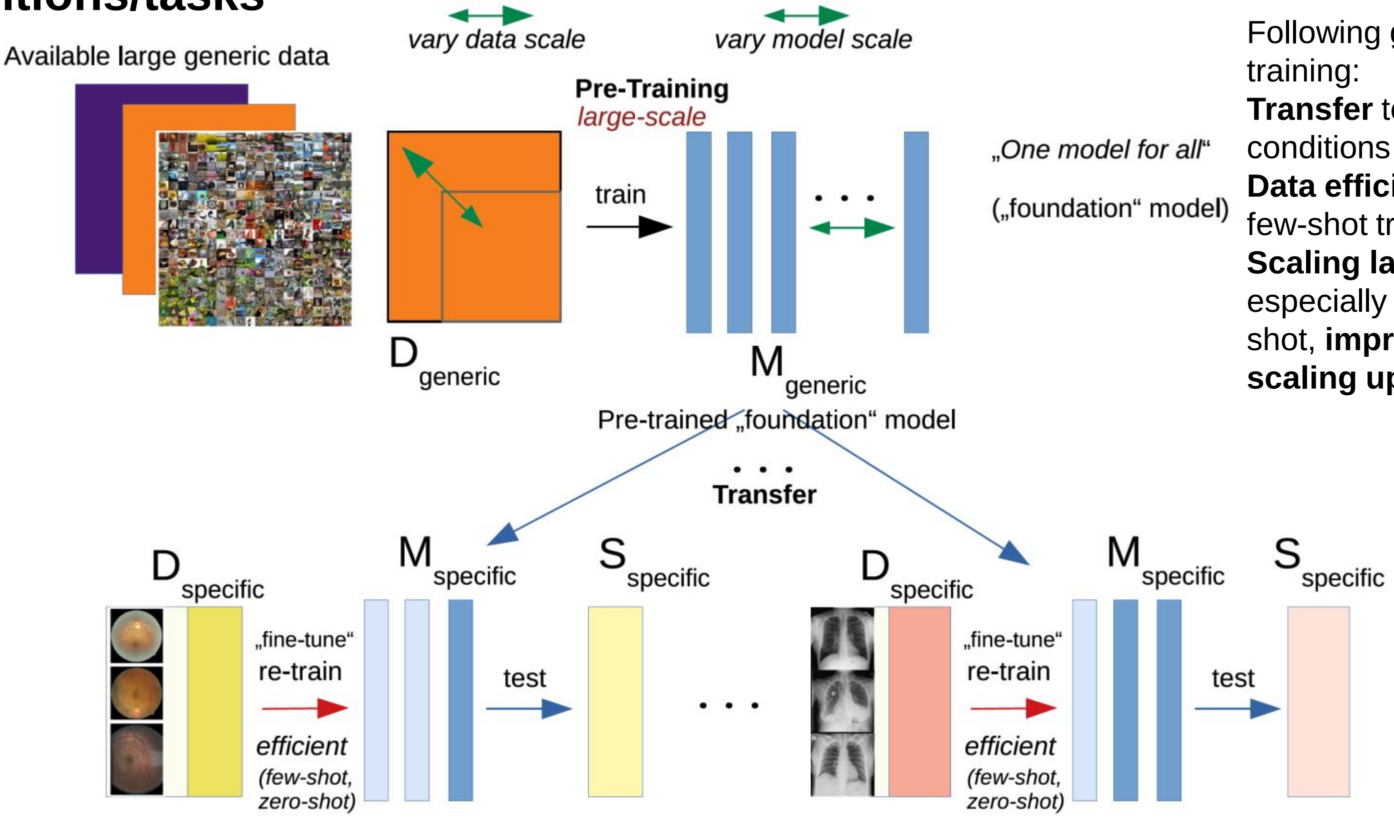
# Open foundation models: scaling laws & generalization

Jülich Supercomputing Center (JSC)
Scalable Learning & Multi-Purpose AI Lab (SLAMPAI)
Large-scale Artificial Intelligence Open Network (LAION)
European Laboratory for Learning and Intelligent Systems (ELLIS)

10. July 2025 | Jenia Jitsev

# Foundation models: generic transferable learning

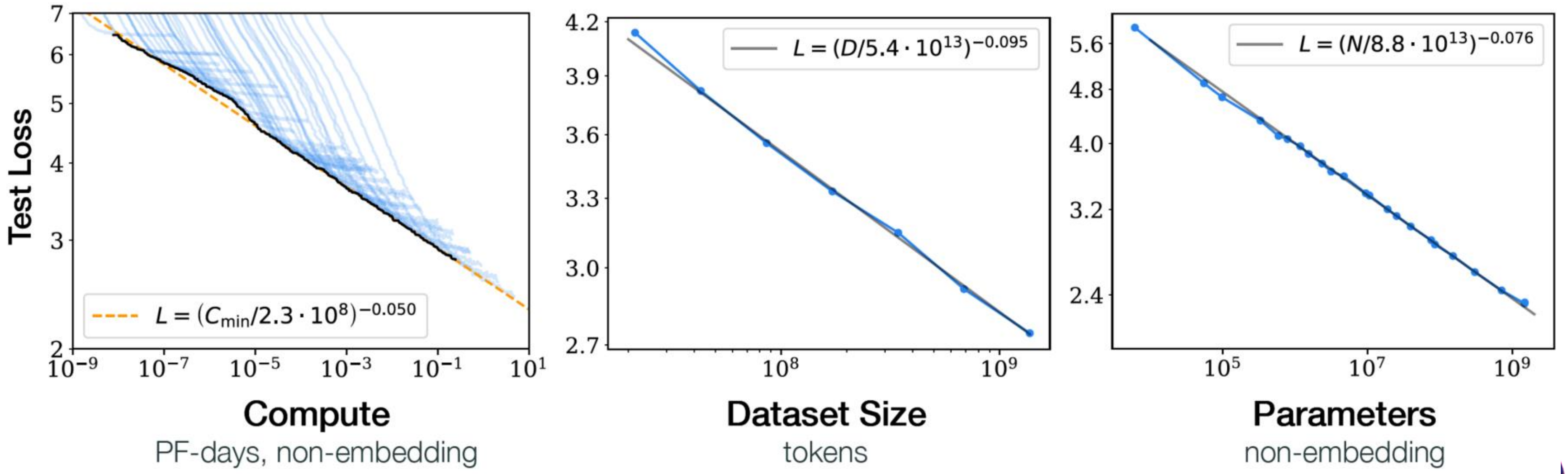
- Core breakthroughs (since ca. 2012): **learning that transfers across condintions/tasks**




Following generic pre-training:
**Transfer** to various conditions and tasks
**Data efficient**: zero- or few-shot transfer
**Scaling laws:** transfer, especially zero and few-shot, **improves when scaling up**

# Foundation models: scaling laws

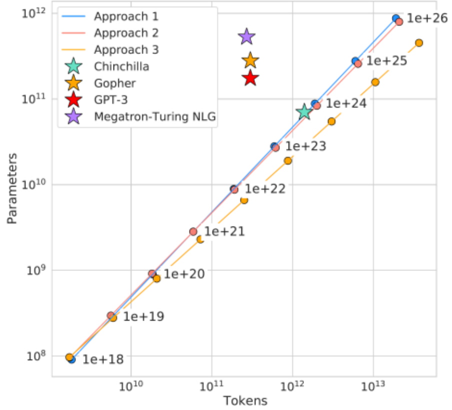
- **Scaling Laws**: larger model, data and compute scale during pre-training – **stronger generalization & transferability**
- **No change** in core algorithmic procedure required! Scaling up alone improves important core functions




Kaplan et al, 2020

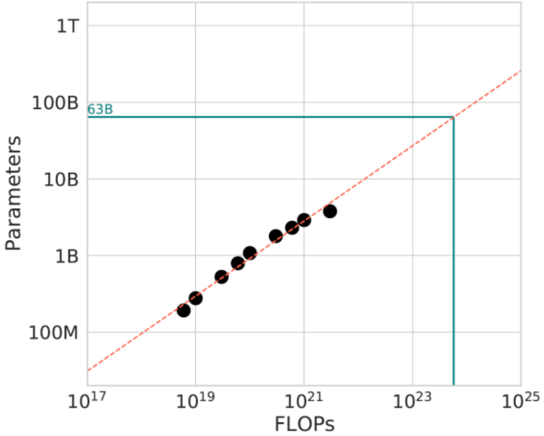
# Foundation models: scaling laws

- **Scaling Laws**: predicting model properties and function across scales

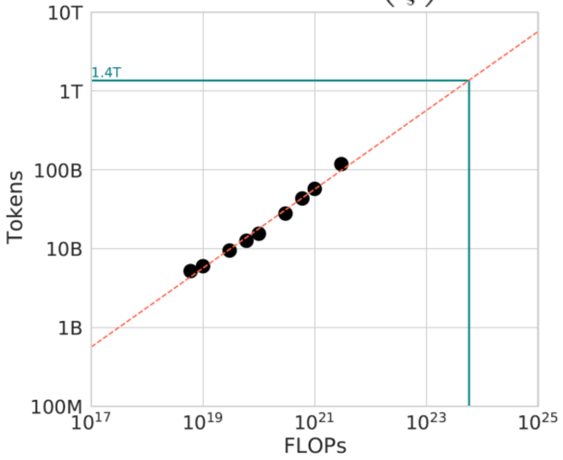



| Parameters | FLOPs | FLOPs (in *Gopher* unit) | Tokens |
|---|---|---|---|
| 400 Million | 1.92e+19 | 1/29,968 | 8.0 Billion |
| 1 Billion | 1.21e+20 | 1/4,761 | 20.2 Billion |
| 10 Billion | 1.23e+22 | 1/46 | 205.1 Billion |
| 67 Billion | 5.76e+23 | 1 | 1.5 Trillion |
| 175 Billion | 3.85e+24 | 6.7 | 3.7 Trillion |
| 280 Billion | 9.90e+24 | 17.2 | 5.9 Trillion |
| 520 Billion | 3.43e+25 | 59.5 | 11.0 Trillion |
| 1 Trillion | 1.27e+26 | 221.3 | 21.2 Trillion |
| 10 Trillion | 1.30e+28 | 22515.9 | 216.2 Trillion |

$$N^*(C) = G\left(\frac{C}{\xi}\right)^a$$



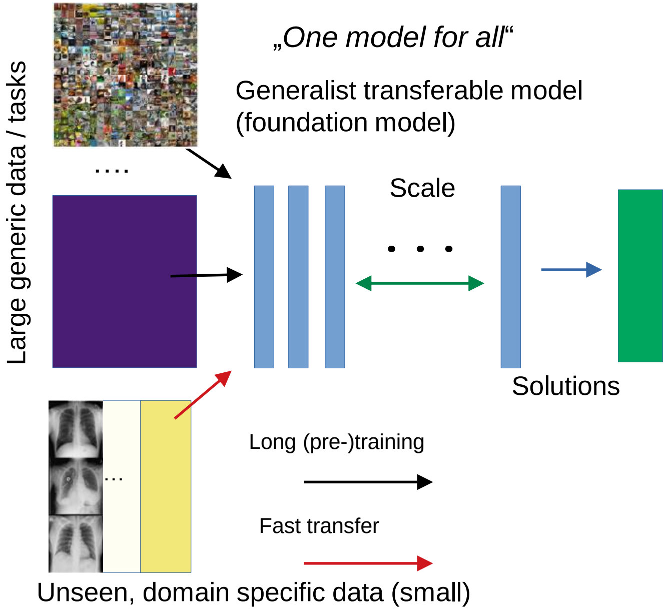

$$D^*(C) = G^{-1}\left(\frac{C}{\xi}\right)^b$$




Hoffmann et al, arXiv:2203.15556, 2022

# Foundation models: reproducibility & progress

- **Problem**: research on foundation models, datasets & scaling laws reproducible only by few large industry labs (Google; openAI; Microsoft; Meta; NVIDIA; …)
- **Important large foundation models:** GPT-3/4, PaLM, DALL-E 2/3, Flamingo, CLIP **- closed to public research**
- **Datasets** used to train those models**: REQUIRED! closed**
- Majority of strong foundation models**: Non-reproducible (by independent parties), intransparent artefacts**



# Research communities for open foundation models

- Rise of **grassroot research communities** to open-source and study foundation models & datasets required for their training
- **EleutherAI** (USA, 2020): language – Pile, Pythia, LM-Eval-Harness
- **BigScience** (EU, France, 2021): language, code, language-vision - BLOOM, StarCoder, Idefix, smolLM (mostly driven by HuggingFace)
- **LAION** (EU, Germany, 2021; **important hub** at **JSC**): multi-modal language-vision, language-audio – LAION-400M/5B, openCLIP, DataComp, Open Assistant, CLAP, openFlamingo, DCLM, CLIP-Benchmarks
- **Open large datasets and foundation models: reproducibility !**
  - **joint efforts accross institutions/organisations boundaries**

# Open-source foundation models & datasets

- Making **whole pipeline** – dataset composition, model training, benchmarks & evaluation – **fully reproducible**

OPEN-SOURCE
Dataset &
Dataset composition

OPEN-SOURCE
Training procedure,
model weights,
checkpoints

OPEN-SOURCE
Evaluation benchmarks,
downstream transfer procedures

**Supercomputers and experts handling them required!**

Re-LAION-5B,
DataComp-1B,
DCLM-baselines
OpenThoughts

https://github.com/mlfoundations/datacomp/

OpenCLIP,
openFlamingo,
DCLM
OpenThinker

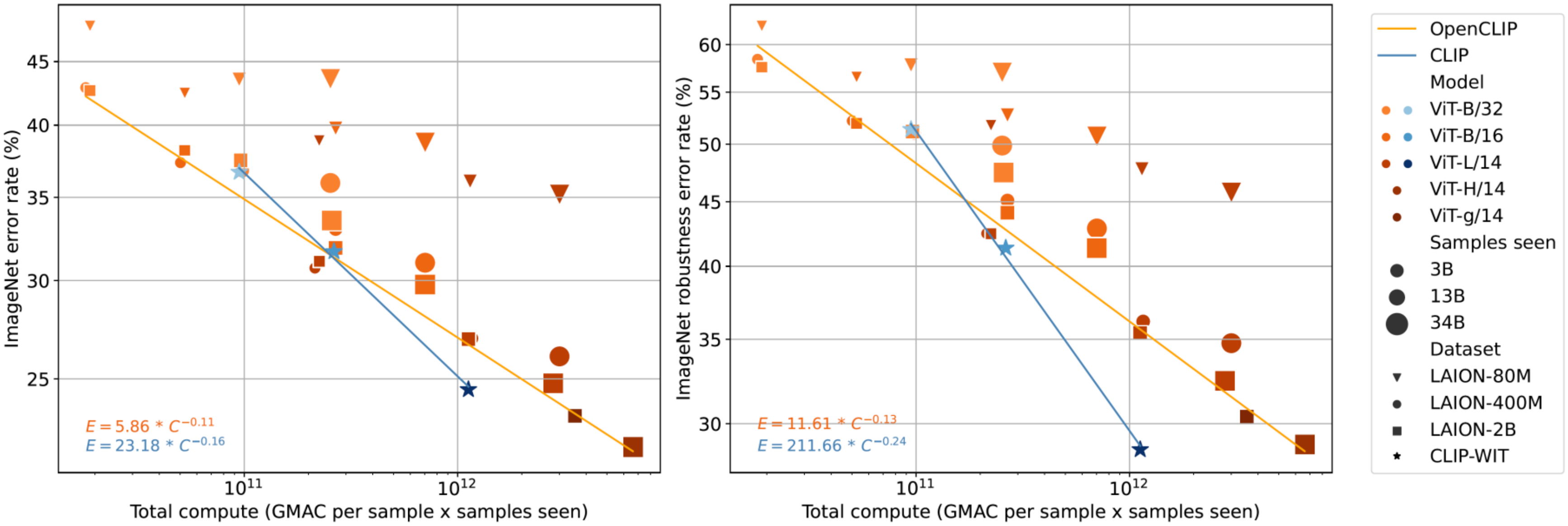
https://github.com/mlfoundations/open_clip

openCLIP Benchmarks,
EvalChemy,
AIW problems: generalization,
reasoning evals

https://github.com/LAION-AI/CLIP_benchmark/

# Reproducible scaling laws for foundation models

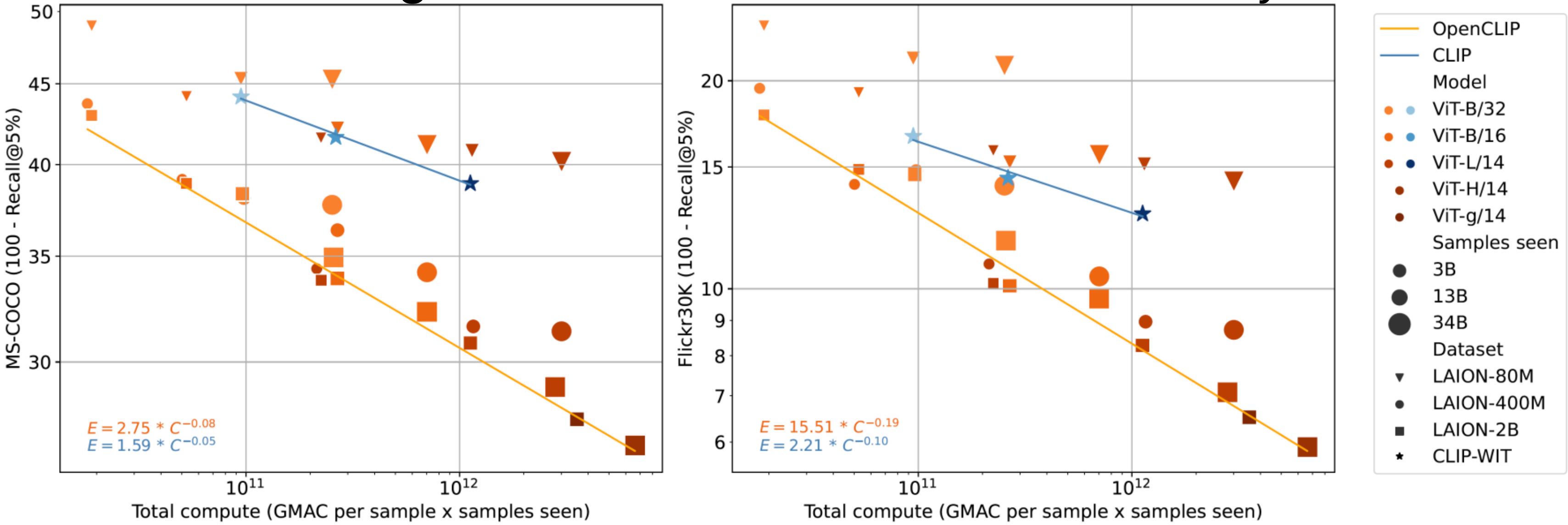
- Scaling laws with LAION-400M/2B and openCLIP: open-source data, models and code - **reproducible** science of foundation models
- Below: zero-shot image classification, ImageNet-1k & robustness sets




Schuhmann et al, NeurIPS, 2022; Cherti et al, CVPR 2023

# Scaling laws for open foundation models

- Comparing LAION-400M/2B (LAION) and WIT (openAI)
- **Matching or outperforming strong closed models** by using open data
  - LAION as a **open frontier lab**: building **open** foundation models that match **strongest** state-of-the-art from closed industry labs

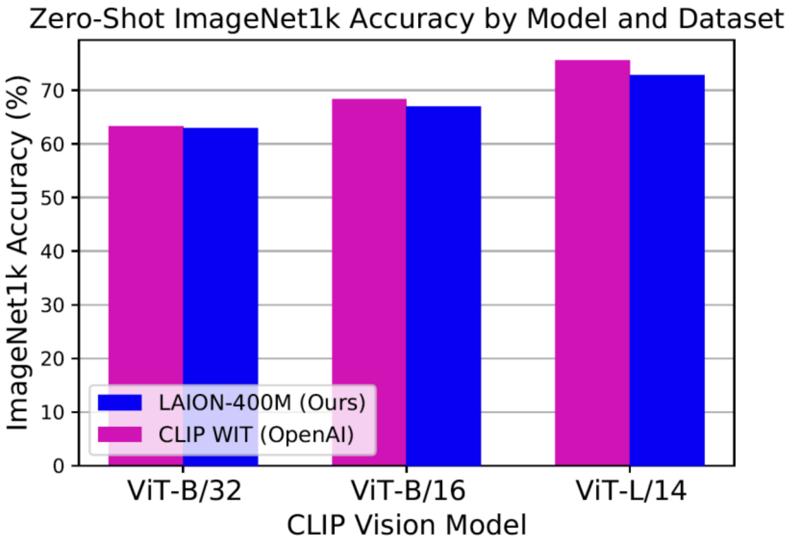



Schuhmann et al, NeurIPS, 2022; Cherti et al, CVPR 2023

https://github.com/mlfoundations/open_clip

# Open foundation models & datasets

- **Predictably outperforming strong closed models** by using open data
- LAION as an **open frontier lab**: building **open** foundation models that match **strongest** state-of-the-art from closed industry labs

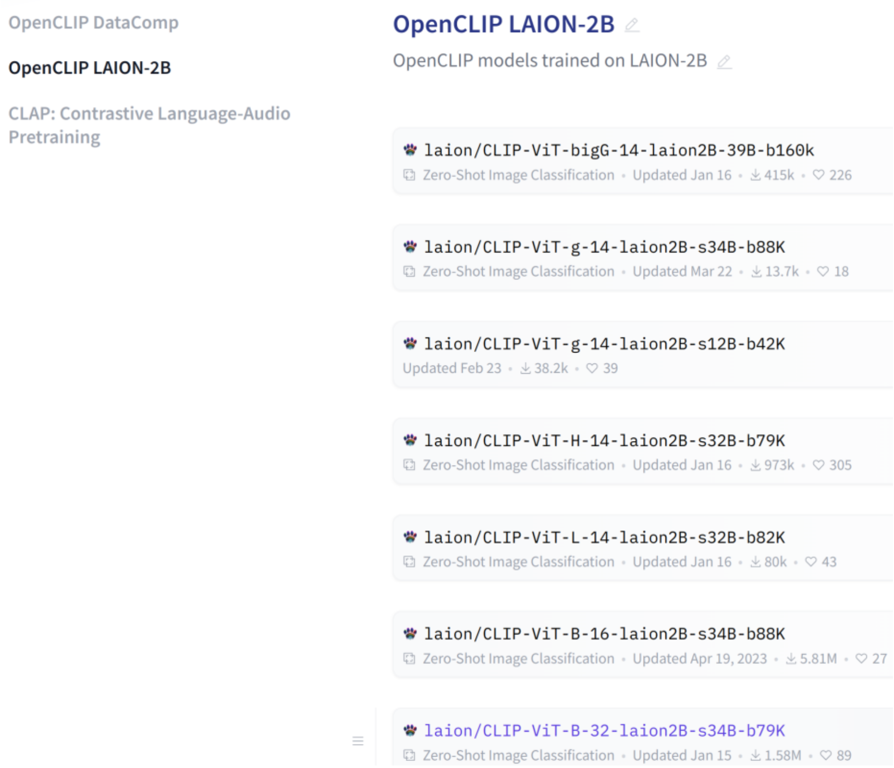


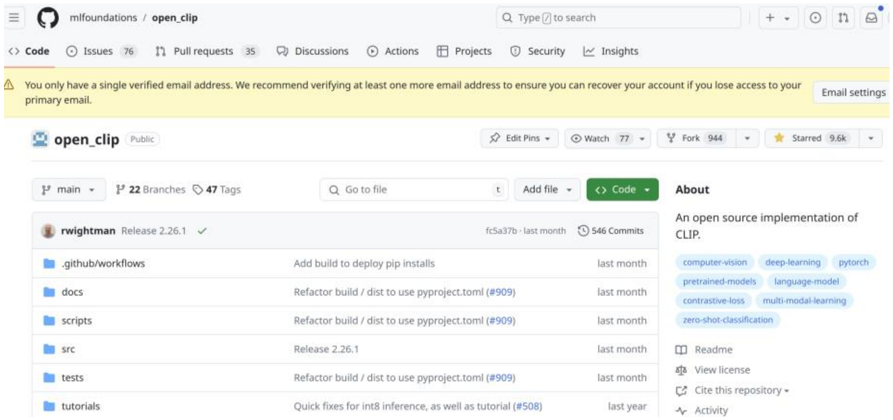

| Dataset | # English Img-Txt Pairs |
|---|---|
| **Public Datasets** | |
| MS-COCO | 330K |
| CC3M | 3M |
| Visual Genome | 5.4M |
| WIT | 5.5M |
| CC12M | 12M |
| RedCaps | 12M |
| YFCC100M | 100M[2] |
| **LAION-5B (Ours)** | **2.3B** |
| **Private Datasets** | |
| CLIP WIT (OpenAI) | 400M |
| ALIGN | 1.8B |
| BASIC | 6.6B |

Schuhmann et al, NeurIPS, 2022; Cherti et al, CVPR 2023

# Open foundation models & datasets

- Open-source releases: > 100M of downloads for pre-trained openCLIP models; >10k stars for code repository



# Open foundation models & datasets

- DataComp-LM: fully open, reproducible pipeline for language modelling; fully open data (DCLM-Baseline, 4.4T tokens in total) & models (DCLM-1B/7B); predictably match/outperform SOTA models (eg Llama-3-8B)

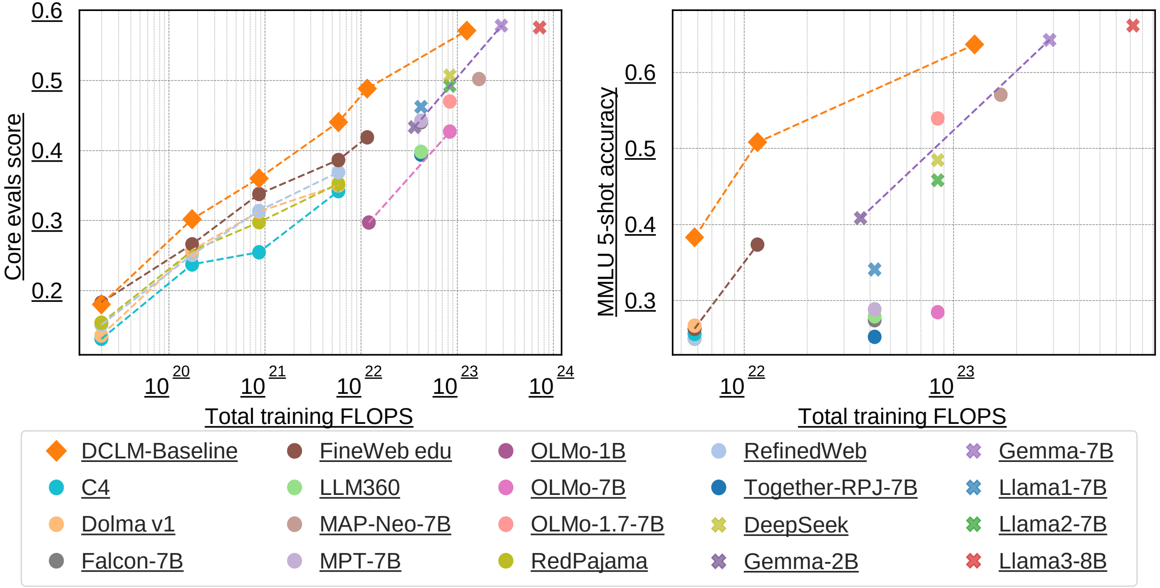


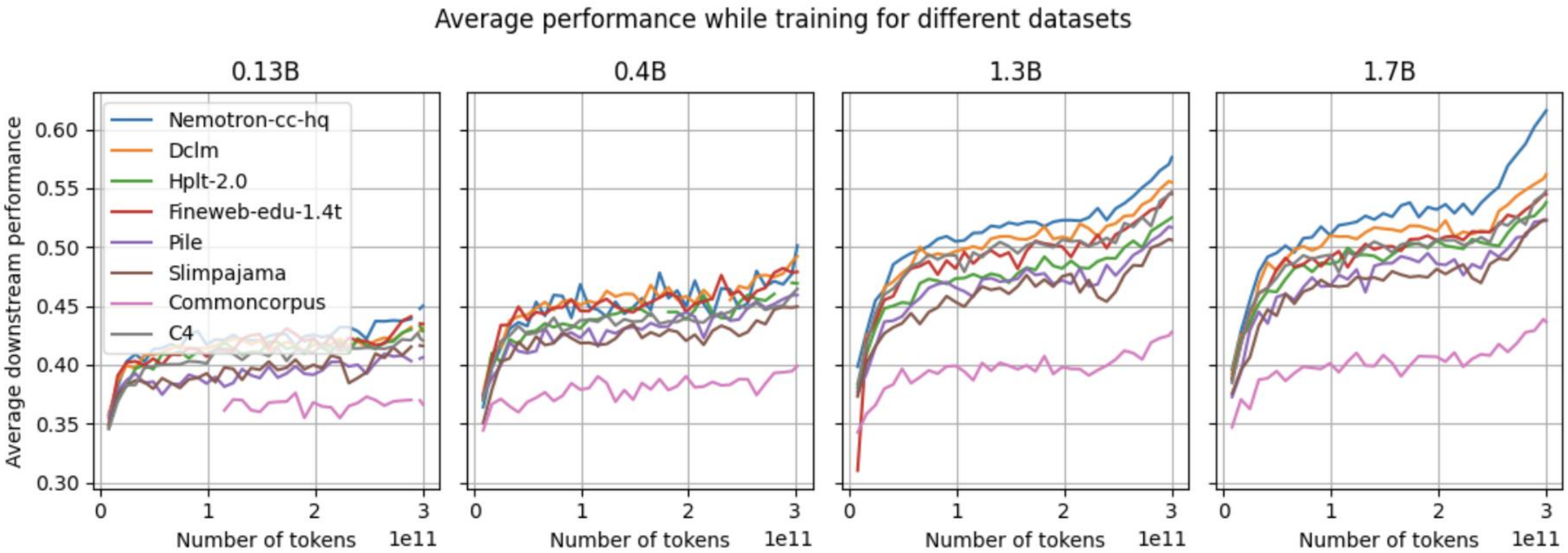

Figure 1: **Improving training sets leads to better models that are cheaper to train.**

---

Li et al, ArXiv:2406.11794, NeurIPS, 2024 https://github.com/mlfoundations/dclm

# Open foundation models & datasets

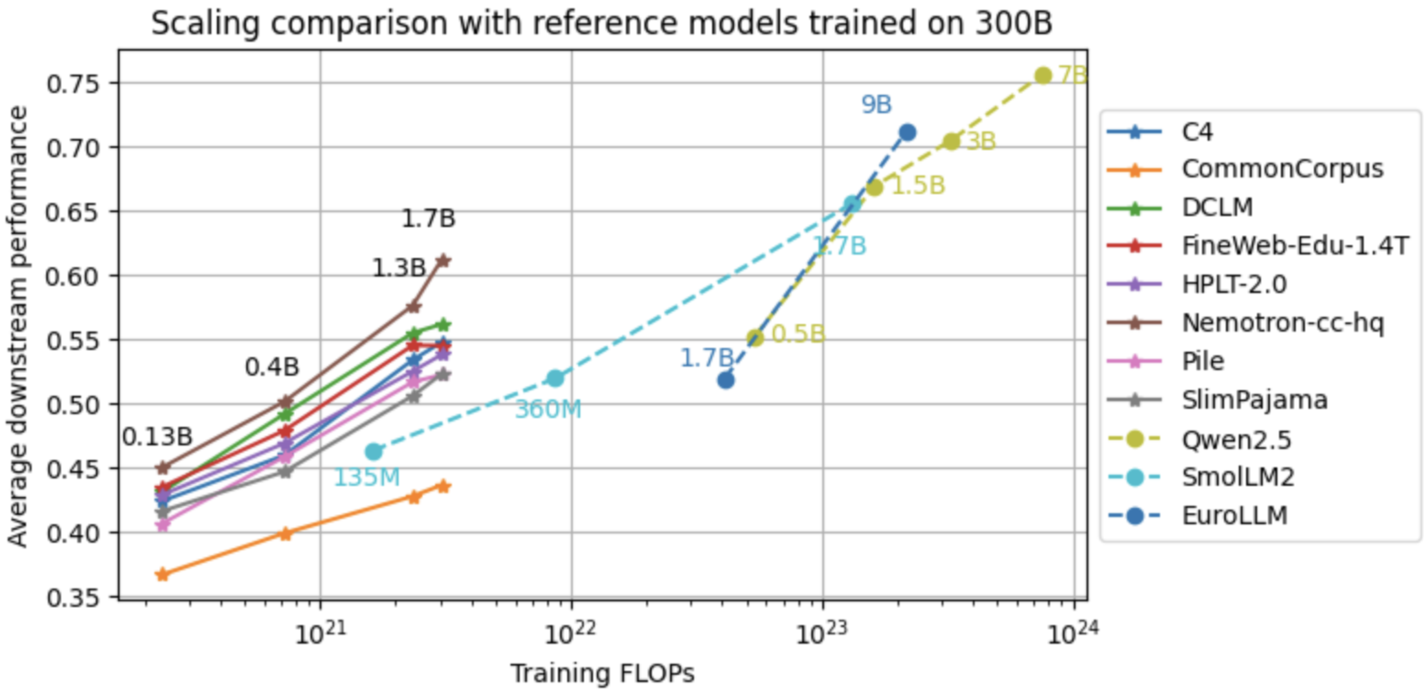
- Open-sci-ref-0.01 : set of reference baseline models to provide grounds for sanity checks and allow fair comparison on aligned compute/data




open-sci collective

# Open foundation models & datasets

- Open-sci-ref-0.01 : fair comparison on aligned compute/data




open-sci collective

OPEN EURO LLM

LAION

# Open foundation models with strong reasoning

https://arxiv.org/abs/2506.04178

open-sci
collective

LAION

Open Thoughts

DATA RECIPES FOR REASONING MODELS

**Etash Guha[*1,2], Ryan Marten[*3], Sedrick Keh[*4], Negin Raoof[*5], Georgios Smyrnis[*6], Hritik Bansal[ς7], Marianna Nezhurina[ς8,9,16], Jean Mercat[ς4], Trung Vu[ς3], Zayne Sprague[ς6], Ashima Suvarna[7], Benjamin Feuer[10], Liangyu Chen[1], Zaid Khan[11], Eric Frankel[2], Sachin Grover[12], Caroline Choi[1], Niklas Muennighoff[1], Shiye Su[1], Wanjia Zhao[1], John Yang[1], Shreyas Pimpalgaonkar[3], Kartik Sharma[3], Charlie Cheng-Jie Ji[3], Yichuan Deng[2], Sarah Pratt[2], Vivek Ramanujan[2], Jon Saad-Falcon[1], Jeffrey Li[2], Achal Dave, Alon Albalak[13], Kushal Arora[4], Blake Wulfe[4], Chinmay Hegde[10], Greg Durrett[6], Sewoong Oh[2], Mohit Bansal[11], Saadia Gabriel[7], Aditya Grover[7], Kai-Wei Chang[7], Vaishaal Shankar, Aaron Gokaslan[14], Mike A. Merrill[1], Tatsunori Hashimoto[1], Yejin Choi[1], Jenia Jitsev[8,9,16], Reinhard Heckel[15], Maheswaran Sathiamoorthy[3], Alexandros G. Dimakis[†3,5], Ludwig Schmidt[†1]**

[1]Stanford University, [2]University of Washington, [3]BespokeLabs.ai, [4]Toyota Research Institute, [5]UC Berkeley, [6]UT Austin, [7]UCLA, [8]JSC, [9]LAION, [10]NYU, [11]UNC Chapel Hill, [12]ASU, [13]Lila Sciences, [14]Cornell Tech [15]TUM [16]Open-Ψ (Open-Sci) Collective

# Open foundation models with strong reasoning

Making **whole pipeline** for reasoning foundation models – dataset composition, model training, benchmarks & evaluation – **fully reproducible**

open-sci collective

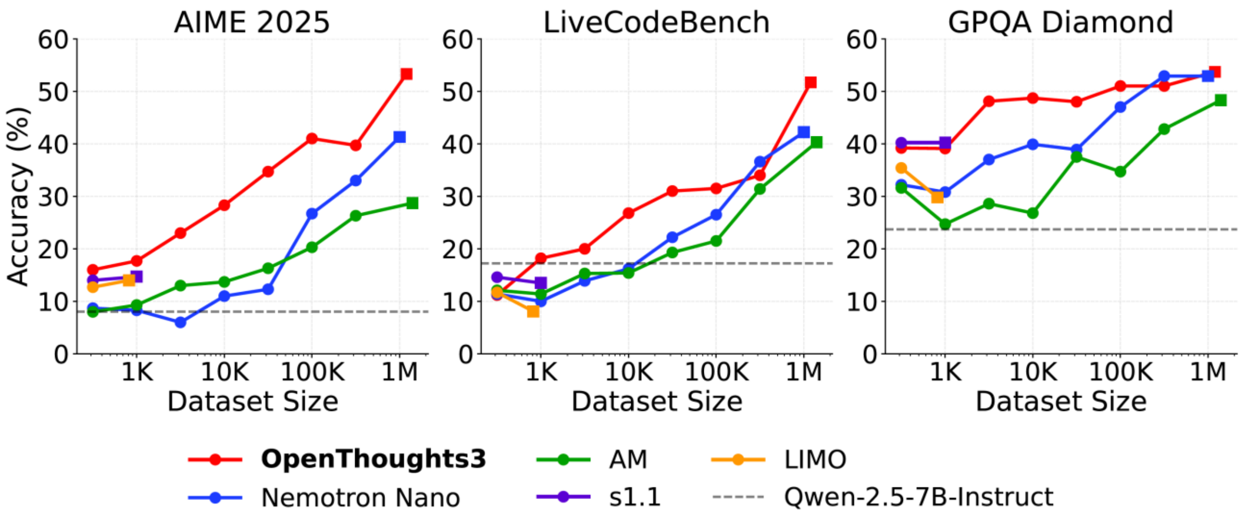


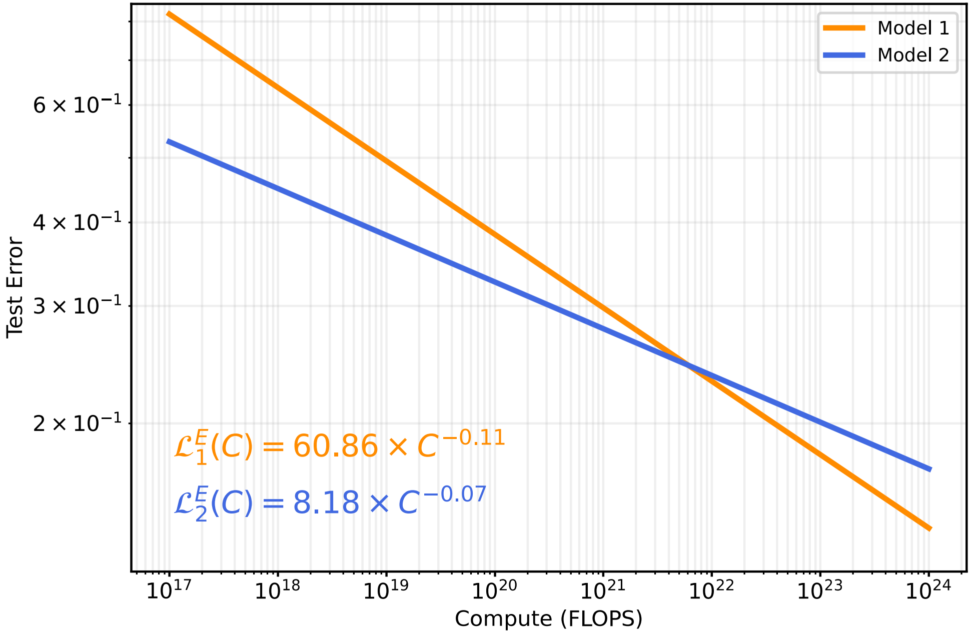

Figure 1: **OpenThoughts3 outperforms existing SFT reasoning datasets across data scales.** All models are finetuned from Qwen-2.5-7B-Instruct. We compare to large SFT datasets (AM, Nemotron Nano) and small curated datasets (s1.1, LIMO) on AIME 2025 (left), LiveCodeBench 06/24-01/25 (middle), and GPQA Diamond (right). Scaling curves for all evaluation benchmarks are in Figure 8.

Guha et al, ArXiv:2506.04178, 2025

# Scaling laws: learning procedure comparison

- Comparison requires scaling law derivation using standardized open procedures
  - measuring sufficient scaling span instead a single reference point
  - conducting by fully controlling dataset composition, training, transfer/evals

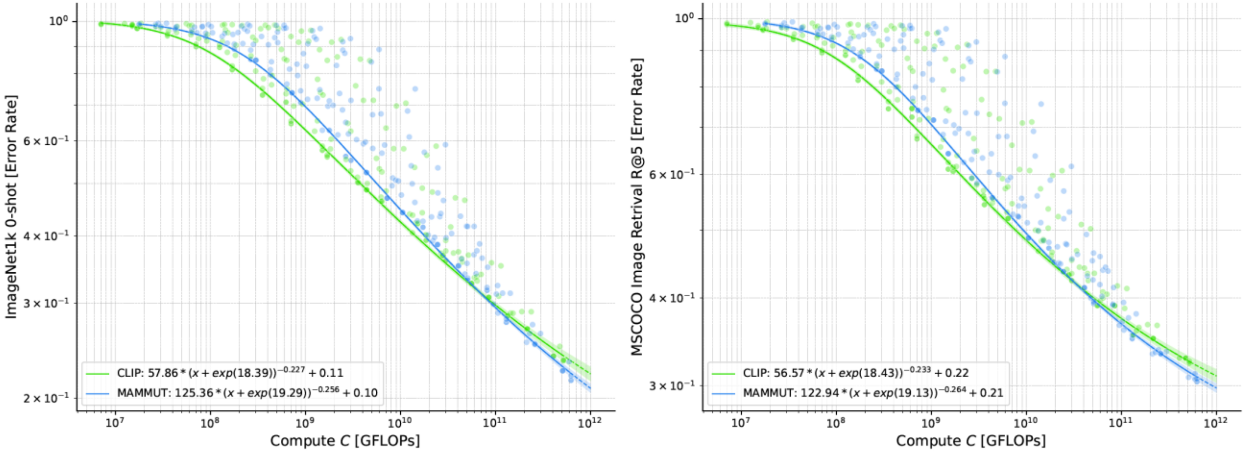
$$\mathcal{L}(C) = C_c \cdot C^{-\alpha_C} + L_\epsilon$$




- Learning procedure 1 vs Learning procedure 2
- Scenarios:
  - Comparing Model 1 vs Model 1 while fixing same open data
  - Comparing open Dataset 1 vs Dataset 2 while fixing same open training/model
  - ...

# Scaling laws: learning procedure comparison

- Comparing foundation models/datasets via scaling law derivation using open pipelines




(a) ImageNet-1k 0-shot classification

(b) MS-COCO image R@5

Figure 1: **Scaling on DataComp-1.4B.** Comparison of CLIP and MaMMUT via scaling laws on DataComp-1.4B. Error rate on downstream tasks is plotted against compute. MaMMUT outperforms CLIP in terms of scalability, indicated by crossing scaling law fit lines, where MaMMUT takes over CLIP in performance from larger compute scale $> 10^{11}$ GFLOPS on.

Nezhurina et al, ArXiv:2506.04598, 2025

# Scaling laws: learning procedure comparison

- Comparing foundation models/datasets via scaling law derivation using open pipelines

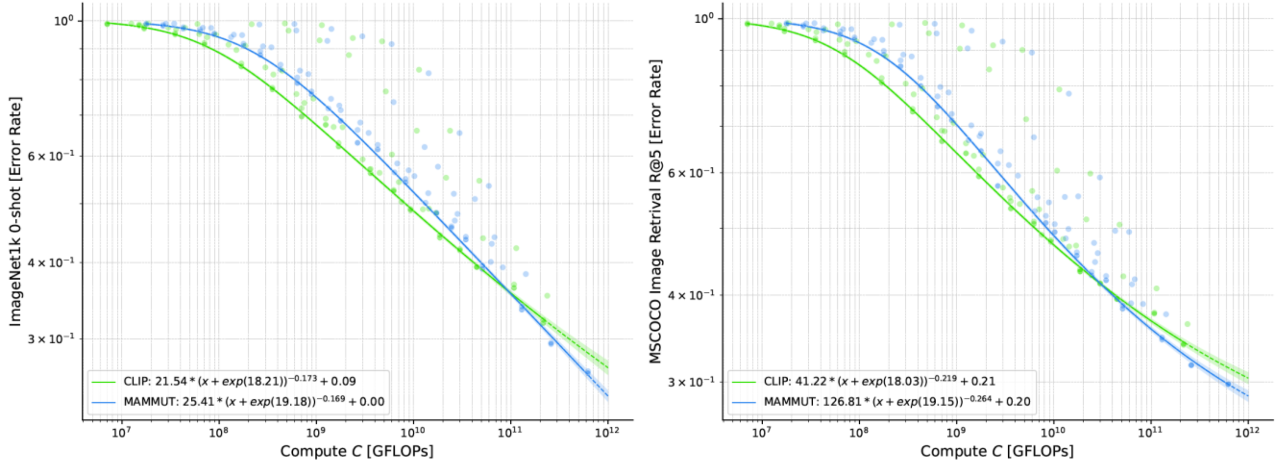
| Model | Samples Seen | GFLOPs | IN1k 0-shot acc | Predicted IN1k 0-shot acc (95% CI) | Predicted (more points) IN1k 0-shot acc (95% CI) |
|---|---|---|---|---|---|
| **CLIP** | | | | | |
| ViT-L-16 | 3.07e+9 | 4.07e+11 | 0.761 | 0.747 (0.738, 0.755) | – |
| ViT-L-14 | 3.07e+9 | 5.18e+11 | 0.766 | 0.753 (0.744, 0.762) | 0.759 (0.751, 0.766) |
| ViT-H-14 | 3.07e+9 | 1.14e+12 | 0.784 | 0.773 (0.761, 0.784) | 0.779 (0.770, 0.789) |
| *RMSE: 1.26e-02* | *RMSE (more points): 5.90e-03* | | | | |
| **MaMMUT** | | | | | |
| mammut-ViT-L-14 | 1.28e+9 | 2.59e+11 | 0.749 | 0.743 (0.737, 0.748) | – |
| mammut-ViT-L-14 | 3.07e+9 | 6.22e+11 | 0.784 | 0.773 (0.765, 0.781) | 0.777 (0.771, 0.783) |
| mammut-ViT-H-14 | 3.07e+9 | 1.43e+12 | 0.796 | 0.797 (0.787, 0.807) | 0.801 (0.793, 0.809) |
| *RMSE: 7.57e-03* | *RMSE (more points): 7.57e-03* | | | | |

Table 8: Predictions for different values of $C_{\text{threshold}}$ for the functional form with double saturation (Eq. 1). Scaling law derivation on DataComp-1.4B. The last column shows updated predictions made after additional data points. Both confidence interval and RMSE decrease as we take more points. RMSE is consistently lower than RMSE measured for functional form without irreducible error (Tab. 9).

Nezhurina et al, ArXiv:2506.04598, 2025

# Scaling laws: learning procedure comparison

- Comparing foundation models/datasets via scaling law derivation



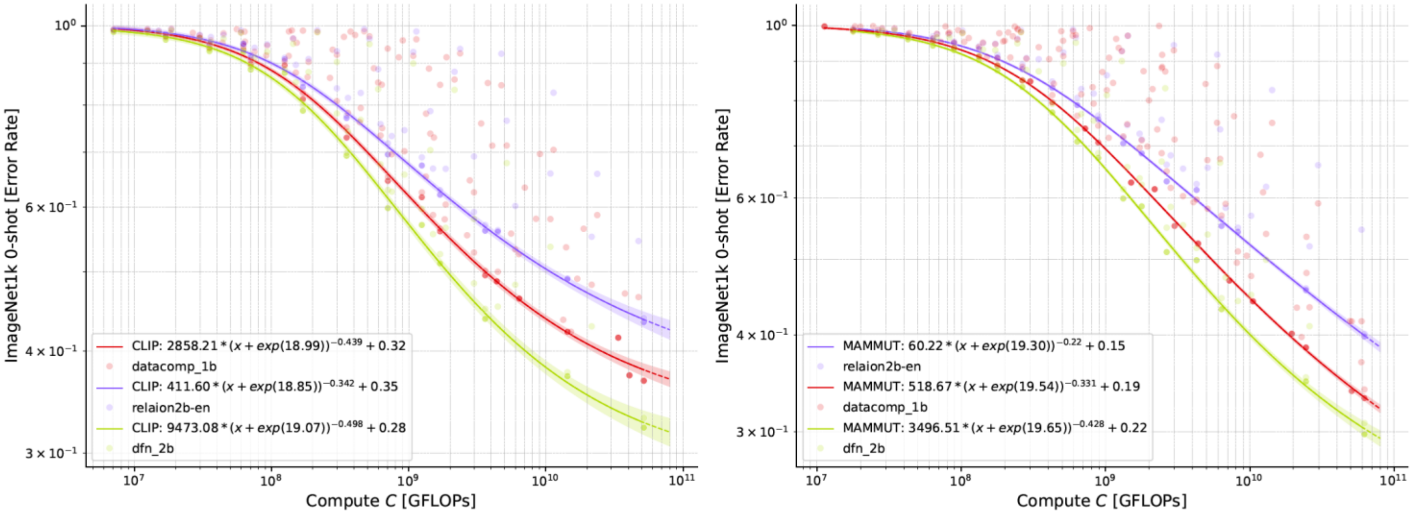

(a) ImageNet-1k 0-shot classification (b) MS-COCO image R@5

Figure 2: **Scaling on Re-LAION-1.4B.** Comparison of CLIP and MaMMUT via scaling laws on Re-LAION-1.4B. Error rate on downstream tasks is plotted against compute. MaMMUT outperforms CLIP in terms of scalability, indicated by crossing scaling law fit lines, where MaMMUT takes over CLIP in performance from larger compute scale $> 10^{11}$ GFLOPS on, showing similar trends as on DataComp-1.4B.

Nezhurina et al, ArXiv:2506.04598, 2025

# Scaling laws: learning procedure comparison

- Comparing foundation models/datasets via scaling law derivation



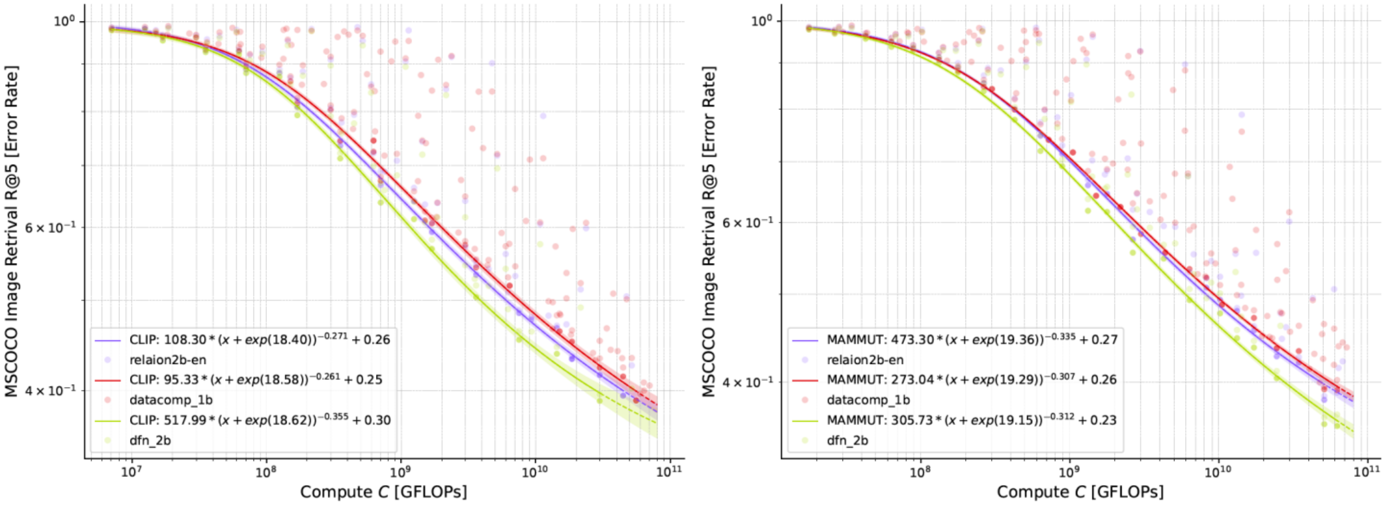

(a) IN-1k 0-shot error rate for openCLIP (b) IN-1k 0-shot error rate for openMaMMUT

Figure 6: Scaling laws for IN1k 0-shot performance of openCLIP (left) and openMaMMUT (right), comparing training on Re-LAION-1.4B, DataComp-1.4B and DFN-1.4B. Training on DFN-1.4B results in superior performance across scales consistently for both architectures.

Nezhurina et al, ArXiv:2506.04598, 2025

# Scaling laws: learning procedure comparison

- Comparing foundation models/datasets via scaling law derivation




(a) Error Rate for CLIP

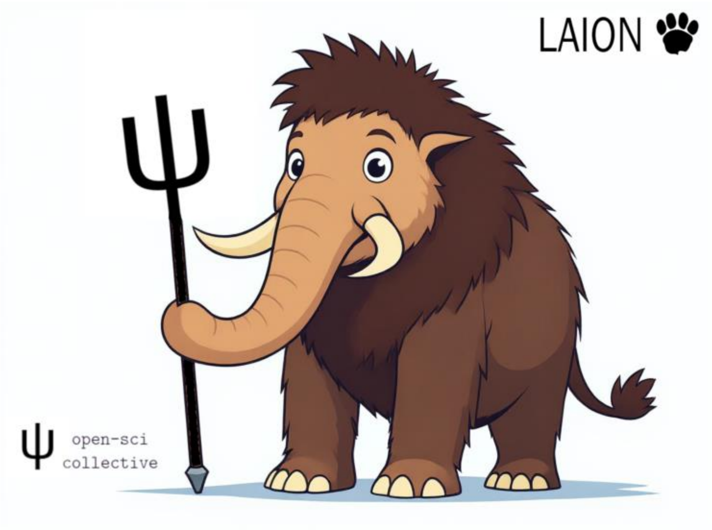
(b) Error Rate for MaMMUT

Figure 7: Scaling laws for MS-COCO image retrieval performance (1- Recall@5) of openCLIP (left) and openMaMMUT (right), comparing training on Re-LAION-1.4B, DataComp-1.4B and DFN-1.4B. Training on DFN-1.4B results again in superior performance across scales consistently for both architectures.

Nezhurina et al, ArXiv:2506.04598, 2025

# Open foundation models with stronger scalability

- LAION as open frontiers lab: openMaMMUT predictably matching or outperforming SOTA of closed labs




| ViT | Res. | Seq. | Model | Dataset | #Samples | ImageNet-1k val | ImageNet-1k v2 | COCO T→I | COCO I→T |
|---|---|---|---|---|---|---|---|---|---|
| L/16 | 256 | 256 | SigLIP [18] | WebLI-10B | 40B | 80.44 | 73.76 | 75.26 | 88.40 |
| | | | SigLIP 2 [14] | WebLI-10B | 40B | 82.35 | 76.66 | 76.84 | 90.44 |
| L/14 | 224 | 256 | OpenCLIP [10] | LAION-2B | 34B | 75.24 | 67.73 | 70.46 | 84.30 |
| | | | CLIP [7] | WIT-400M | 12.8B | 75.54 | 69.84 | 59.95 | 79.56 |
| | | | MetaCLIP [45] | MetaCLIP-2.5B | 12.8B | 79.19 | 72.64 | 71.36 | 84.94 |
| | | | EVA-CLIP [46] | Merged-2B | 4B* | 79.75* | 72.92* | 70.68 | 85.26 |
| | | | DFN [20] | DFN-2B | 13B | **81.41*** | **74.58*** | **73.19*** | **86.20*** |
| | | | DataComp [19] | DataComp-1.4B | 12.8B | 79.19 | 72.06 | 69.86 | 84.64 |
| | | | **OpenMaMMUT (Ours)** | DataComp-1.4B | 12.8B | 80.34 | 73.78 | 71.19 | 85.88 |

Table 3: Zero-shot classification (accuracy) and retrieval (R@5) results. DFN used ImageNet/MS-COCO-finetuned model for data filtering; EVA-CLIP was initialized from models pre-trained on ImageNet. We use **bold** for best overall results, gray for models involving ImageNet/MS-COCO data as training data in pipeline, and underlined for best results without ImageNet/MS-COCO involvement.

Nezhurina et al, ArXiv:2506.04598, 2025

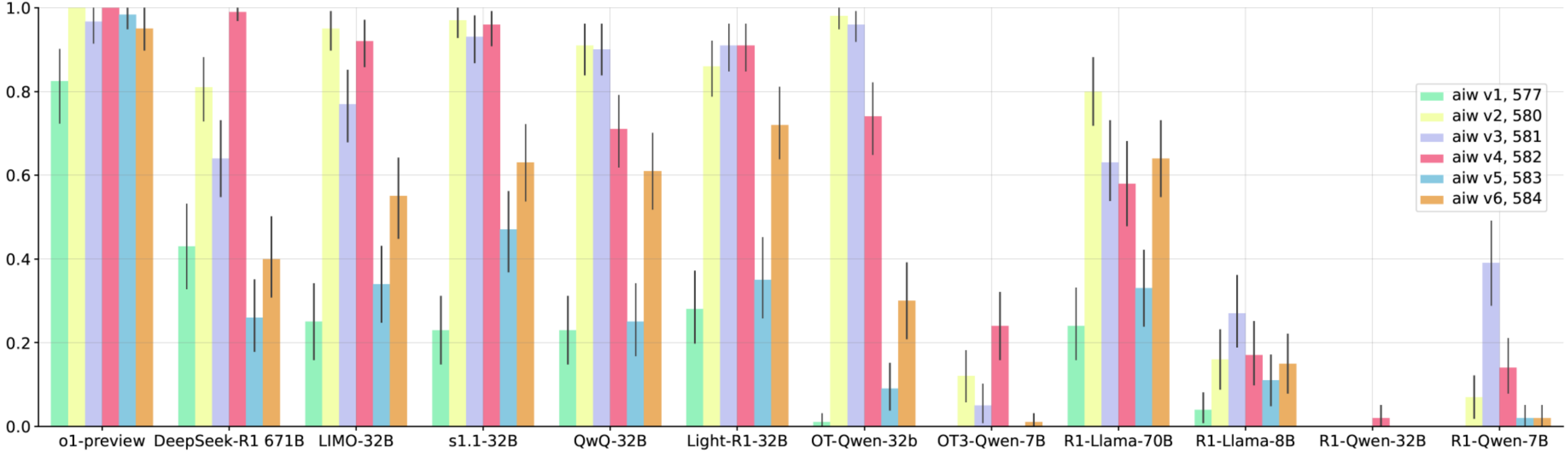
# Open foundation models: measuring it right

- Improve generalization & reasoning evals!
- Testing claims of  strong function (olympiad & graduate level) with simple problems & their variations (AIW problems)

AIW Friends, Variations 1-6, Prompt IDs: 577 580 581 582 583 584

Variation 1: Alice has **3 male friends** and she also has **6 female friends**. [Correct answer: **7** ]
Variation 2: Alice has **2 female friends** and she also has **4 male friends**. [Correct answer: **3** ]
Variation 3: Alice has **4 female friends** and she also has **1 male friend**. [Correct answer: **5** ]
Variation 4: Alice has **4 male friends** and she also has **1 female friend**. [Correct answer: **2** ]
Variation 5: Alice has **2 male friends** and she also has **3 female friends**. [Correct answer: **4** ]
Variation 6: Alice has **5 female friends** and she also has **3 male friends**. [Correct answer: **6** ]

*All mentioned persons are friends with each other and have no other friends aside.*
*How many female friends does male friend of Alice have?*

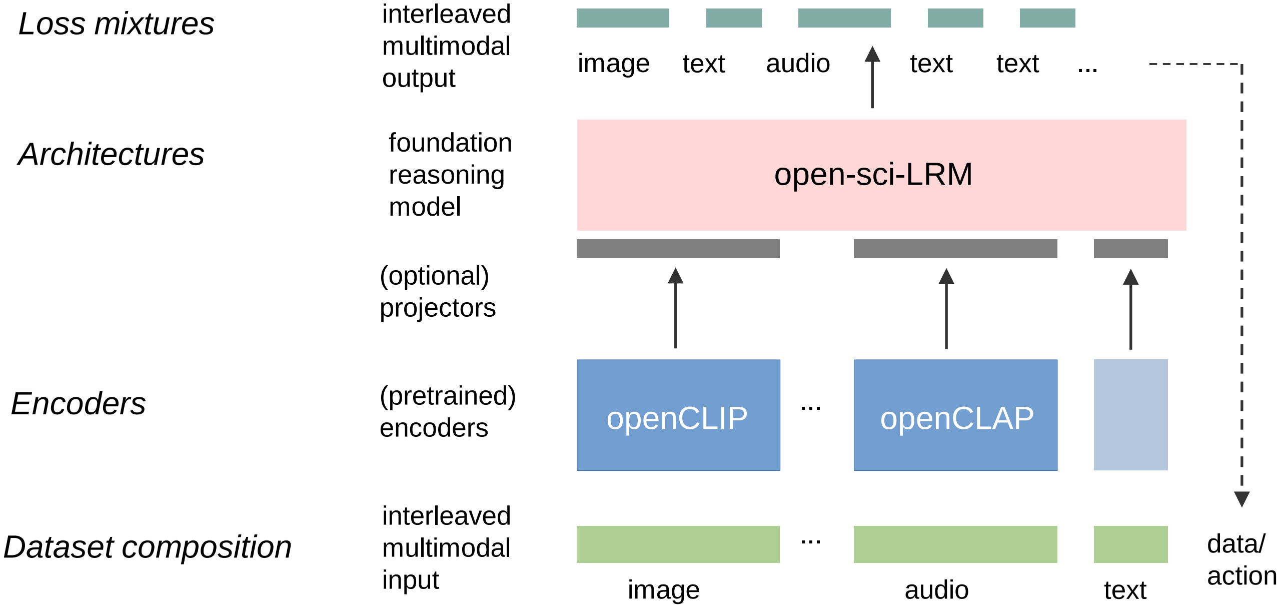



Nezhurina et al, ArXiv:2406.02061, NeurIPS SciForDL, 2024; Guha et al, ArXiv:2506.04178, 2025

LAION open-sci collective Open Thoughts

# Open multi-modal foundation models: progress

- Reference scaling laws for guided search of scalable open FoMos
- Comparison to reference scaling laws for established FoMo designs
  - eg Llava-Next: pretrained FoMos, post-training on smaller scale multi-modal instruction data




OPEN EURO LLM

open-sci collective

ELLIOT

LAION

LAION

# Open foundation models: outlook

- „Moonshot“: **open-sci-MMA – strong open multi-modal foundation action model family, learning with any modality – text, vision, audio, ...**
  - Securing souvereignity in basic research on foundations of ML/AI
  - Requires dedicated, large-scale compute!

- BigScience BLOOM: GPT-3 replication, dedicated partition of 480 GPUs (Jean Zay, Paris Saclay). Back 2021 → ca. 650K A100 GPU hours; ca. 3 months training

- Now: DeepSeek R1 level models (optimized), language only: ca. 4M H100 GPU hours → ca. 1 week on **whole** JUPITER for **single training run ...**

- Multi-modal foundation models: at least 10x more compute → almost **6 months** for single training run taking **whole** JUPITER (24k H100 GPUs)

Without dedicated partitions / machines : **basic research impossible**

# Open foundation models and datasets: alliance

- **OSFoMo Alliance** : **Coordination of colab and resource acquisition** for open source foundation models and datasets R & D
- Build by orgas with strong track of record researching and building open FoMos
  - HuggingFace (**EU**), BlackForestLabs (**EU**), PriorLabs (**EU**), LAION (**EU**), TogetherAI, EleutherAI, AllenAI, ...
- Define important open FoMo & datasets to be researched & maintained as open-source
- **Common grant applications for compute** and fund resources
- Possible milestones
  - Open foundation reasoning models & datasets (DeepSeek level), strong reasoning and generalization
  - Open multi-modal language action models & datasets (transferable backbone for agents, open OS for robotics & automonous systems)

# Acknowledgements


Dr. Mehdi Cherti, Marianna Nezhurina, JSC

Visit https://laion.ai/
Join public LAION Discord server for more projects and research tracks
> 30k members !

LAION community & friends (Romain Beaumont, Ross Wightmann, Irina Rish, ...)

Prof. Ludwig Schmidt, Stanford

Christoph Schumann

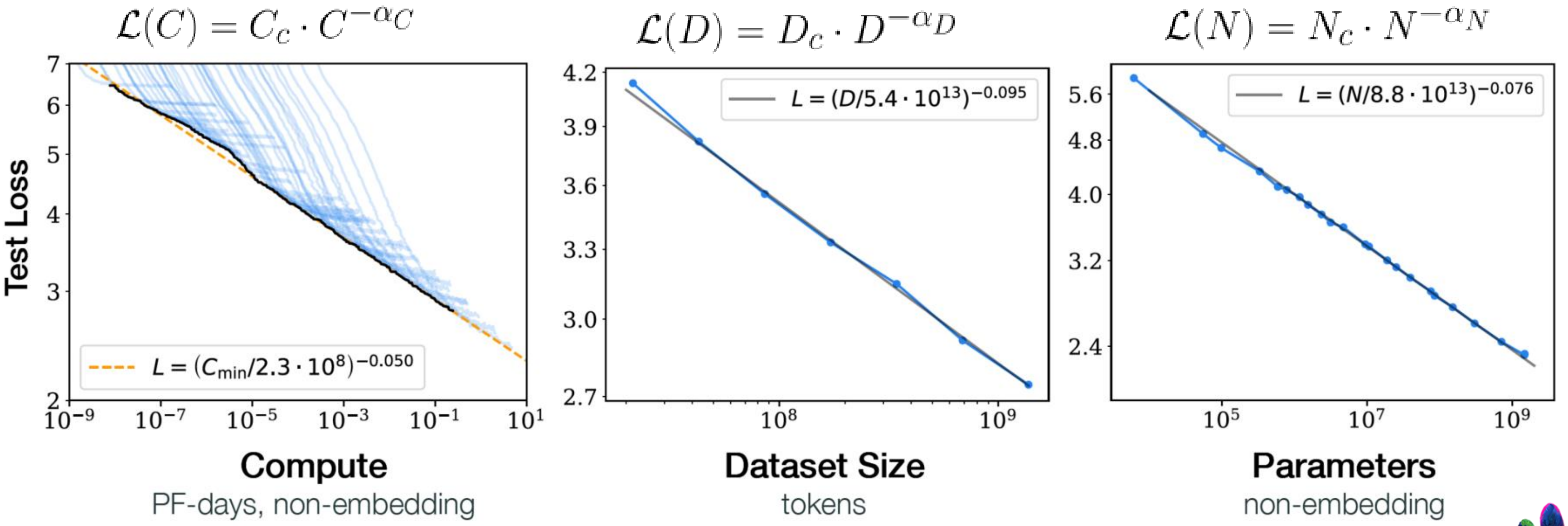
**Let's build open, robust, safe AI foundations together!**

Thanks
for
your
**Attention**

# Supplementary Material

# Foundation models: scaling laws

- **Scaling Laws**: predicting model properties and function across scales




Kaplan et al, 2020

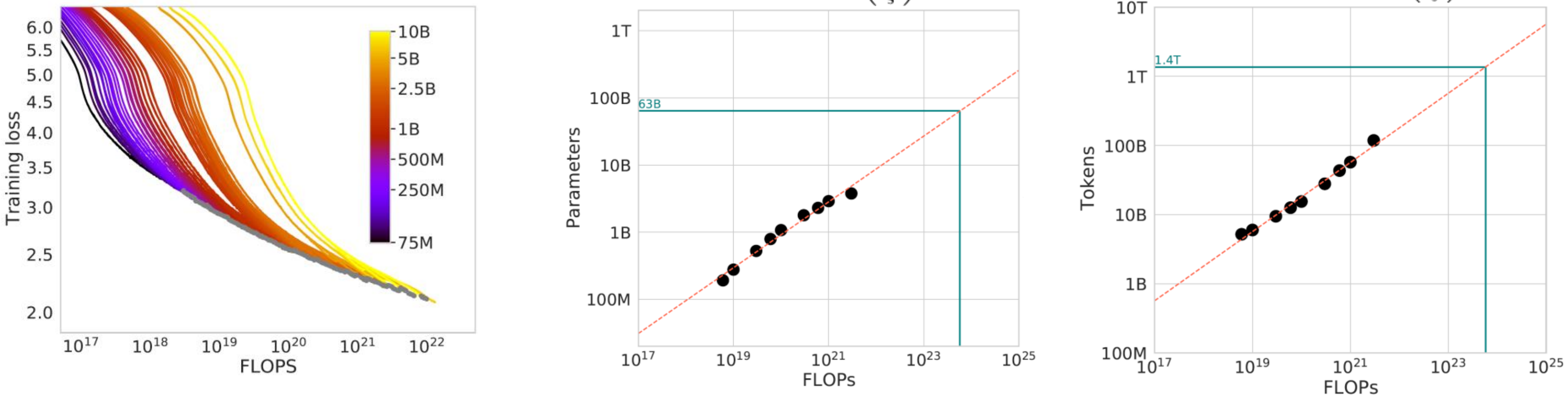
# Foundation models: scaling laws

- **Scaling Laws**: predicting model properties and function across scales

$$\mathcal{L}(C) = C_c \cdot C^{-\alpha_C} + L_\epsilon$$

$$N^*(C) = G\left(\frac{C}{\xi}\right)^a$$

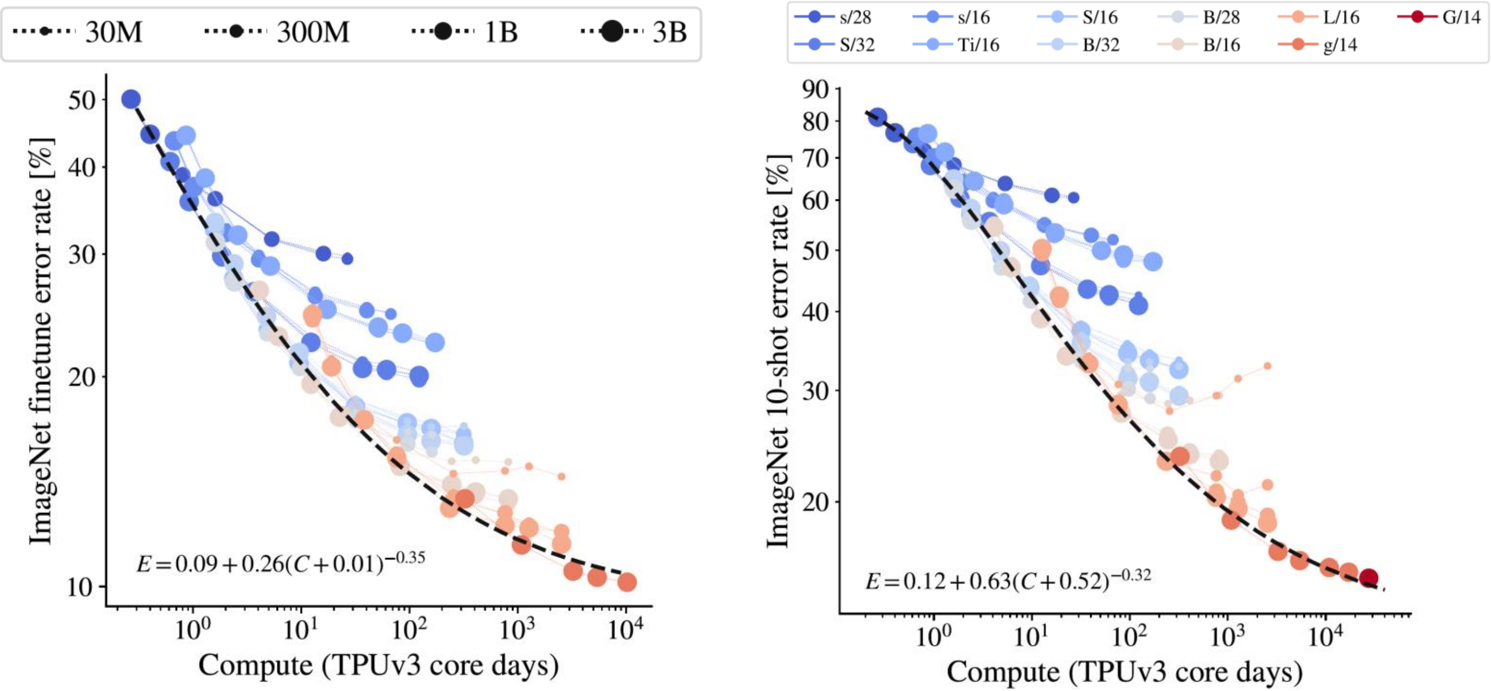
$$D^*(C) = G^{-1}\left(\frac{C}{\xi}\right)^b$$




Hoffmann et al, arXiv:2203.15556, NeurIPS, 2022

# Foundation models: scaling laws

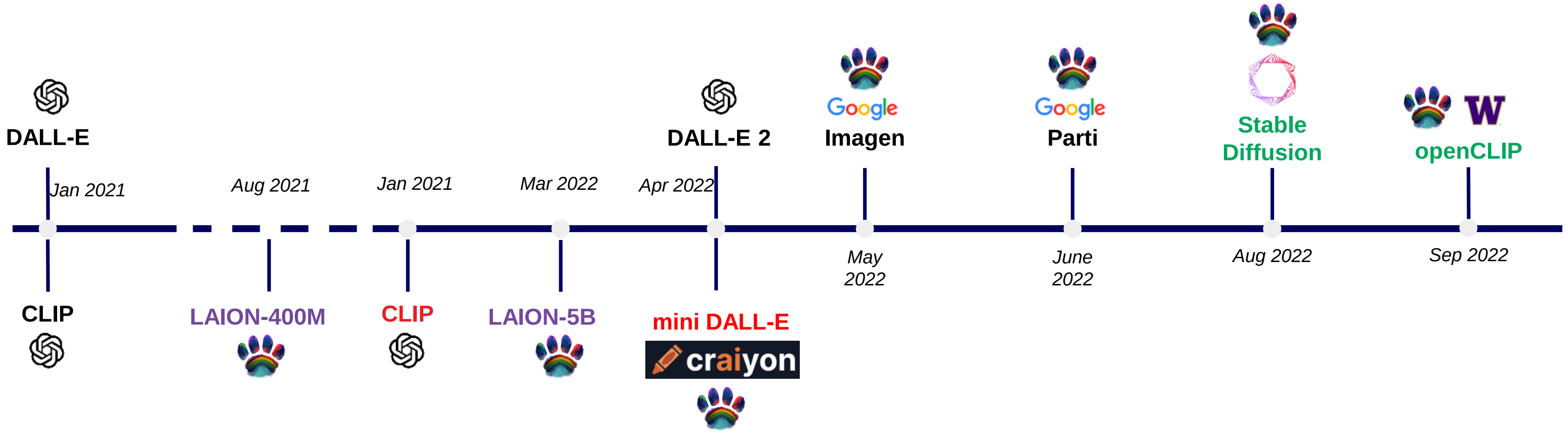
- Scaling Laws: exist for various generalist learning procedures
- Example: Supervised classification, ViT (JFT-3B dataset)




Zhai et al, arXiv:2106.04560, CVPR, 2022

# From closed to open data and models: a timeline

- Open-source releases fertilize research and technology development




*Closed model in black*
*Open release pre-trained models in red*
*Open data in purple*
*Open foundation models in green*

Adapted from State of AI report, 2022

# Open foundation models: building on foundations

**Taming Transformers for High-Resolution Image Synthesis**

Patrick Esser* Robin Rombach* Björn Ommer
Heidelberg Collaboratory for Image Processing, IWR, Heidelberg University, Germany
*Both authors contributed equally to this work

**CVPR, 2021** VQGAN encoder/decoder: open-source release

**High-Resolution Image Synthesis with Latent Diffusion Models**

Robin Rombach[1] * Andreas Blattmann[1] * Dominik Lorenz[1] Patrick Esser[R] Björn Ommer[1]
[1]Ludwig Maximilian University of Munich & IWR, Heidelberg University, Germany [R]Runway ML

**CVPR, 2022**
Latent Diffusion model: open-source release

**+**

**NeurIPS, 2022, (Outstanding paper award)**

**LAION-5B: A NEW ERA OF OPEN LARGE-SCALE MULTI-MODAL DATASETS**

**Reproducible scaling laws for contrastive language-image learning**

Mehdi Cherti[1,5] §§ Romain Beaumont[1] §§ Ross Wightman[1,3] §§
Mitchell Wortsman[4] §§ Gabriel Ilharco[4] §§ Cade Gordon[2]
Christoph Schuhmann[1] Ludwig Schmidt[1,4] °° Jenia Jitsev[1,5] §§°°
LAION[1] UC Berkeley[2] HuggingFace[3] University of Washington[4]
Juelich Supercomputing Center (JSC), Research Center Juelich (FZJ)[5]
contact@laion.ai, {m.cherti,j.jitsev}@fz-juelich.de
§§ Equal first contributions, °° Equal senior contributions

**CVPR, 2023**

LAION-5B image-text dataset, openCLIP models: open-source release

**Open-source power**

**=**

Stable Diffusion: **Latent Diffusion** + **openCLIP** + **LAION datasets**

*Stable Diffusion 1.5, trained on **LAION-5B** image-text dataset.*
*Prompt: "An epic scene of a supercomputing center building of the future, embedded in a rich wild green exotic blooming jungle forest, nearby a lake“*

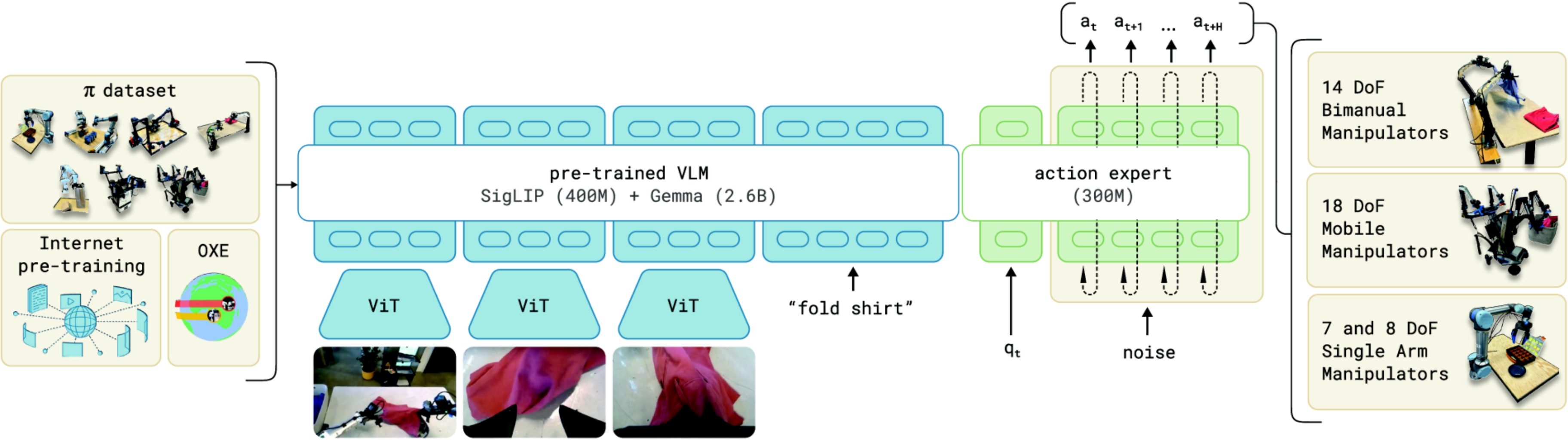
# Open science for large-scale foundation models

- Open-sourcing whole foundation model research pipeline, case LAION-openCLIP studies

| | |
|---|---|
| Dataset curation & composition | Open-source (img2dataset, datacomp) |
| Dataset | Publicly accessible (ReLAION-5B) |
| Model training | Open-source (OpenCLIP) |
| Model evaluation | Open-source (CLIPBenchmark) |
| Model weights | Open-weights (LAION CLIP) |

# Foundation models from re-usable components

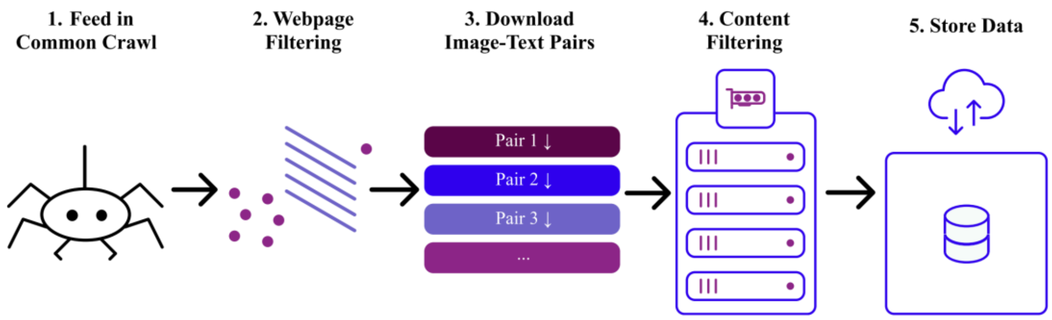
- Combining pre-trained models into multi-modal generalist foundation models (no or little adaptation required): Flamingo, BLIP-2, ImageBind, LENS, LlaVA, EMU, MM-1, PaliGemma, ...




Black et al, Physical Intelligence, 2024

# Open large-scale reference/foundation data

- LAION-400M/5B: Open sourcing data collection procedures - transparent dataset, open source toolsets, reproducible training across various scales (**NeurIPS Outstanding Paper Award 2022**)
- Open dataset: collection of text and links to images on public Internet




| Dataset | # English Img-Txt Pairs |
|---|---|
| **Public Datasets** | |
| MS-COCO | 330K |
| CC3M | 3M |
| Visual Genome | 5.4M |
| WIT | 5.5M |
| CC12M | 12M |
| RedCaps | 12M |
| YFCC100M | 100M[2] |
| **LAION-5B (Ours)** | **2.3B** |
| **Private Datasets** | |
| CLIP WIT (OpenAI) | 400M |
| ALIGN | 1.8B |
| BASIC | 6.6B |

Schuhmann et al, NeurIPS, 2022

# Open large-scale reference/foundation data

- LAION-400M/5B: Open sourcing data collection procedures - transparent dataset, open source toolsets, reproducible training across various scales

C: Green Apple Chair

C: sun snow dog

C: pink, japan, aesthetic image

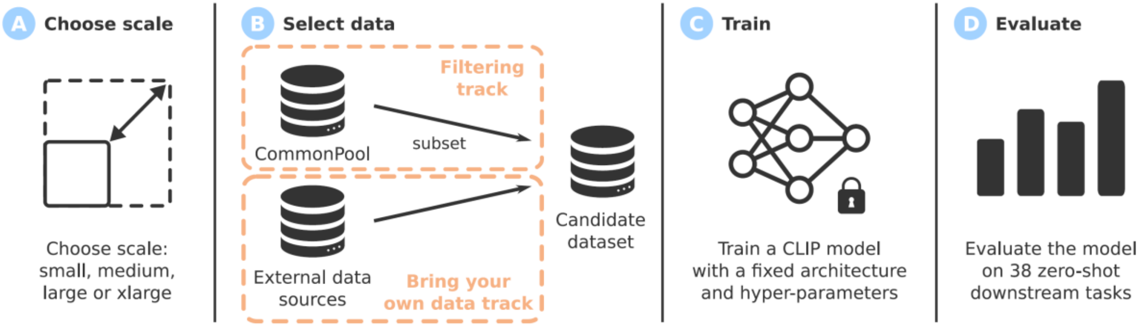
| Dataset | # English Img-Txt Pairs |
|---|---|
| **Public Datasets** | |
| MS-COCO | 330K |
| CC3M | 3M |
| Visual Genome | 5.4M |
| WIT | 5.5M |
| CC12M | 12M |
| RedCaps | 12M |
| YFCC100M | 100M[2] |
| **LAION-5B (Ours)** | **2.3B** |
| **Private Datasets** | |
| CLIP WIT (OpenAI) | 400M |
| ALIGN | 1.8B |
| BASIC | 6.6B |

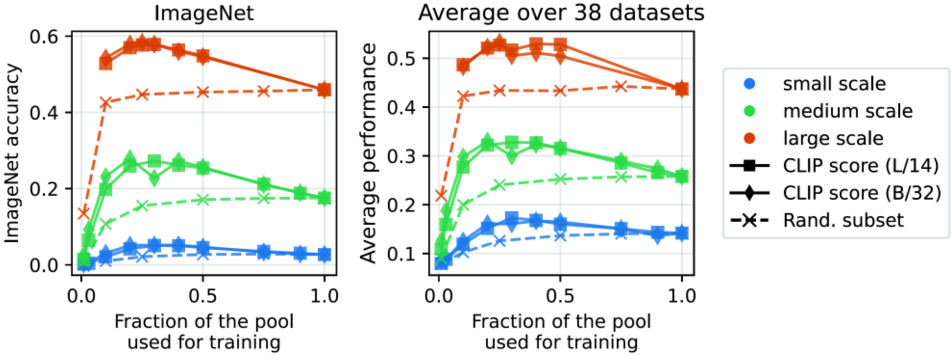
- Follow-ups: DataComp-1B; Re-LAION (safety revision update, Aug 2024)

Schuhmann et al, NeurIPS, 2022

# Data-centric scaling law interventions

- DataComp, DataComp-LM: what constitutes good data for FM training?




| Dataset | Dataset size | # samples seen | Architecture | Train compute (MACs) | ImageNet accuracy |
|---|---|---|---|---|---|
| OpenAI's WIT [111] | 0.4B | 13B | ViT-L/14 | $1.1 \times 10^{21}$ | 75.5 |
| LAION-400M [128, 28] | 0.4B | 13B | ViT-L/14 | $1.1 \times 10^{21}$ | 72.8 |
| LAION-2B [129, 28] | 2.3B | 13B | ViT-L/14 | $1.1 \times 10^{21}$ | 73.1 |
| LAION-2B [129, 28] | 2.3B | 34B | ViT-H/14 | $6.5 \times 10^{21}$ | 78.0 |
| LAION-2B [129, 28] | 2.3B | 34B | ViT-g/14 | $9.9 \times 10^{21}$ | 78.5 |
| DATACOMP-1B (ours) | 1.4B | 13B | ViT-L/14 | $1.1 \times 10^{21}$ | **79.2** |

Gadre et al, NeurIPS 2023 (Oral); Li et al, NeurIPS 2024

# Open foundation models: reproducibility

- Ingredients for an reproducible, open foundation model
  - open **large-scale dataset** & open dataset composition
  - open **pre-training** procedure (**compute intensive - supercomputers**)
  - open **transfer** procedures (zero-shot, linear probing, fine-tuning, ...)
  - open **standardized evaluation benchmarks** (eg: https://github.com/LAION-AI/CLIP_benchmark, https://github.com/EleutherAI/lm-evaluation-harness

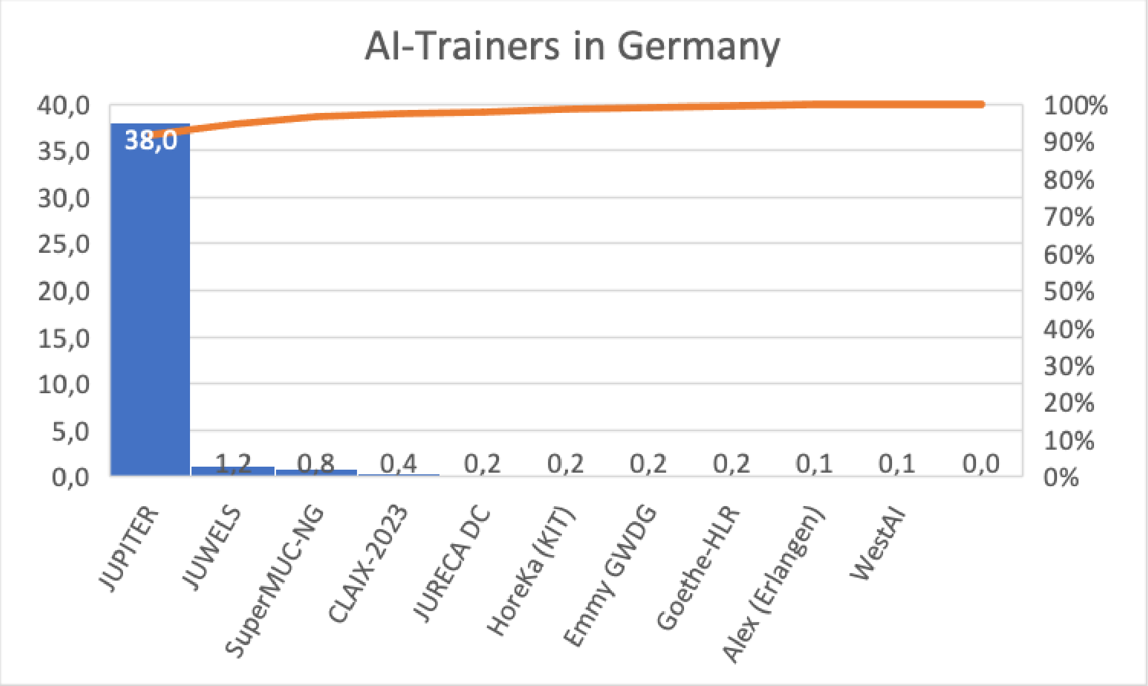
- → Enables **reproducible scaling laws** that can be used to
  - Perform learning procedure comparison
  - Guide search towards stronger scalable learning procedures

# Open science for large-scale foundation models

- **<u>Compute</u>**: using publicly funded supercomputers at JSC
    - **JUWELS Booster**: 3700 A100 GPUs, 40 GB per GPU
    - **JUPITER:** 24000 H100 GPUs (> 6x), 96 GB per GPU (Q3 2025)

# Open science for large-scale foundation models

- **Compute**: using publicly funded supercomputers at JSC
  - **JUWELS Booster**: 3700 A100, 1.2 ExaFLOPs, fp16
  - **JUPITER:** 24000 H100 GPUs, 38 ExaFLOPs, fp8



# Supercomputers for distributed training

- Distributed training on supercomputers requires scalable code

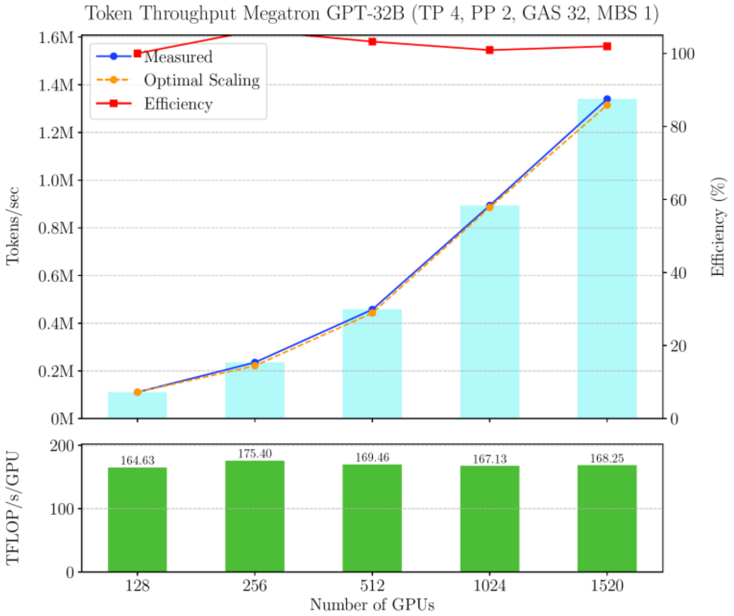



Figure 3: Throughput scalability of a 32B parameter GPT pretraining on 32 to 380 nodes on JUWELS Booster using MegaTron-LM, see also Suppl. Tab. 4. GPU utilization (A100 40GB) and token throughput achieve high numbers across various node configurations.

| Nodes | GPUs | Global BS | Tokens/Step | s/Step | TFLOP/s/GPU | Tokens/s | Efficiency (%) |
|---|---|---|---|---|---|---|---|
| 32 | 128 | 512 | 2,097,152 | 18.941 | 164.63 | 110,722 | 100.0 |
| 64 | 256 | 1024 | 4,194,304 | 17.830 | 175.40 | 235,234 | 106.2 |
| 128 | 512 | 2048 | 8,388,608 | 18.348 | 169.46 | 457,195 | 103.2 |
| 256 | 1024 | 4096 | 16,777,216 | 18.773 | 167.13 | 893,673 | 100.9 |
| 380 | 1520 | 6080 | 24,903,680 | 18.582 | 168.25 | 1,340,238 | 101.9 |