A Cartography of Open Collaboration in Open Source AI:

Mapping Collaboration throughout the Development and Reuse Lifecycle of 14 Open Large Language Models

Johan Linåker, PhD & Cailean Osborne, PhD

OFA Symposium 2025

FGV Law Rio, Rio de Janeiro, Brazil

Team

Johan Linåker

RISE - Research Institutes of Sweden

Cailean Osborne

University of Oxford

Jennifer Ding

Boundary Object Studio

Ben Burtenshaw

Hugging Face

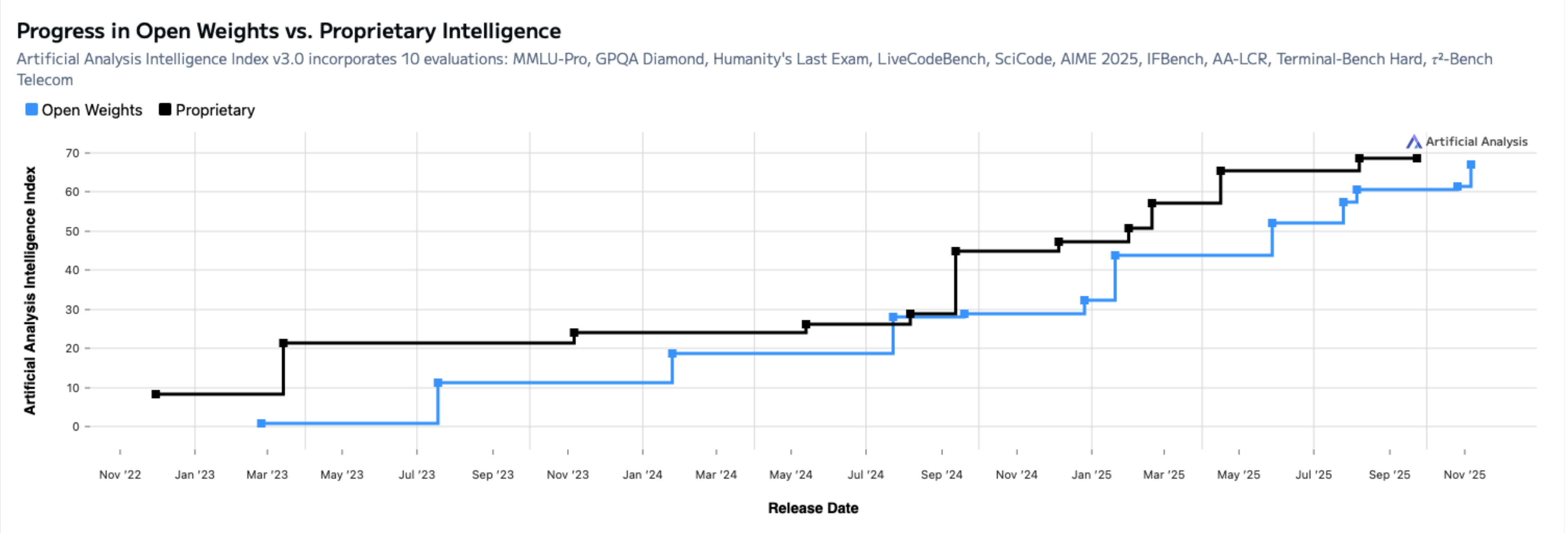
The explosion of open source AI in numbers

~5K models added to Hugging Face Hub every day

2.2M public model repositories on Hugging Face Hub

66 models have >=100M downloads

Open source is eating the world



Progress in Open Weights vs. Proprietary Intelligence

Artificial Analysis Intelligence Index v3.0 incorporates 10 evaluations: MMLU-Pro, GPQA Diamond, Humanity's Last Exam, LiveCodeBench, SciCode, AIME 2025, IFBench, AA-LCR, Terminal-Bench Hard, τ^2-Bench Telecom

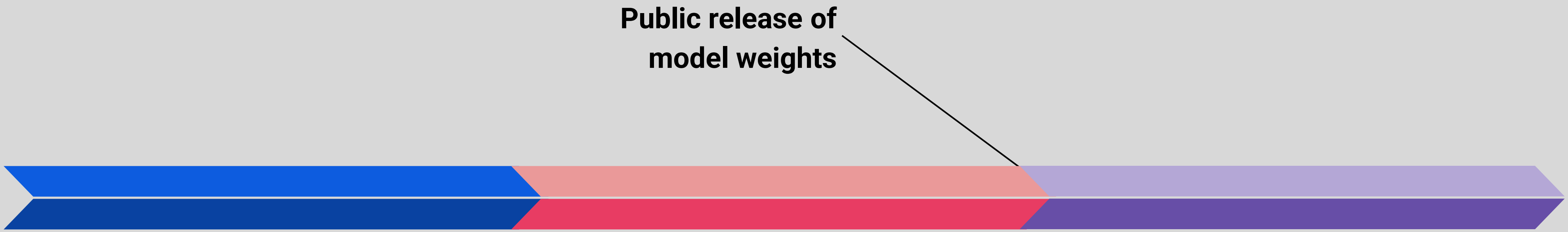
Source: https://artificialanalysis.ai/trends#progress-in-open-weights-vs-proprietary-intelligence

What is open source in open source AI?

- Open-washing (licenses)
- OSS and AI involve different development processes & resources
- Limited openness before the release of weights
- Open source AI = models or more (e.g. science, software, data)?

- **The license perspective on open source AI:** Open Source AI Definition v1 requires that open source licenses are used for AI systems and their constituent parts (e.g. model weights, code, and training data or sufficiently detailed information of it)

Openness throughout the AI model lifecycle



Public release of model weights

Pre-training

Building a model from scratch, including:

- Model design
- Data collection & curation
- Training code
- Compute access

Post-training

Optimizing model capabilities and behaviours, including:

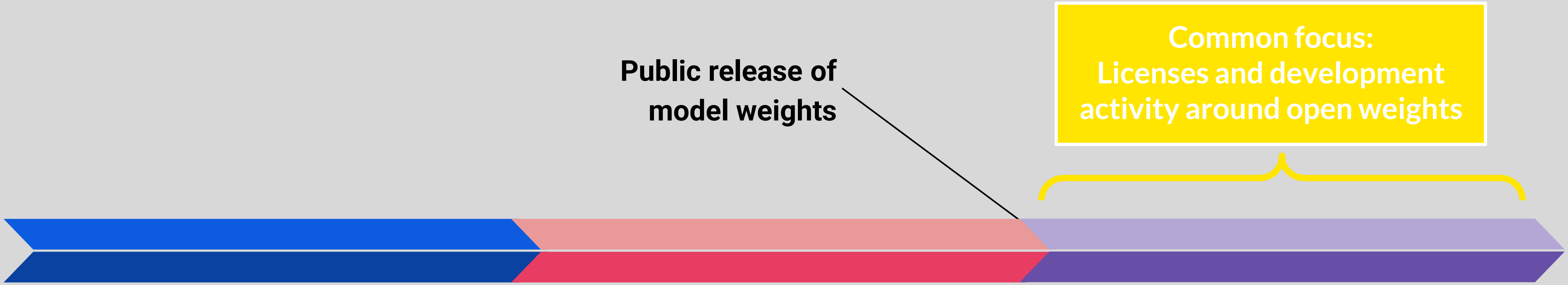
- Domain-specific fine-tuning
- Alignment fine-tuning
- Red-teaming & safety testing
- Data for all of the above

Post-release

Downstream reuse, adaptation, and feedback once released, including:

- Derivative development
- Feedback and bugs
- Upstream contributions
- Research

Openness throughout the AI model lifecycle



Public release of model weights

Common focus: Licenses and development activity around open weights

Pre-training

Building a model from scratch, including:

- Model design
- Data collection & curation
- Training code
- Compute access

Post-training

Optimizing model capabilities and behaviours, including:

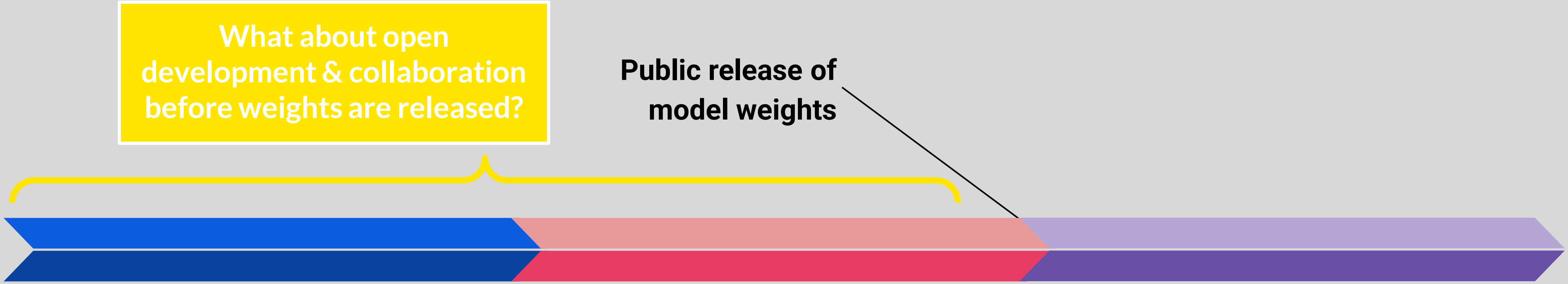
- Domain-specific fine-tuning
- Alignment fine-tuning
- Red-teaming & safety testing
- Data for all of the above

Post-release

Downstream reuse, adaptation, and feedback once released, including:

- Derivative development
- Feedback and bugs
- Upstream contributions
- Research

Openness throughout the AI model lifecycle



What about open development & collaboration before weights are released?

Public release of model weights

Pre-training

Building a model from scratch, including:

- Model design
- Data collection & curation
- Training code
- Compute access

Post-training

Optimizing model capabilities and behaviours, including:

- Domain-specific fine-tuning
- Alignment fine-tuning
- Red-teaming & safety testing
- Data for all of the above

Post-release

Downstream reuse, adaptation, and feedback once released, including:

- Derivative development
- Feedback and bugs
- Upstream contributions
- Research

Research design

Research design

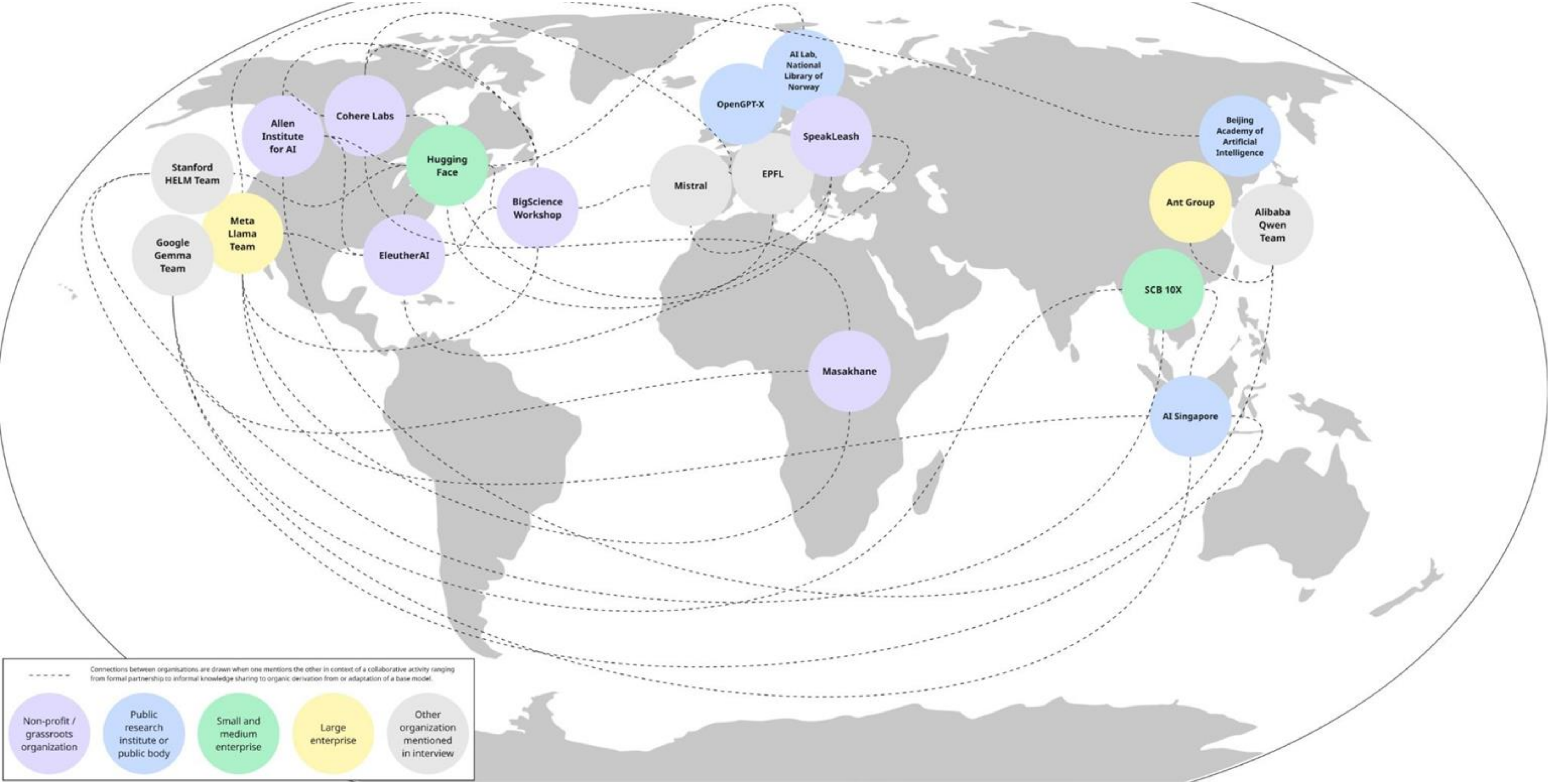
- **Aim:** Explore open collaboration throughout the LLM development and reuse lifecycle

- **Methods & data:** Semi-structured interviews with 17 developers of 14 open LLMs

- **Sample:**
  - Licenses: Permissive and restrictive
  - Regions: North America, Europe, Africa, Asia
  - Projects: Grassroots initiatives, non-profits, research labs, public bodies, startups, big tech

- **Limitations:**
  - Focus on LLMs
  - Missing regions: Latin America!

Where and how does open collaboration take place throughout the development and reuse lifecycle of open LLMs?

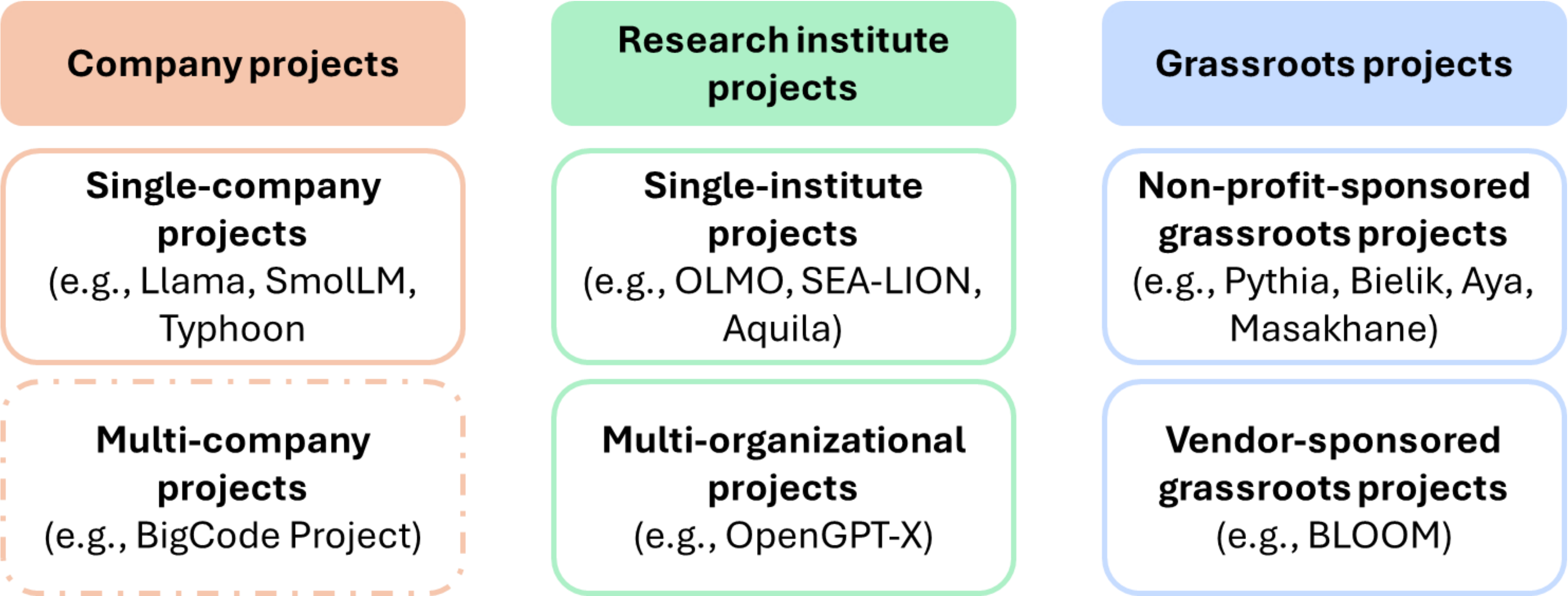


Connections between organisations are drawn when one mentions the other in context of a collaborative activity ranging from formal partnership to informal knowledge sharing to organic derivation from or adaptation of a base model.

Key findings

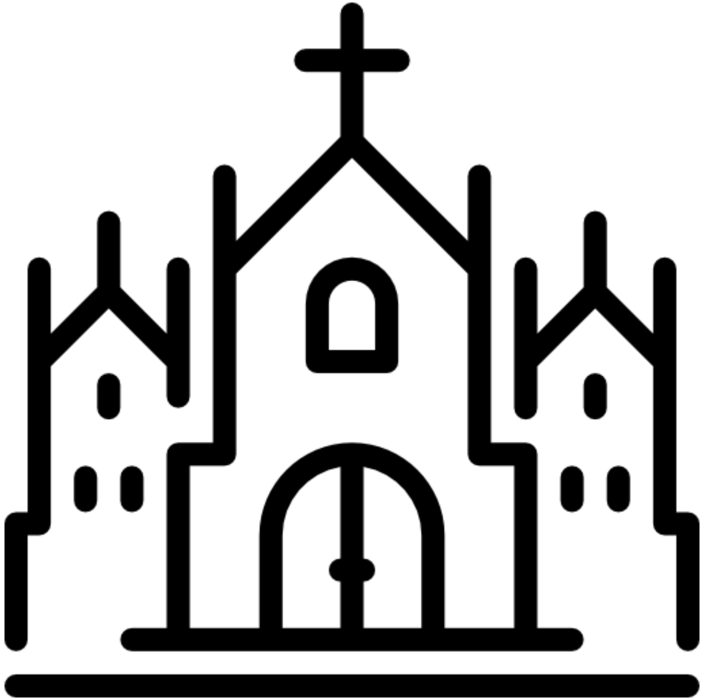
Organisational models in open LLM projects

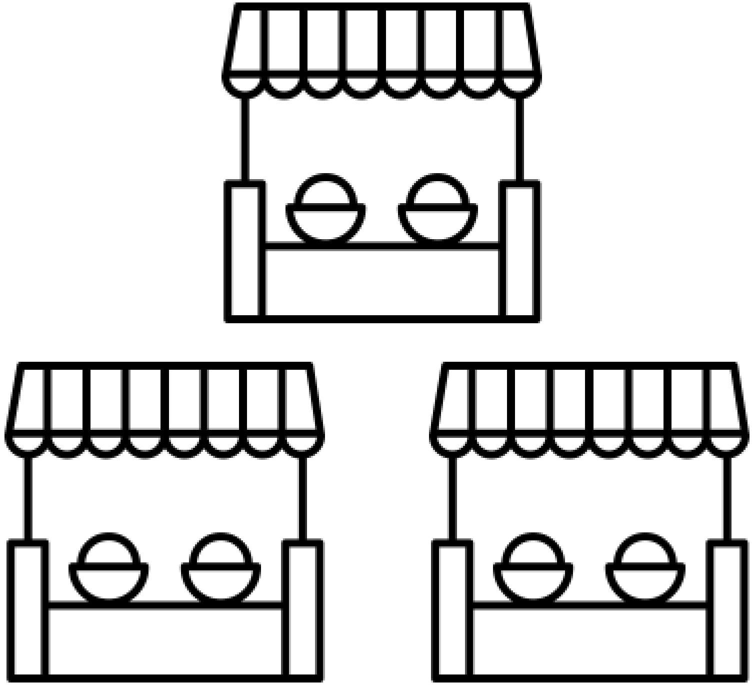
Formal governance observed across cases



Company projects	Research institute projects	Grassroots projects
Single-company projects (e.g., Llama, SmolLM, Typhoon	**Single-institute projects** (e.g., OLMO, SEA-LION, Aquila)	**Non-profit-sponsored grassroots projects** (e.g., Pythia, Bielik, Aya, Masakhane)
Multi-company projects (e.g., BigCode Project)	**Multi-organizational projects** (e.g., OpenGPT-X)	**Vendor-sponsored grassroots projects** (e.g., BLOOM)

Cathedrals and bazaars throughout the open LLM lifecycle





Cathedrals and bazaars across the lifecycle

1 Pre-training

Limited collaboration: Small teams of experts, compute, and coordination challenges.

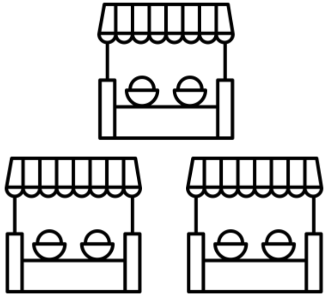
However, open collaboration on resources used to train LLMs, including open training data, open source frameworks, and compute.

2 Post-training

Limited collaboration: Small teams of experts, secrecy, and coordination challenges.

However, some collaboration takes place, including sharing weights with trusted partners and releasing post-training datasets.

3 Post-release



Significant collaboration activity, including derivative development, feedback, research, demos, etc.

However, challenges remain, including gated licenses, firewalls, and sustaining collaboration given speed of LLM releases.

Open collaboration challenges and on-ramps in open LLM projects

Pre-training stage

Pre-training: Collaboration challenges

Companies tend to keep their training techniques secret

Informal chats are possible, but *"pre-training and post-training treated as very closely guarded secrets"* – AI2

Resource requirements lock developers in early

"Collaborations need to be scoped out before we start working because they will majorly alter the experimentation stage" – AI2

Technical expertise is a barrier

"It requires very specialized knowledge... It would be very unoptimized to have 300 people involved" – SpeakLeash Foundation

Data work is undervalued

"Everybody wants to train models. Nobody wants to do the data work" – National Library of Norway

Curated data is expensive and source of competitive advantage

Companies are reluctant to share training data because *"they invested so much money in it" – National Library of Norway*

Hesitation around data sharing due to copyright concerns

"Increased attention from rights holders to clarify what may be used for training and what may be released" – National Library of Norway

RI.
SE

Pre-training: Collaboration on-ramps

Low & medium-resource language datasets

"Building datasets is one of the easiest things for us to collaborate on. Thai is a medium to low resource language, there aren't a lot of high quality datasets, and the research community understands this..." – Typhoon, SCB 10X

"Many African datasets have been collected and Africans do not have access to them and have to pay... Eventually, we decided to address this by creating datasets" - Masakhane

Sharing open training datasets

"[DeepSeek] had a really good math model... but they did not release [the math dataset]. So, we rebuilt datasets similar to what they have, except that it is open and now everyone can train on it and get really good math performance" - Hugging Face

Compute partnerships

"The origin of the BigScience Workshop was an invitation from an administrator of Jean Zay, a French public cluster, for Hugging Face to stress-test it" - Hugging Face

Contributing to open source frameworks

Users of EleutherAI's GPT-NeoX have contributed code to the main branch, including mixture of experts architecture and RLHF finetuning support, after developing for their own needs.

Improving quality and legality of open training datasets

EleutherAI reverse-engineered licenses in CommonCrawl data repository, given the problem of "license laundering".

Post-training stage

Post-training: Challenges and on-ramps

Challenges

Competitive secrets

Companies tend to guard post-training techniques, as performance optimization differentiates models.

Rapid iteration cycles

Fast experimentation cycles make it difficult to coordinate with external partners.

On-ramps

Sharing weights with trusted partners for testing

"We shared intermediate checkpoints of the model with other startups, so they could test them and see if we could add new capabilities during post-training" – Hugging Face

Releasing open post-training datasets

"We released a post-training dataset called Infinity Instruct. After three months, it had been used by external developers to post-train over 130 models" – BAAI

Post-release stage

Post-release: Collaboration on-ramps

Distribution and discovery

"We have been collaborating with Hugging Face on model dissemination... The Hugging Face platform is the de facto way to share your models and to make things accessible to people" – EleutherAI

Derivative development

Lugha Llama was a *"collaborative effort to finetune Llama for 20 African languages... It outperformed Llama on Global MMLU"* – Masakhane

Catalysing follow-up projects

"The BigCode project came out of BigScience and was able to learn from what did not work in BigScience, including having more direction from the start and less experimental working groups" – Hugging Face

Community feedback

SpeakLeash Foundation launched a toxicity assessment survey for their Bielik model and already received over 60,000 responses.

Benchmarking

Masakhane maintains AfroBench, a public leaderboard on Hugging Face that tracks the performance of multilingual models across 64 African languages, 15 NLP tasks, and 22 datasets. Developers can submit evaluation results for models to it.

Regional adaptation and use

GoTo (super app in Indonesia) now uses AI Singapore's SEA-LION LLM to power customer service in Javanese, Sundanese, and Indonesian.

Post-release: Collaboration challenges

Gated licenses exclude regions

"I cannot apply for the Llama license because my identity says that I'm in mainland China" – Ant Group

Making code reusable is intensive and not always a priority

"Absolutely everything is open, but that is more like a principle, right? I think there are 2, 3, maybe 4 people that have actually taken that code and used it" – National Library of Norway

Network firewalls

"We work internally because there is a firewall. We developed an internal model scope in our internal GitLab, and regularly we upload our updates to GitHub and Hugging Face" – Ant Group

Sustaining collaboration is hard given rapid speed of new releases

"Many AI open source projects are very short living. They do not live long. They just appear for one or two months and they will disappear forever…" – Ant Group

RI.
SE

Summary and food for thought

Summary and food for thought

- Open collaboration in open LLM projects takes place throughout the LLM lifecycle (pre-training, post-training, post-release) and extends far beyond LLMs themselves (e.g. open data, open source software, open benchmarks, open science, compute partnerships).

⇒ How, if at all, does this change your perspective on openness in AI?

⇒ What lessons can we learn from these national, regional, and global projects for designing future projects?

⇒ What policy levers could facilitate the development, maintenance, and sustainability of open technology communities and ecosystems powering public interest AI research and innovation?

Thanks for listening!