deduplication, decontamination...

Contradictory findings sit unreconciled

Four Things We Could Do Together

A shared permissive SFT corpus

The MMLU-CoT collapse happens because permissively licensed reasoning traces barely exist.

Language ID in the tail

Every 1000+ language claim, ours included, rests on LID nobody has validated below roughly the top 200.

Attribution you can actually query

An European OlmoTrace?

Let's chat?

Currently working on / interested in:

Reasoning and planning for
multimodal LLMs

Faithfulness of closed-book
question answering in
uncertainty situations

Multilingual instruction-tuning
and reasoning-oriented RL

AI x Fundamental rights x Regulation, or how
to build models, frameworks & laws that
protect privacy, dignity, non-discrimination...

**Thanks for your
attention!**