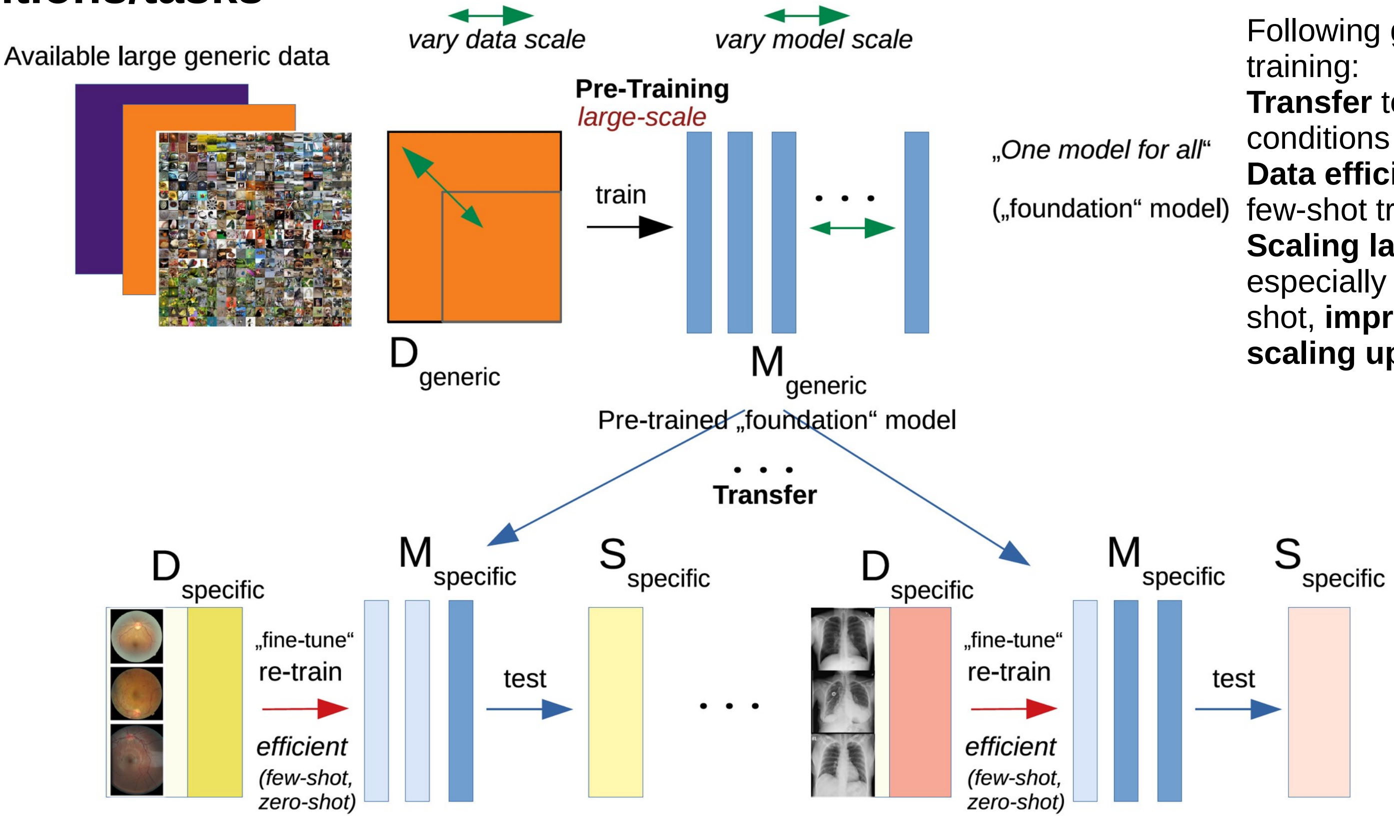
Open foundation models: scaling laws & generalization

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)

17. October 2025 | Jenia Jitsev

Foundation models: generic transferable learning

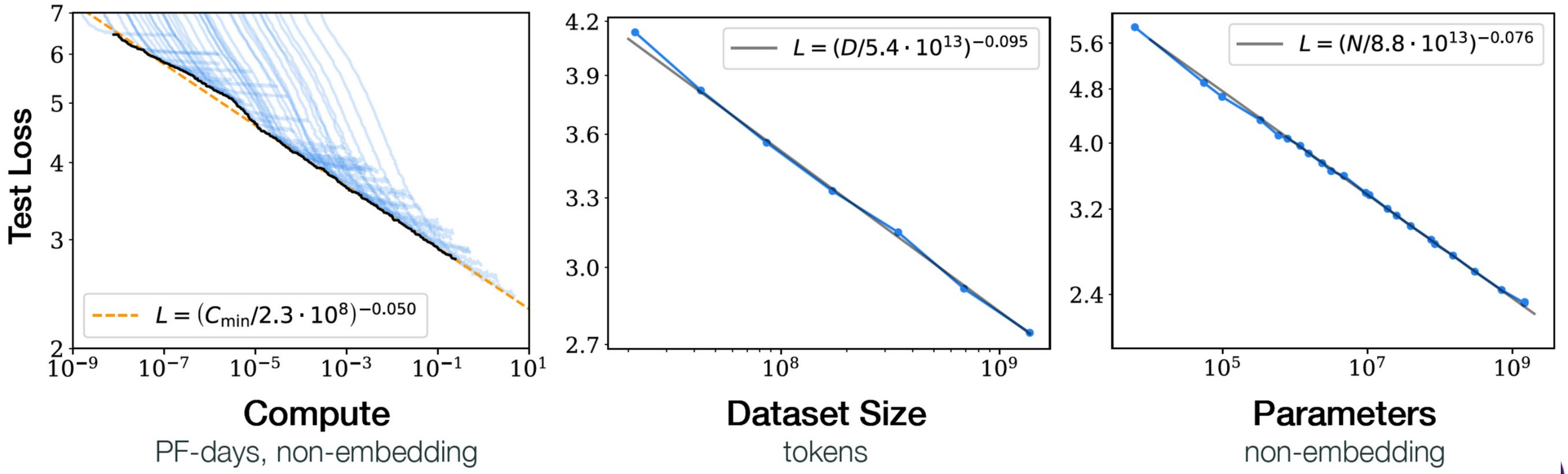
- Core breakthroughs (since ca. 2012): **learning that transfers across condintions/tasks**



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



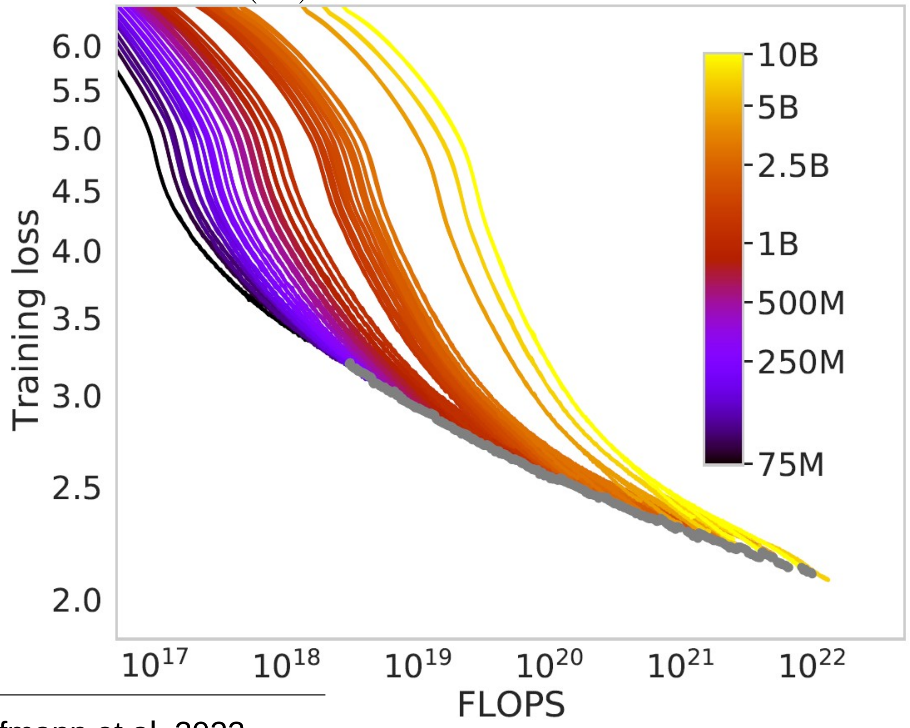
Kaplan et al, 2020

Foundation models: scaling laws

- Scaling law: fitting on Pareto front

Min Loss for given compute: compute-optimal scaling law

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



Hoffmann et al, 2022

Approx. for dense transformer C=6ND (Kaplan et al, 2020)

$$C = \xi N D$$

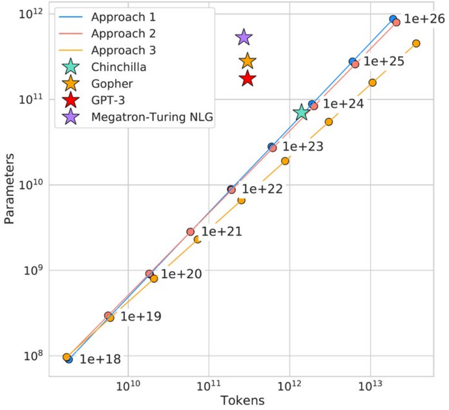
- Measure loss for various N, D combinations
  - Eg, fix N, go through increasing D, hypers tuning on a grid for each N,D combo training
- For each C(N,D), - **tuned hypers**! - get min L(C(N,D))
- Fit L(C) through those points

LM, text data tokens

Model N / Data (D)	75M	250M	500M	1B	2.5B	5B	10B
10B	L(N,D)	L(N,D)	L(N,D)	L(N,D)	L(N,D)	L(N,D)	L(N,D)
20B	L(N,D)	L(N,D)	L(N,D)	L(N,D)	L(N,D)	L(N,D)	L(N,D)
50B	L(N,D)	L(N,D)	L(N,D)	L(N,D)	L(N,D)	L(N,D)	L(N,D)
100B	L(N,D)	L(N,D)	L(N,D)	L(N,D)	L(N,D)	L(N,D)	L(N,D)
300B	L(N,D)	L(N,D)	L(N,D)	L(N,D)	L(N,D)	L(N,D)	L(N,D)

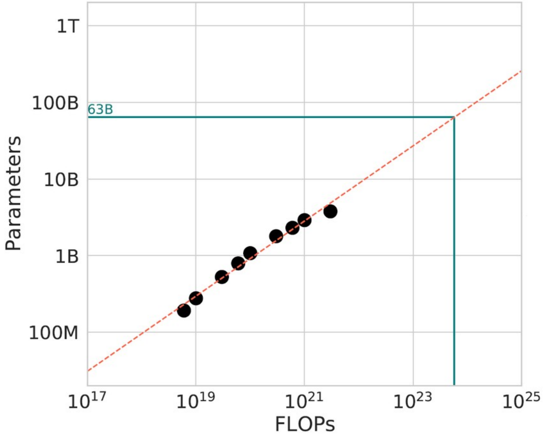
Foundation models: scaling laws

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

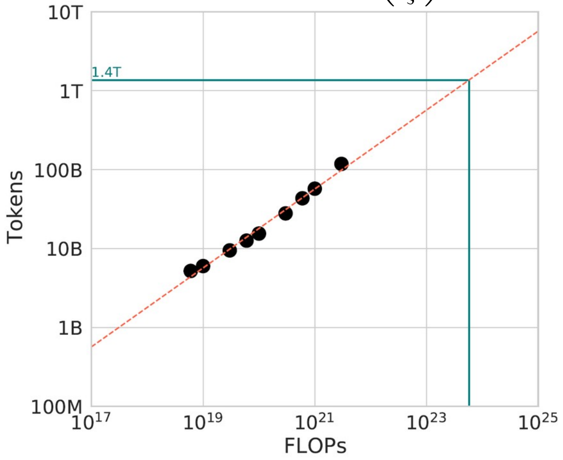


Parameters	FLOPs	FLOPs (in *Gopher* unit)	Tokens
400 Million	1.92e+19	1/29,968	8.0 Billion
1 Billion	1.21e+20	1/4,761	20.2 Billion
10 Billion	1.23e+22	1/46	205.1 Billion
67 Billion	5.76e+23	1	1.5 Trillion
175 Billion	3.85e+24	6.7	3.7 Trillion
280 Billion	9.90e+24	17.2	5.9 Trillion
520 Billion	3.43e+25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

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



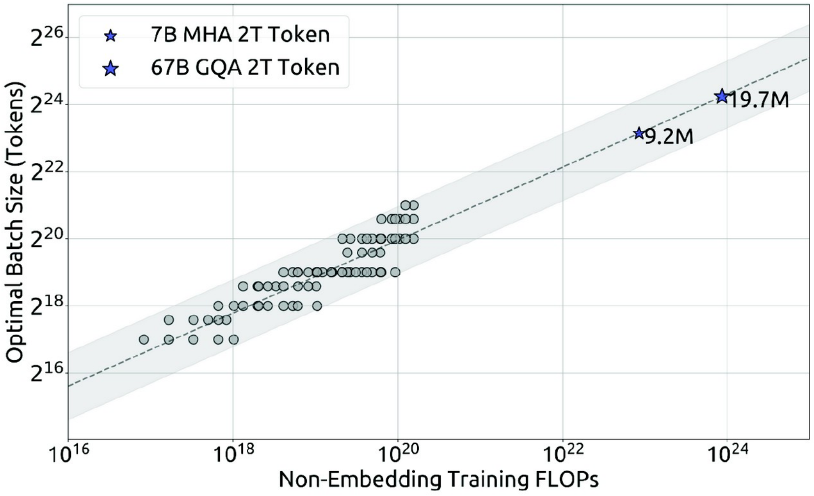
$$D^*(C) = G^{-1}\left(\frac{C}{\xi}\right)^b$$



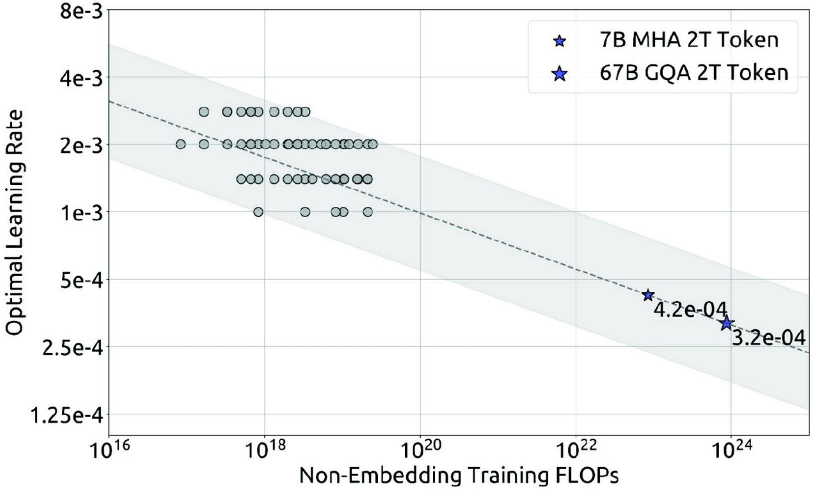
Hoffmann et al, arXiv:2203.15556, 2022

Foundation models: scaling laws

- Scaling law: predicting training/model properties and function
- Predictions are only accurate IF scaling law derivation is done properly!
- EXTREMELY IMPORTANT: TUNE hyperparams for each measurement



(a) Batch size scaling curve

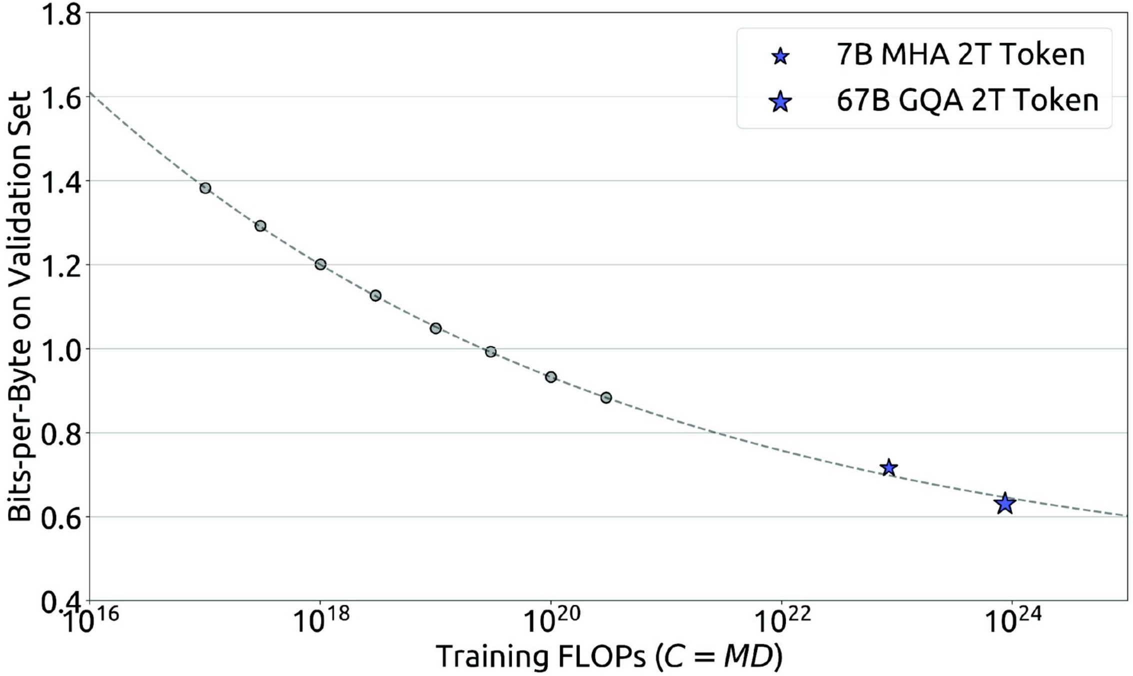


(b) Learning rate scaling curve

Bi et al, DeepSeek, 2024

Foundation models: scaling laws

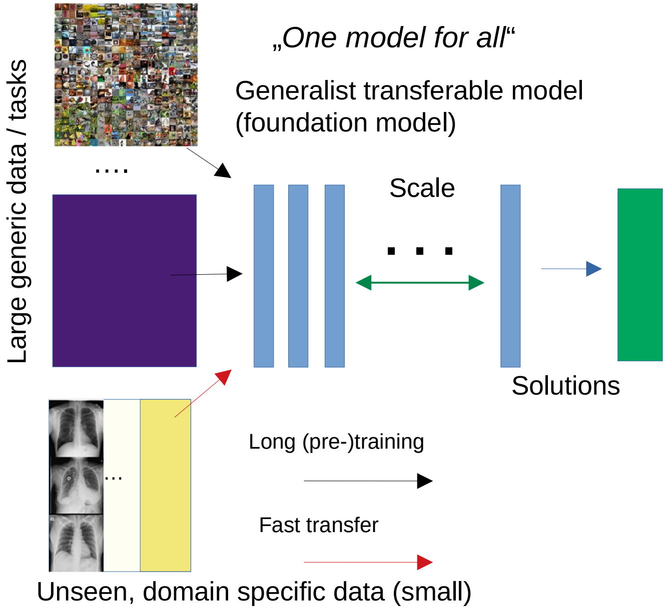
- Scaling law: predicting training/model properties and function
- Predictions are accurate if scaling law derivation is done properly



Bi et al, DeepSeek, 2024

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**

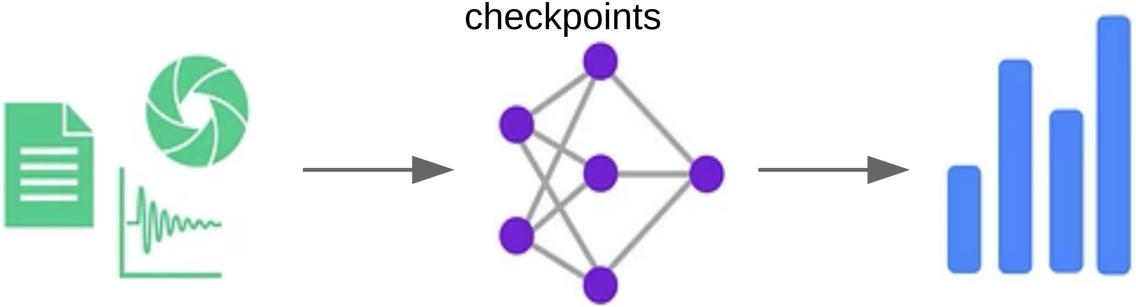
Open-source foundation models & datasets

- 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

OpenCLIP,
openFlamingo,
DCLM
OpenThinker

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

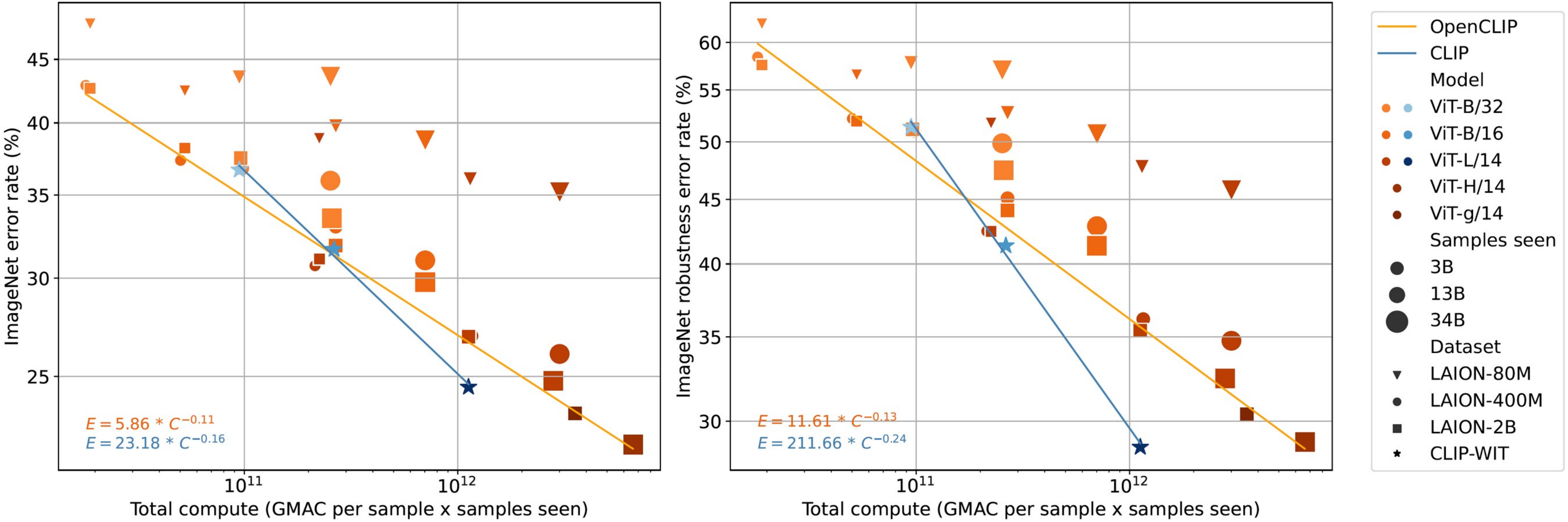
https://github.com/mlfoundations/datacomp/

https://github.com/mlfoundations/open_clip

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

Reproducible scaling laws for foundation models

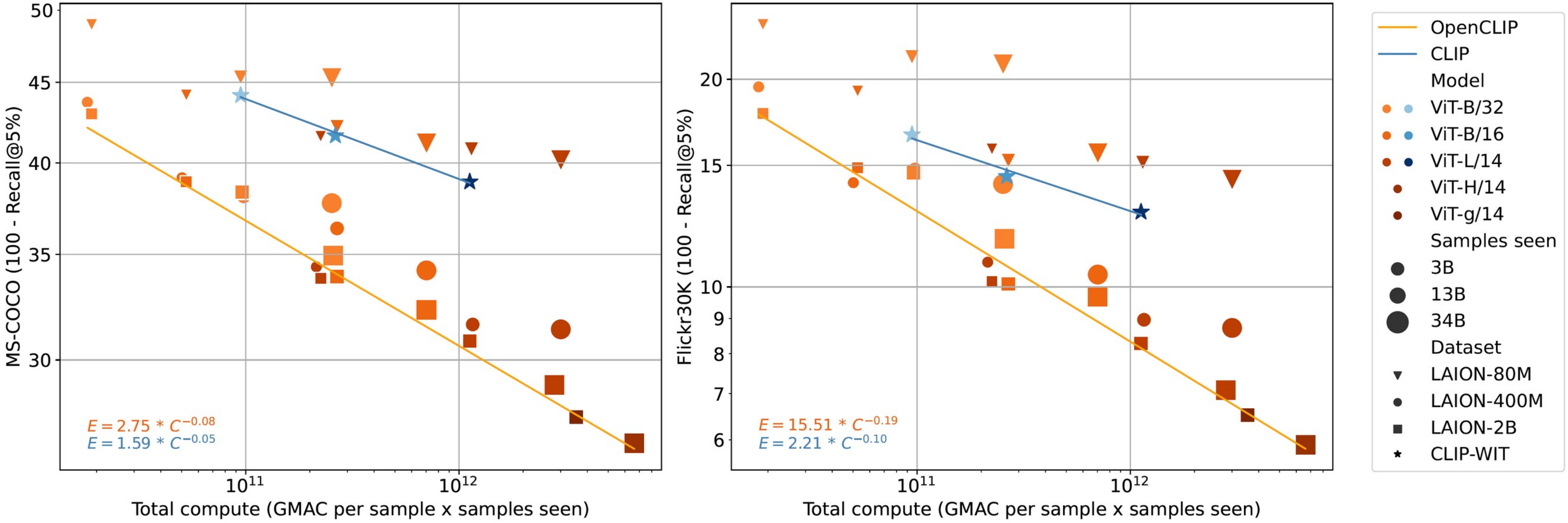
- Scaling laws with LAION-400M/2B and openCLIP: open-source data, models and code - **reproducible** science of foundation models
- Below: zero-shot image classification, ImageNet-1k & robustness sets



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

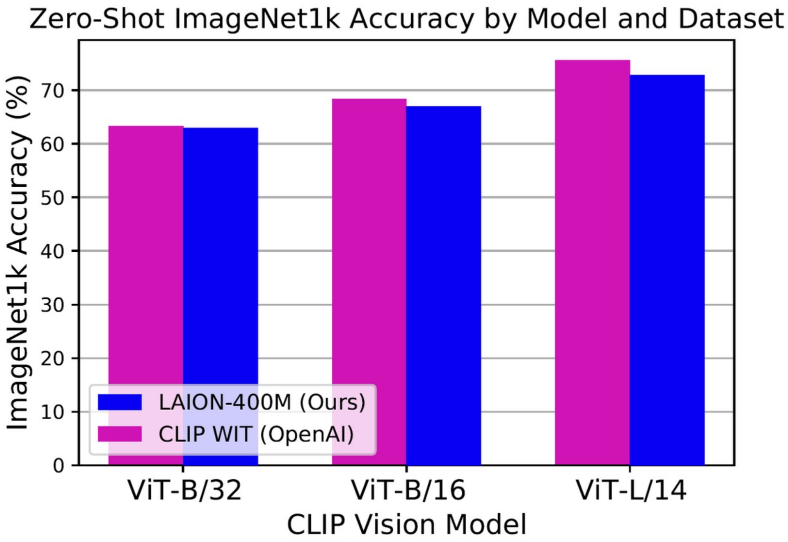


Schuhmann et al, NeurIPS, 2022; Cherti et al, CVPR 2023

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

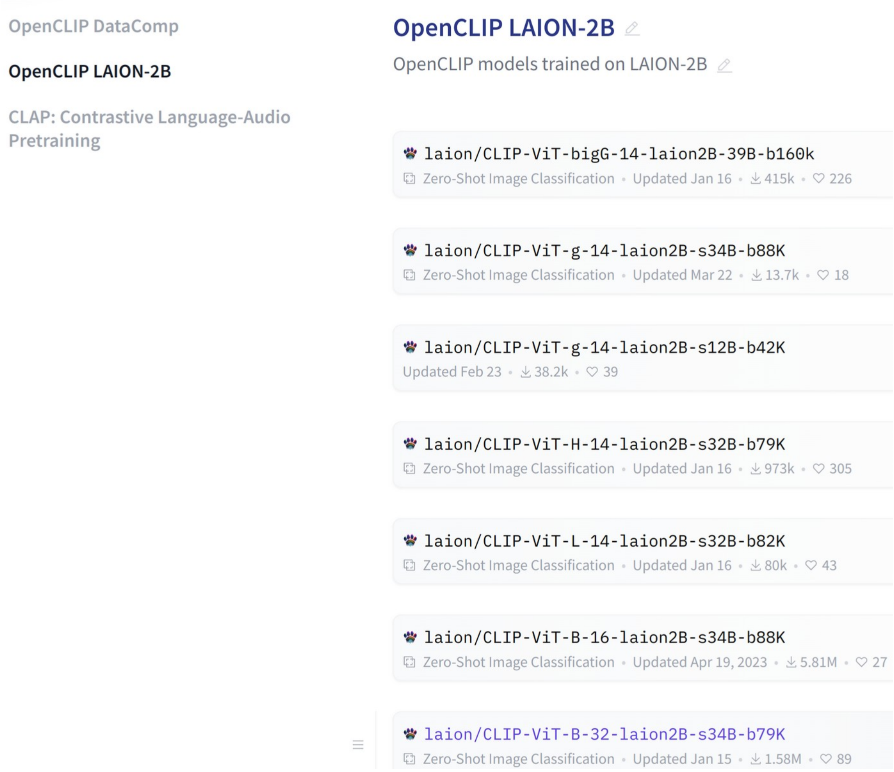


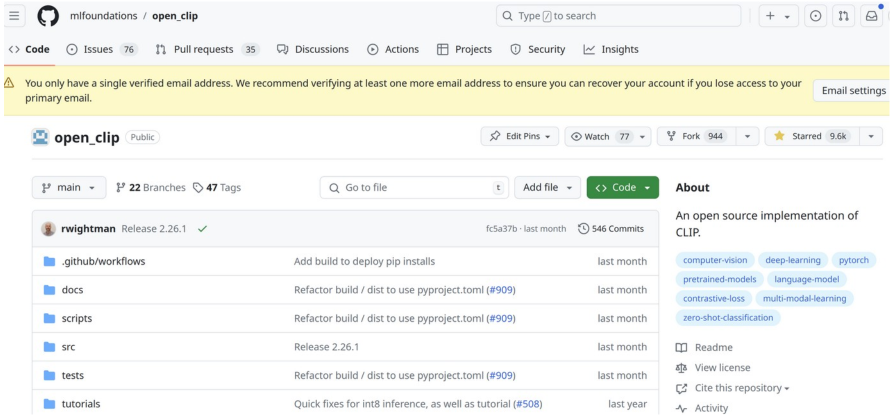
Dataset	# English Img-Txt Pairs
Public Datasets	
MS-COCO	330K
CC3M	3M
Visual Genome	5.4M
WIT	5.5M
CC12M	12M
RedCaps	12M
YFCC100M	100M[2]
LAION-5B (Ours)	**2.3B**
Private Datasets	
CLIP WIT (OpenAI)	400M
ALIGN	1.8B
BASIC	6.6B

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)

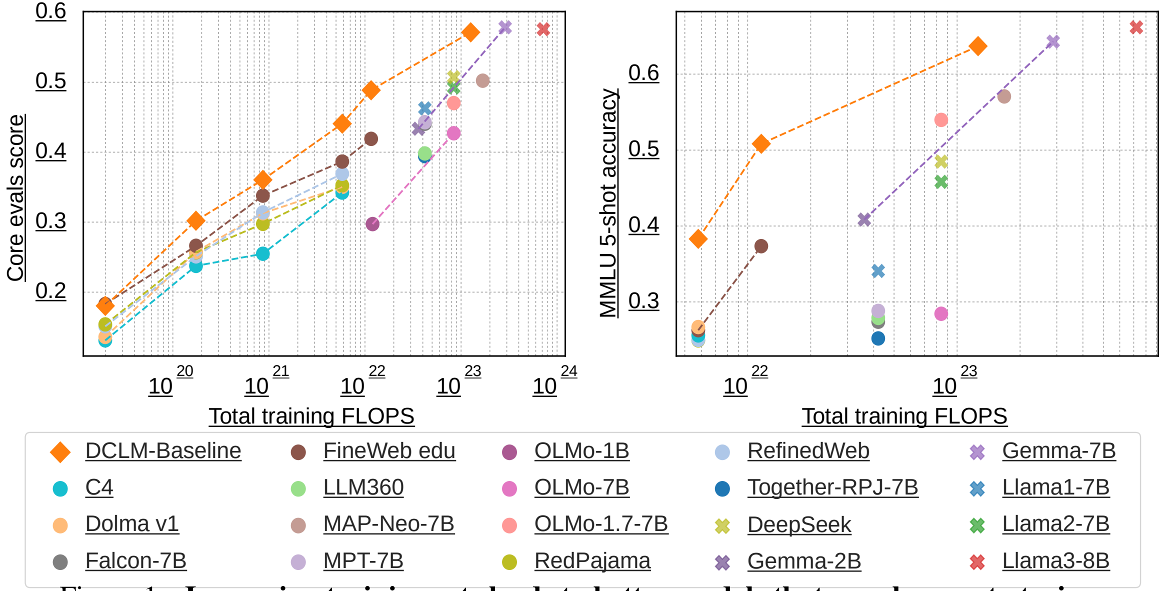


Figure 1: **Improving training sets leads to better models that are cheaper to train.**

Li et al, ArXiv:2406.11794, NeurIPS, 2024

https://github.com/mlfoundations/dclm

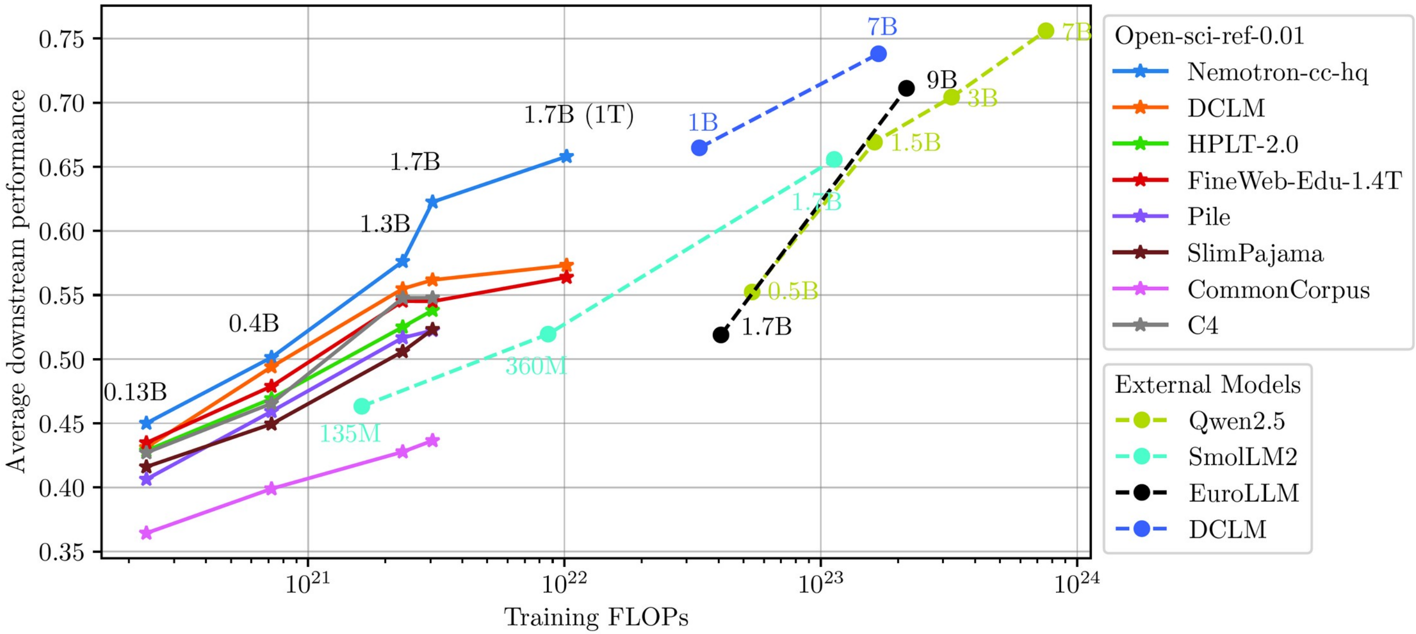
Open foundation models & datasets

- 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 : comparison on aligned compute



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

LAION

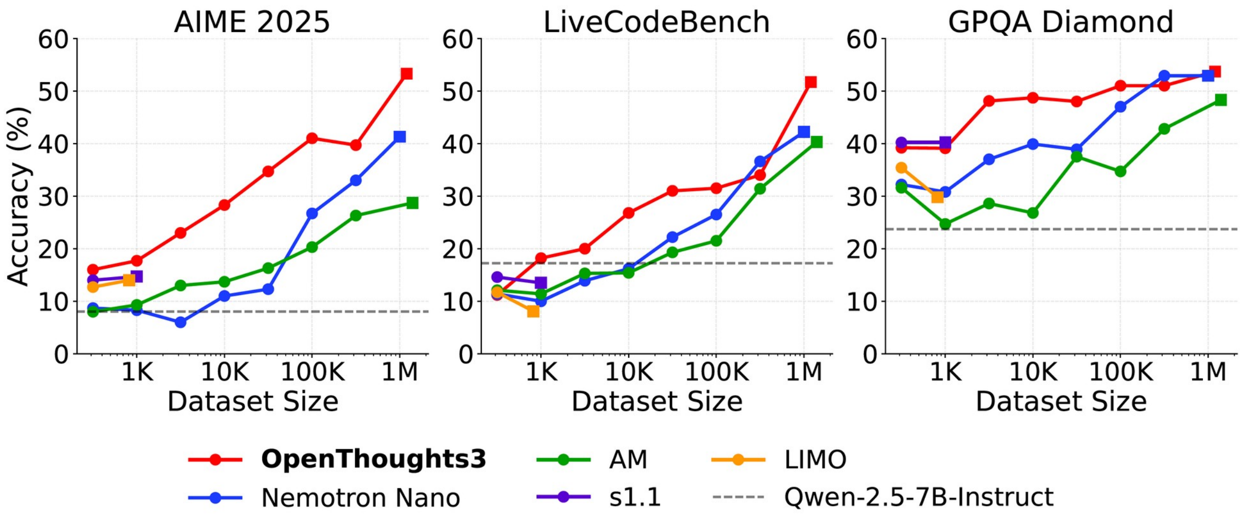
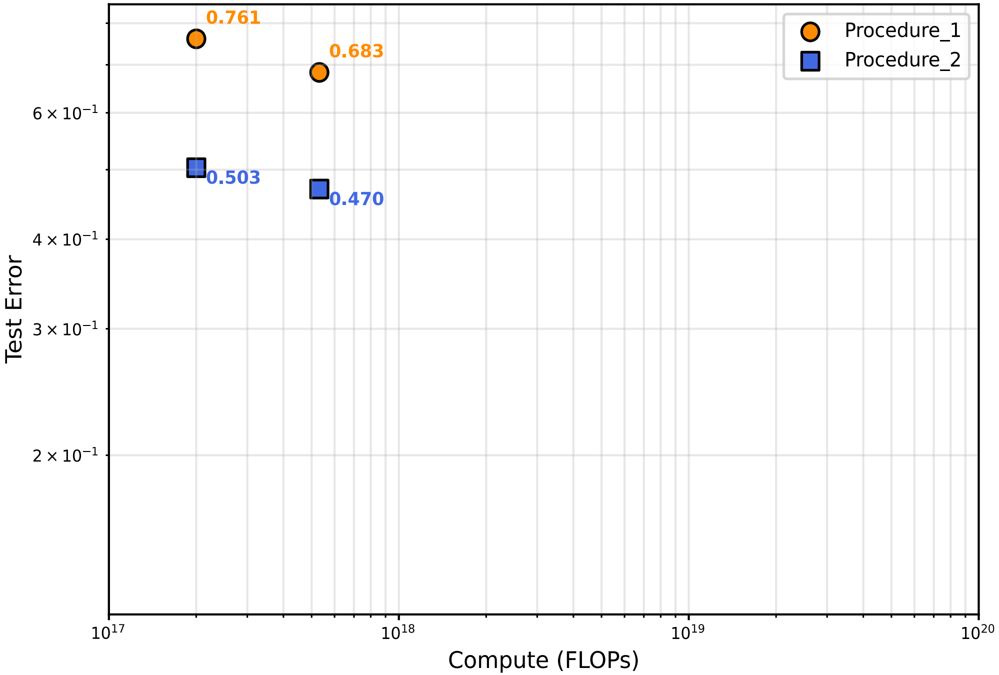


Figure 1: **OpenThoughts3 outperforms existing SFT reasoning datasets across data scales.** All models are finetuned from Qwen-2.5-7B-Instruct. We compare to large SFT datasets (AM, Nemotron Nano) and small curated datasets (s1.1, LIMO) on AIME 2025 (left), LiveCodeBench 06/24-01/25 (middle), and GPQA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

Guha et al, ArXiv:2506.04178, 2025

Improving foundation models: comparison

- Is blue procedure to be preferred over orange procedure?
- Assume already handled: hyperparam tuning, combination model scale / samples scale, aligned on compute, ...

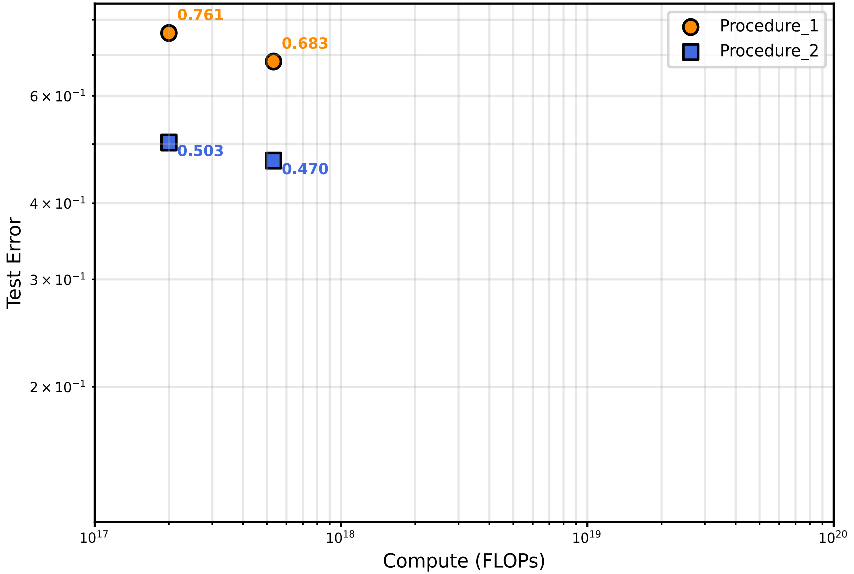


Procedure	Compute (FLOPs)	Test Error
Procedure 1	2.00e+17	0.761
	5.32e+17	0.683
Procedure 2	2.00e+17	0.503
	5.32e+17	0.470

Improving foundation models: comparison

- How to determine which learning procedure leads to better foundation models: what interventions matter, which procedure is worth scaling up?

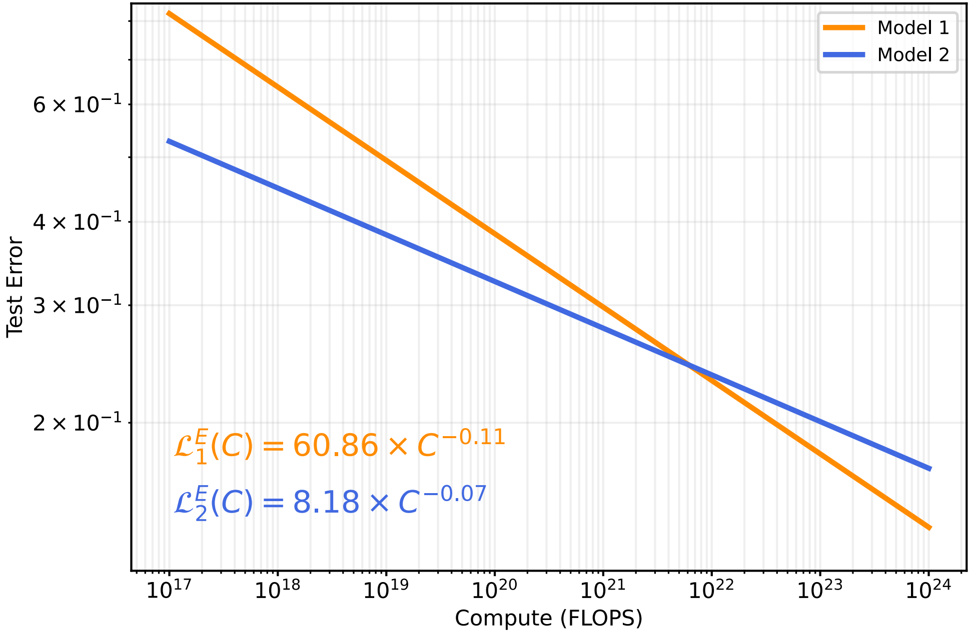
Procedure	Arch	Params	Dataset	Samples Seen	Compute (FLOPs)	Test Error
Procedure 1	Arch 1	0.5B	Dataset X	30M	2.00e+17	0.761
	Arch 1	0.5B	Dataset X	80M	5.32e+17	0.683
Procedure 2	Arch 2	0.5B	Dataset X	30M	2.00e+17	0.503
	Arch 2	0.5B	Dataset X	80M	5.32e+17	0.470



Scaling laws: learning procedure comparison

- Comparison using single isolated points can be highly misleading

Procedure	Arch	Params	Dataset	Samples Seen	Compute (FLOPs)	Test Error
Procedure 1	Arch 1	0.5B	Dataset X	30M	2.00e+17	0.761
	Arch 1	0.5B	Dataset X	80M	5.32e+17	0.683
Procedure 2	Arch 2	0.5B	Dataset X	30M	2.00e+17	0.503
	Arch 2	0.5B	Dataset X	80M	5.32e+17	0.470

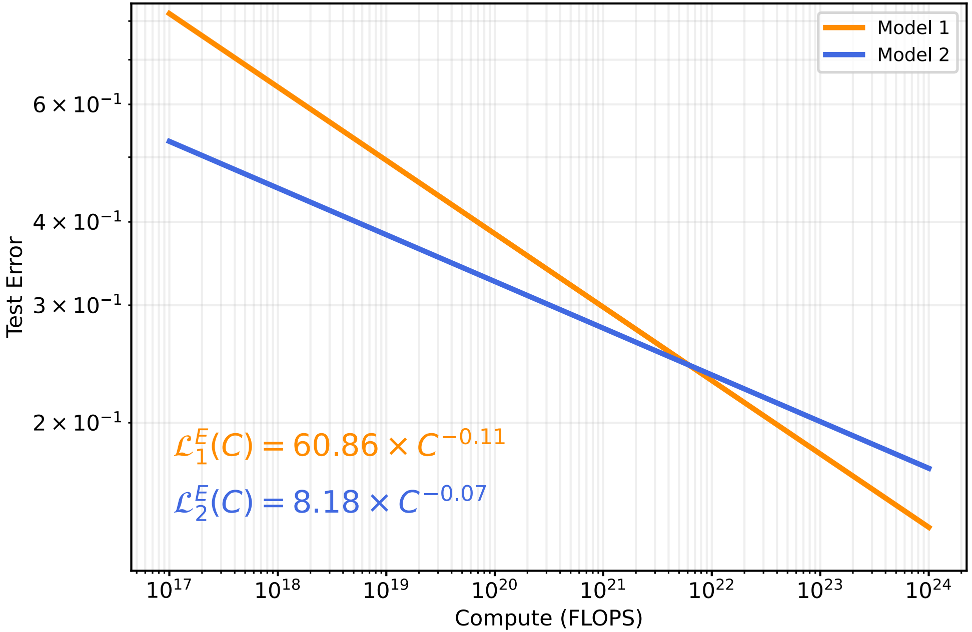


$$\mathcal{L}(C) = C_c \cdot C^{-\alpha_C} + L_\epsilon$$

Scaling laws: learning procedure comparison

- Comparison requires scaling law derivation using standardized open procedures
  - measuring scaling span instead a single reference point, predicting scaling up
  - 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 procedure / model arch
  - ...

Scaling laws: learning procedure comparison

- Comparing foundation models/datasets via scaling law derivation using open pipelines (CLIP vs. MaMMUT; open datasets - DataComp, Re-LAION, DFN)

Scaling Laws for Robust Comparison of Open Foundation Language-Vision Models and Datasets

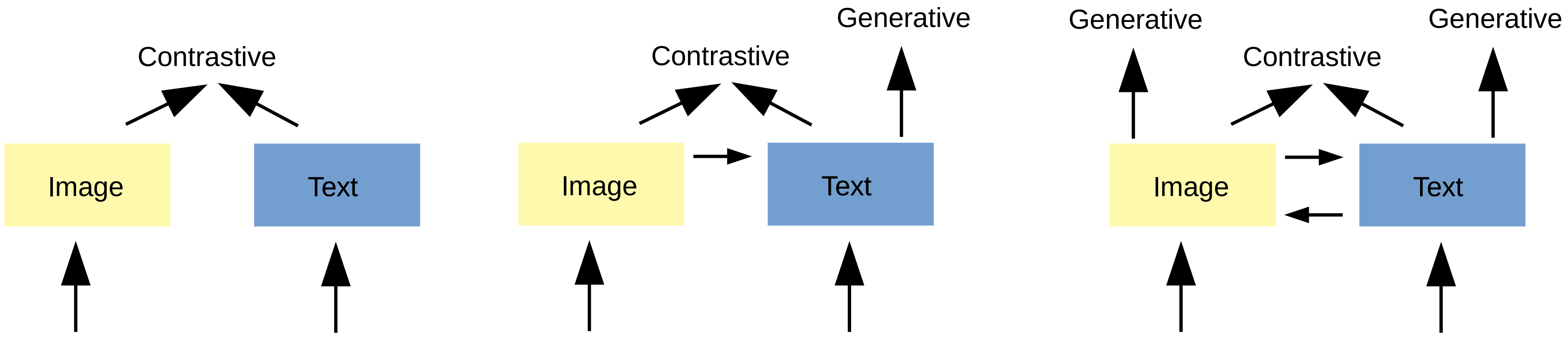
Marianna Nezhurina[1,2,5§*] **Tomer Porian**[1,2,5*] **Giovanni Pucceti**[3] **Tommie Kerssies**[1,4]
Romain Beaumont[1] **Mehdi Cherti**[1,2,5§°*] **Jenia Jitsev**[1,2,5°*]

[1]LAION [2]Juelich Supercomputing Center (JSC), Research Center Juelich (FZJ)
[3] Institute of Information Science and Technologies “A. Faedo” - CNR Pisa
[4] Eindhoven University of Technology
[5] Open-Ψ (Open-Sci) Collective

Nezhurina et al, ArXiv:2506.04598, 2025 (to appear in NeurIPS)

Search for stronger scalable foundation models

- Re-LAION, DataComp & DFN: improving datasets for pre-training
- OpenCLIP extensions: improving learning procedure
  - extend for text & image generative losses (CoCa, Mammut)
  - what loss mix might have stronger scaling?



Standard CLIP

Pure Contrastive

CoCa, Mammut

Contrastive + Text Generative

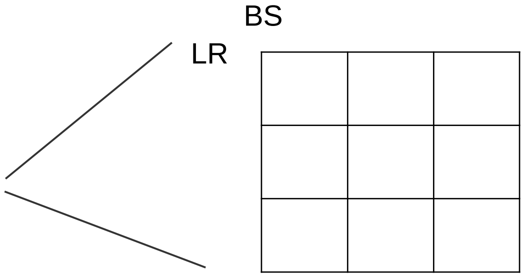
CoBIT, SyCoCa

Contrastive + + Image, Text Generative

Scaling laws: learning procedure comparison

Model N / Data (D)	S/32 (63M)	M/32 (103M)	S/16 (63M)	B/32 (151M)	B/16 (210M)	L/14 (427M)	H/14 (986M)
1.28M	L(C)	L(C)	L(C)	L(C)	L(C)	L(C)	L(C)
3.07M	L(C)	L(C)	L(C)	L(C)	L(C)	L(C)	L(C)
6.4M	L(C)	L(C)	L(C)	L(C)	L(C)	L(C)	L(C)
12.8M	L(C)	L(C)	L(C)	L(C)	L(C)	L(C)	L(C)
30.7M	L(C)	L(C)	L(C)	L(C)	L(C)	L(C)	L(C)
64M	L(C)	L(C)	L(C)	L(C)	L(C)	L(C)	L(C)
128M	L(C)	L(C)	L(C)	L(C)	L(C)	L(C)	L(C)
307M	L(C)	L(C)	L(C)	L(C)	L(C)	L(C)	L(C)
640M	L(C)	L(C)	L(C)	L(C)	L(C)	L(C)	L(C)
1.28B	L(C)	L(C)	L(C)	L(C)	L(C)	L(C)	L(C)
3.07B	L(C)	L(C)	L(C)	L(C)	L(C)	L(C)	L(C)

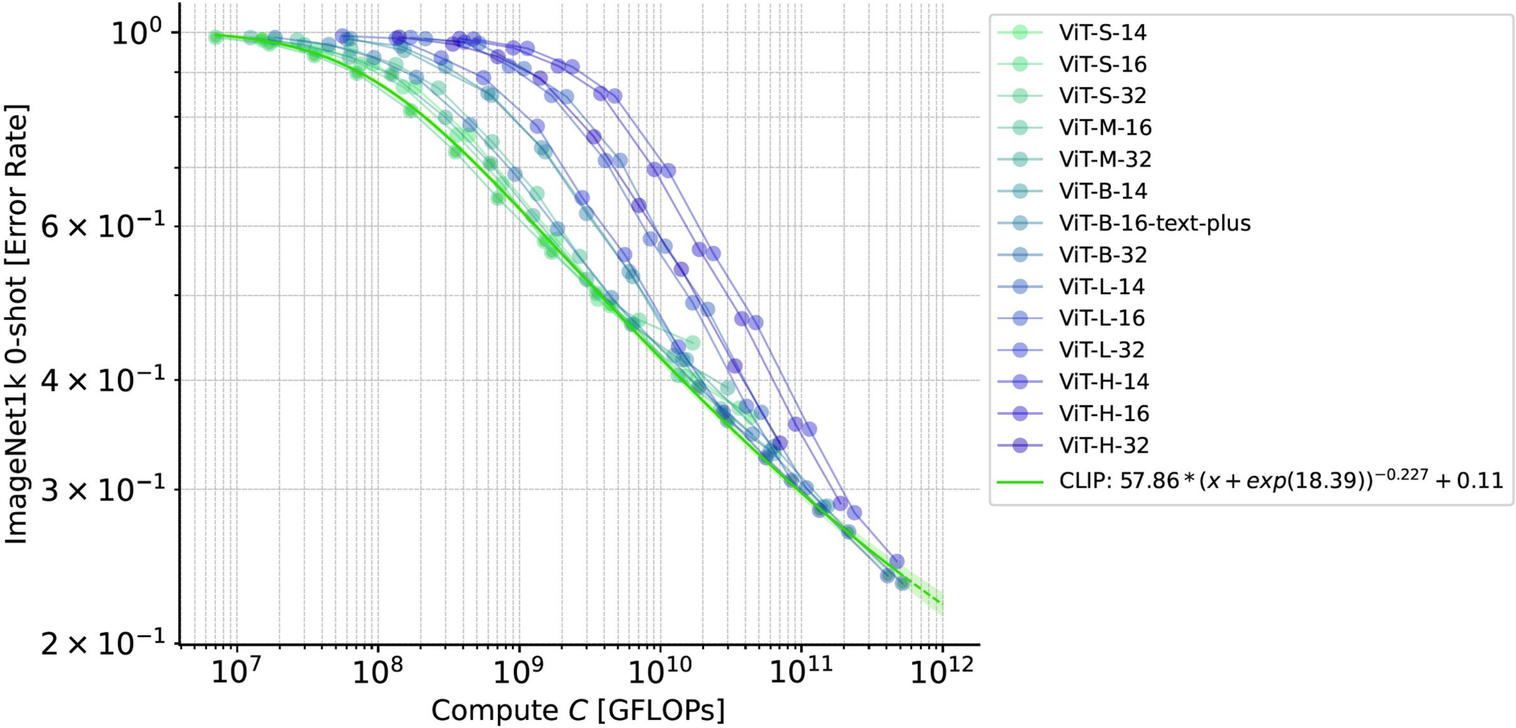
L(C) = **min** L(C,LR,BS)



Nezhurina et al, ArXiv:2506.04598, 2025 (to appear in NeurIPS)

Scaling laws: learning procedure comparison

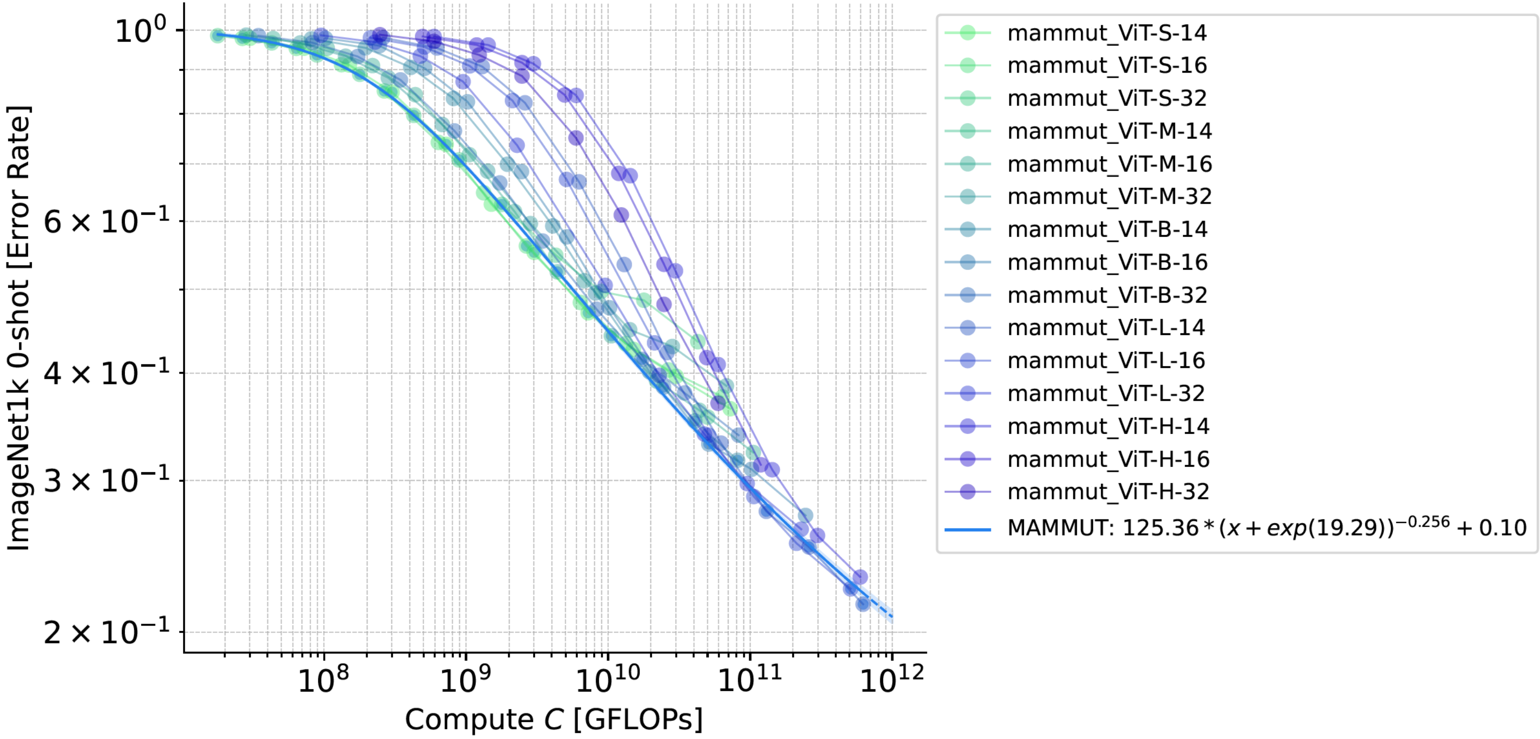
- Scaling law derivation on dense measurements: CLIP (data: DataComp-1b)



Nezhurina et al, ArXiv:2506.04598, 2025 (to appear in NeurIPS)

Scaling laws: learning procedure comparison

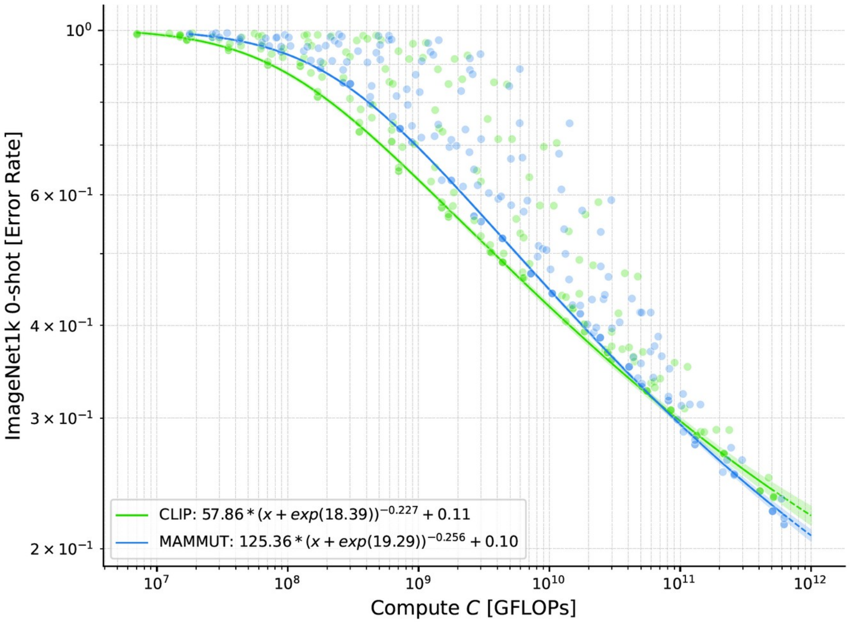
- Scaling law derivation on dense measurements: MaMMUT (data: DataComp-1b)



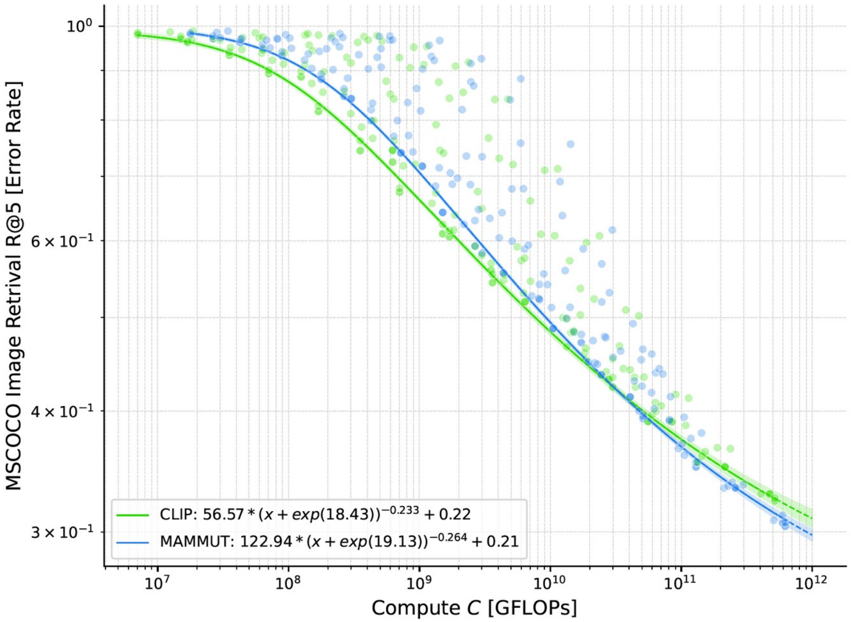
Nezhurina et al, ArXiv:2506.04598, 2025 (to appear in NeurIPS)

Scaling laws: learning procedure comparison

- Comparing CLIP vs. MaMMUT (dataset: DataComp-1.4B)



(a) ImageNet-1k 0-shot classification



(b) MS-COCO image R@5

Nezhurina et al, ArXiv:2506.04598, 2025 (to appear in NeurIPS)

Scaling laws: learning procedure comparison

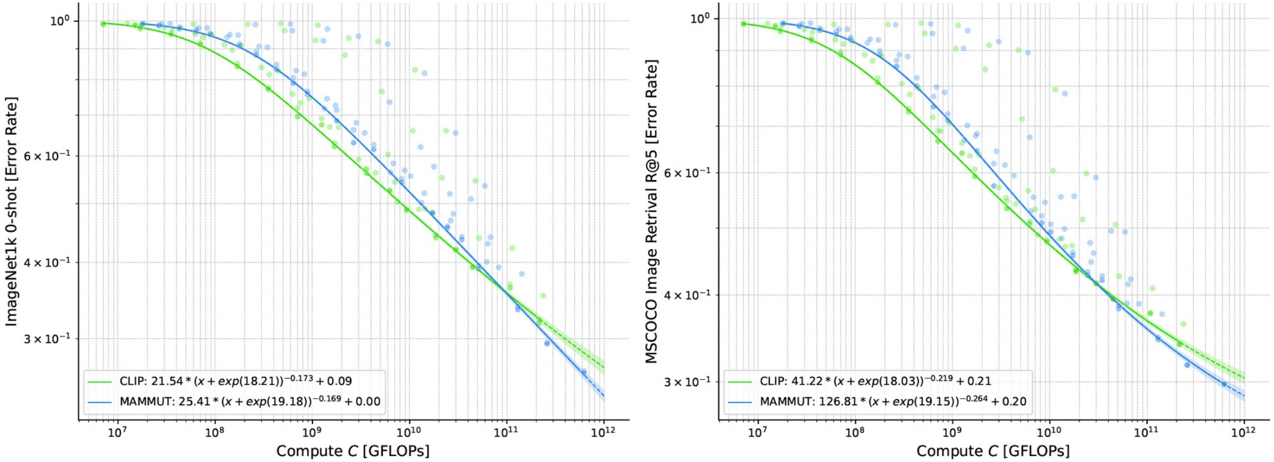
- Checking scaling law fit quality: validating scaling law predictions on held-out points close to compute optimal Pareto front

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	0.798	0.797 (0.787, 0.807)	0.801 (0.793, 0.809)
RMSE: 7.57e-03	*RMSE (more points): 7.57e-03*				

Table C: Predictions for different values of $C_{\text{threshold}} = \{2.5 \cdot 10^{11}, 5 \cdot 10^{11}\}$ GFLOPS. Scaling law derivation on DataComp-1.4B. The last column shows updated predictions made after adding more data points. Both confidence interval and RMSE decrease as we take more points.

Scaling laws: learning procedure comparison

- Comparison via scaling law derivation: consistency across various scenarios



(a) ImageNet-1k 0-shot classification

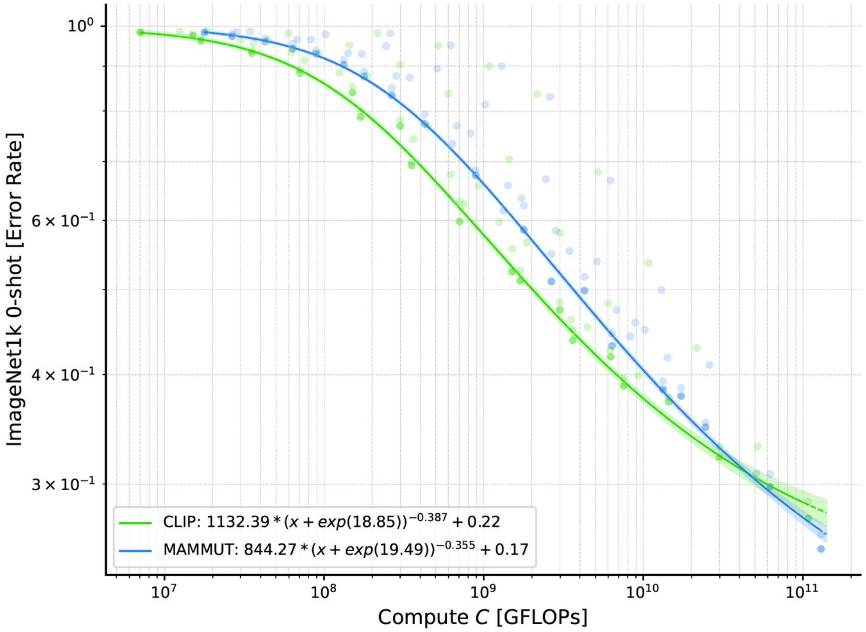
(b) MS-COCO image R@5

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.

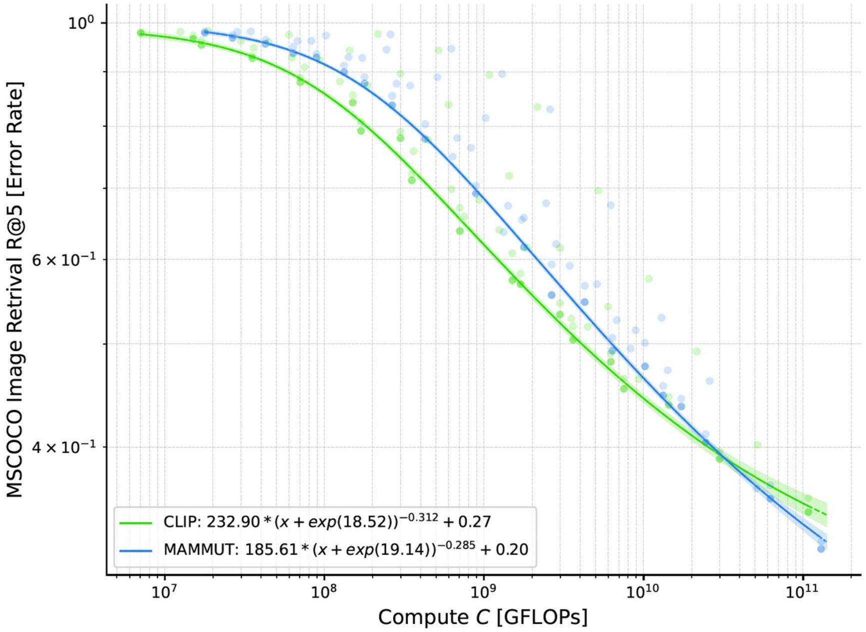
Nezhurina et al, ArXiv:2506.04598, 2025 (to appear in NeurIPS)

Scaling laws: learning procedure comparison

- Comparison via scaling law derivation: consistency across various scenarios (dataset: DFN-1.4B)



(a) ImageNet-1k 0-shot classification

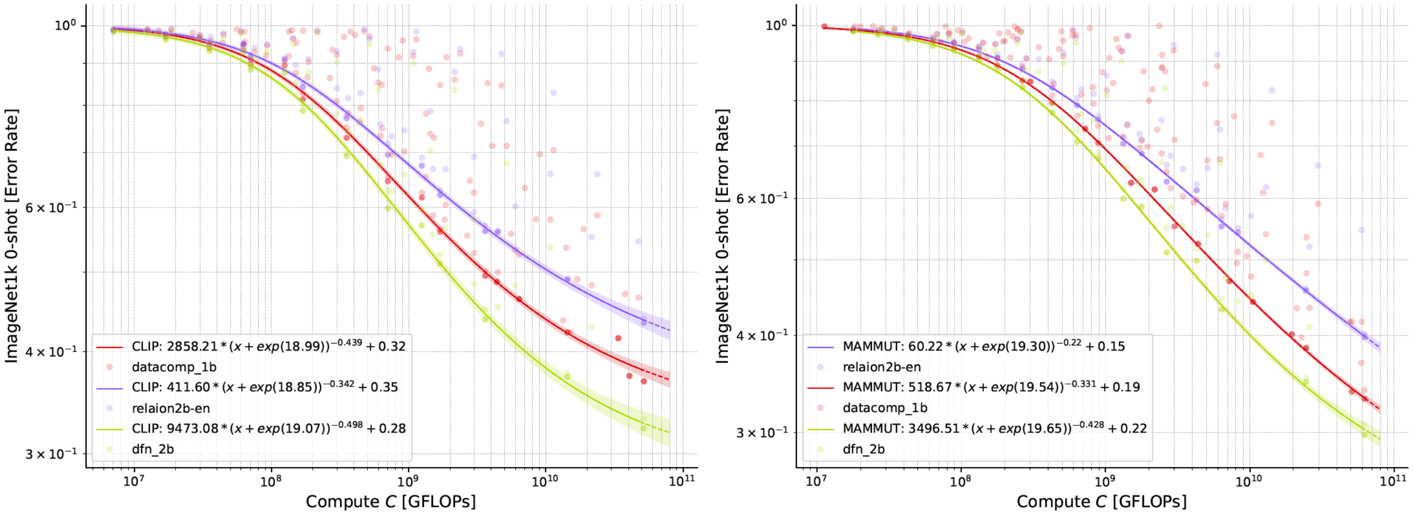


(b) MS-COCO image R@5

Nezhurina et al, ArXiv:2506.04598, 2025 (to appear in NeurIPS)

Scaling laws: learning procedure comparison

- Comparison via scaling law derivation: open dataset comparison



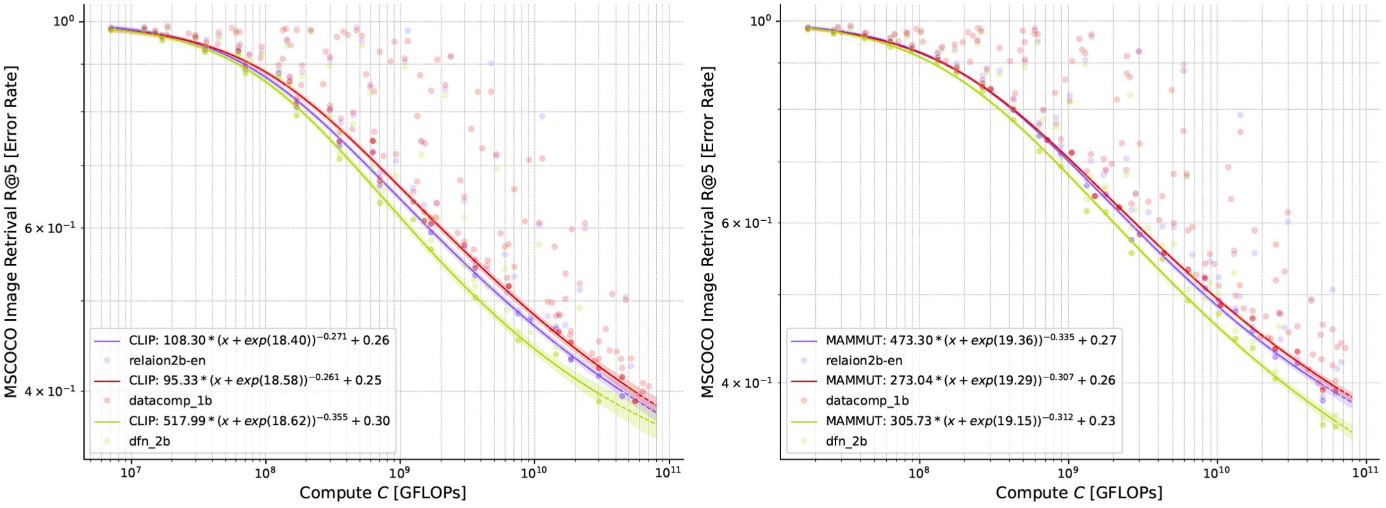
(a) IN-1k 0-shot error rate for openCLIP (b) IN-1k 0-shot error rate for openMaMMUT

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 (to appear in NeurIPS)

Scaling laws: learning procedure comparison

- Comparison via scaling law derivation: open dataset comparison



(a) Error Rate for CLIP

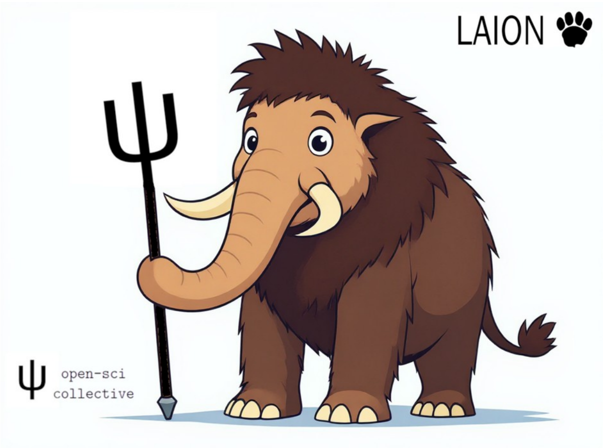
(b) Error Rate for MaMMUT

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 (to appear in NeurIPS)

Open foundation models with stronger scalability

- LAION as open frontiers lab: openMaMMUT predictably matching or outperforming SOTA of closed labs
- Scaling law based comparison: predicting whether an experimental procedure is worth scaling up, leading to stronger models than an already existing reference



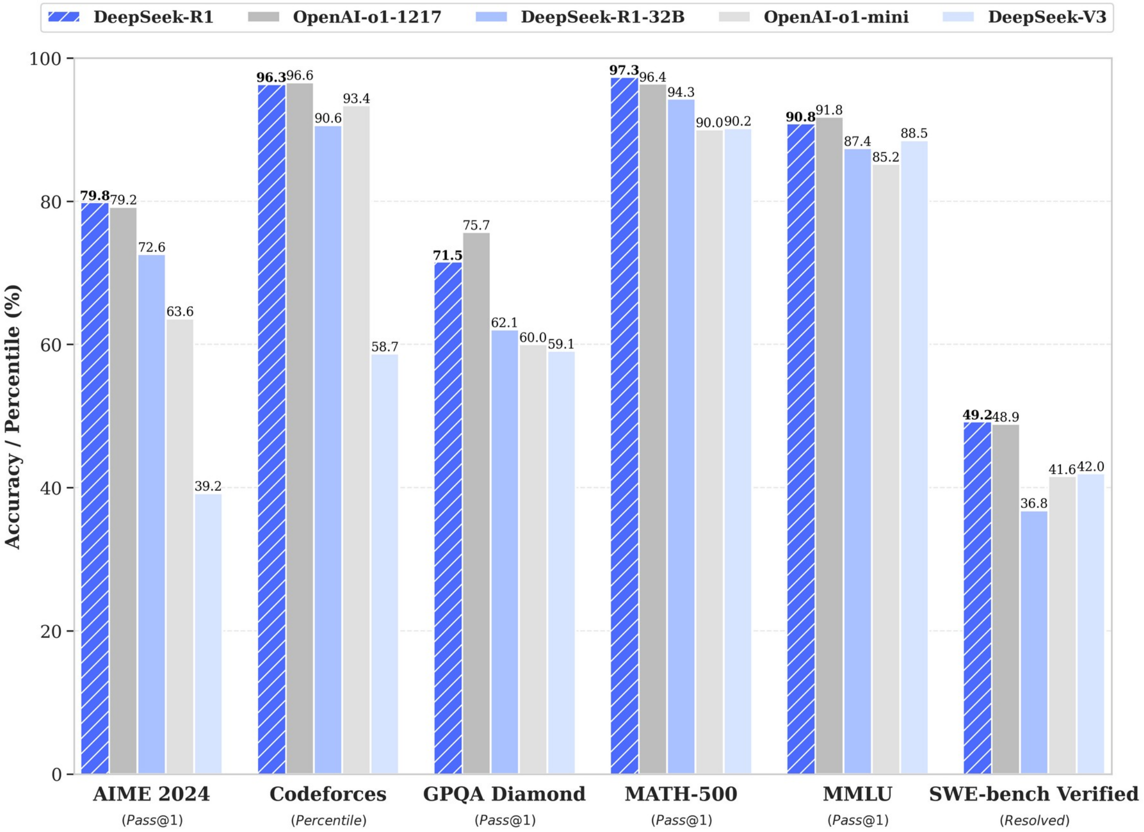
ViT	Res.	Seq.	Model	Dataset	#Samples	ImageNet-1k val	ImageNet-1k v2	COCO T→I	COCO I→T
L/16	256	256	SigLIP [18]	WebLI-10B	40B	80.44	73.76	75.26	88.40
			SigLIP 2 [14]	WebLI-10B	40B	82.35	76.66	76.84	90.44
L/14	224	256	OpenCLIP [10]	LAION-2B	34B	75.24	67.73	70.46	84.30
			CLIP [7]	WIT-400M	12.8B	75.54	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 (to appear in NeurIPS)

Scaling laws: predicting generalization

- Do standardized benchmarks downstream tasks reflect generalization properly?



DeepSeek R1 Technical Report, DeepSeek, 2025

Scaling laws: predicting generalization

- Do benchmark downstream tasks reflect generalization properly?
- Test set leakage, training data contamination: how to test generalization?
- Using variations of simple problem templates to measure model robustness

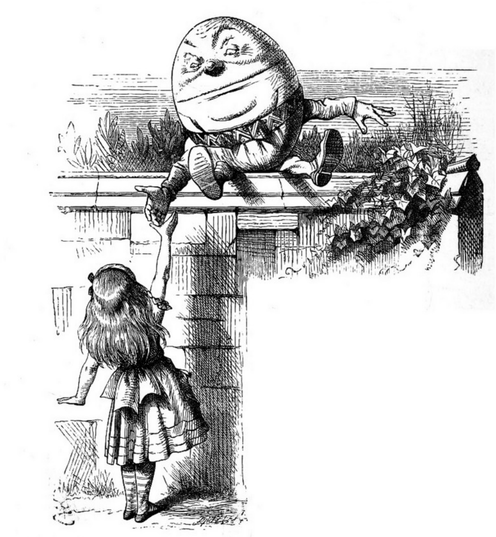


Figure 1: Alice is reasoning: will it break? Illustration of Humpty Dumpty from Through the Looking Glass, by John Tenniel, 1871. Source: Wikipedia.

AIW Original, Variations 1-6. Prompt IDs 264 266 268 270 455 456

Variation 1: Alice has **3 brothers** and she also has **6 sisters**. [Correct answer: **7**]
Variation 2: Alice has **2 sisters** and she also has **4 brothers**. [Correct answer: **3**]
Variation 3: Alice has **4 sisters** and she also has **1 brother**. [Correct answer: **5**]
Variation 4: Alice has **4 brothers** and she also has **1 sister**. [Correct answer: **2**]
Variation 5: Alice has **2 brothers** and she also has **3 sisters**. [Correct answer: **4**]
Variation 6: Alice has **5 sisters** and she also has **3 brothers**. [Correct answer: **6**]

How many sisters does Alice's brother have?

LAION

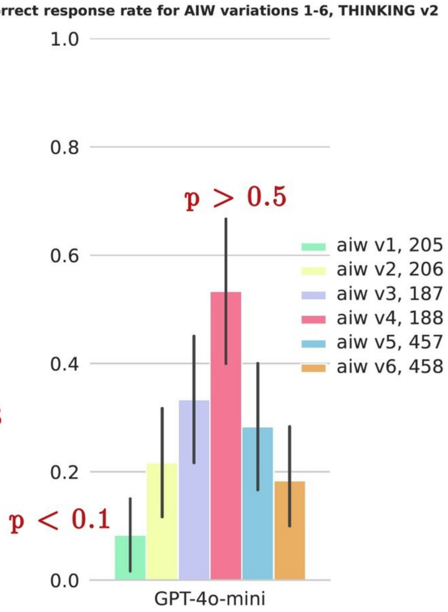
open-sci collective

Nezhurina et al, ArXiv:2406.02061, NeurIPS SciForDL, 2024

Generalization: measuring it right

- SOTA LLMs show strong fluctuations across variations that DO NOT CHANGE problem structure at all

- 60 trials for each AIW variation 1-6
- Measure p, correct response rate, for each AIW variation
- Prompt IDs: 205,206,187,188,457,458



AIW Original, Variations 1-6. Prompt IDs 264 266 268 270 455 456

Variation 1: Alice has **3 brothers** and she also has **6 sisters**. [Correct answer: **7**]
Variation 2: Alice has **2 sisters** and she also has **4 brothers**. [Correct answer: **3**]
Variation 3: Alice has **4 sisters** and she also has **1 brother**. [Correct answer: **5**]
Variation 4: Alice has **4 brothers** and she also has **1 sister**. [Correct answer: **2**]
Variation 5: Alice has **2 brothers** and she also has **3 sisters**. [Correct answer: **4**]
Variation 6: Alice has **5 sisters** and she also has **3 brothers**. [Correct answer: **6**]

How many sisters does Alice's brother have?

Nezhurina et al, ArXiv:2406.02061, NeurIPS SciForDL, 2024

AIW Variations 1-4

Variation 1: Alice has **3 brothers** and she also has **6 sisters**. [Correct answer: **7**]
Variation 2: Alice has **2 sisters** and she also has **4 brothers**. [Correct answer: **3**]
Variation 3: Alice has **4 sisters** and she also has **1 brother**. [Correct answer: **5**]
Variation 4: Alice has **4 brothers** and she also has **1 sister**. [Correct answer: **2**]

How many sisters does Alice's brother have?

AIW Ext Alice and Bob, Alice's Brothers, Variations 1-4

Alice and Bob are sister and brother.

Variation 1: Alice has **3 sisters** and Bob has **6 brothers**. [Correct answer: **7**]
Variation 2: Alice has **2 sisters** and Bob has **2 brothers**. [Correct answer: **3**]
Variation 3: Alice has **1 sister** and Bob has **4 brothers**. [Correct answer: **5**]
Variation 4: Alice has **3 sisters** and Bob has **1 brother**. [Correct answer: **2**]

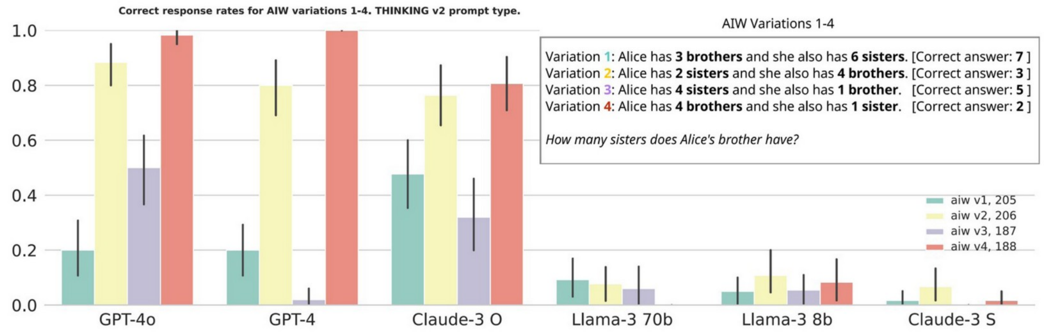
How many brothers does Alice have?

AIW Male and Female Friends, Variations 1-4

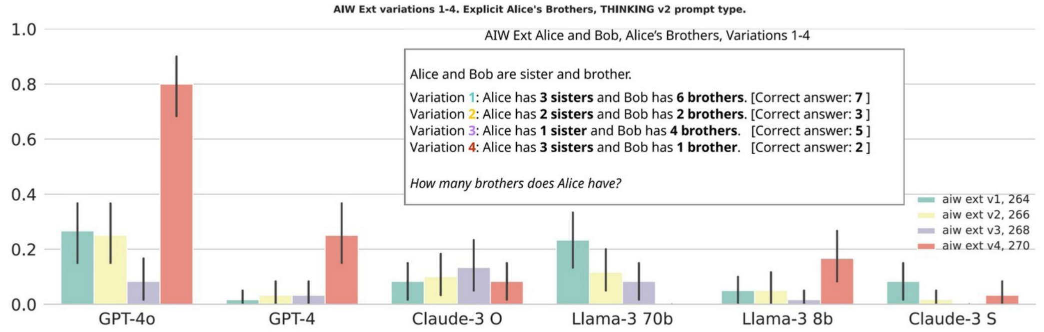
Variation 1: Alice has **3 male friends** and she also has **6 female friends**. [Correct answer: **7**]
Variation 2: Alice has **2 female friends** and she also has **4 male friends**. [Correct answer: **3**]
Variation 3: Alice has **4 female friends** and she also has **1 male friend**. [Correct answer: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

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

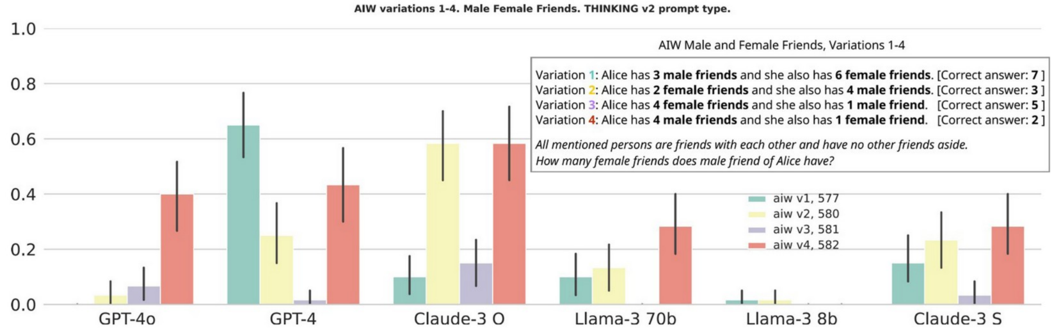
A



B



C



Generalization: measuring it right

- Control problems (AIW Light): ruling out low-level issues

AIW Variations, Original and AIW Light Control

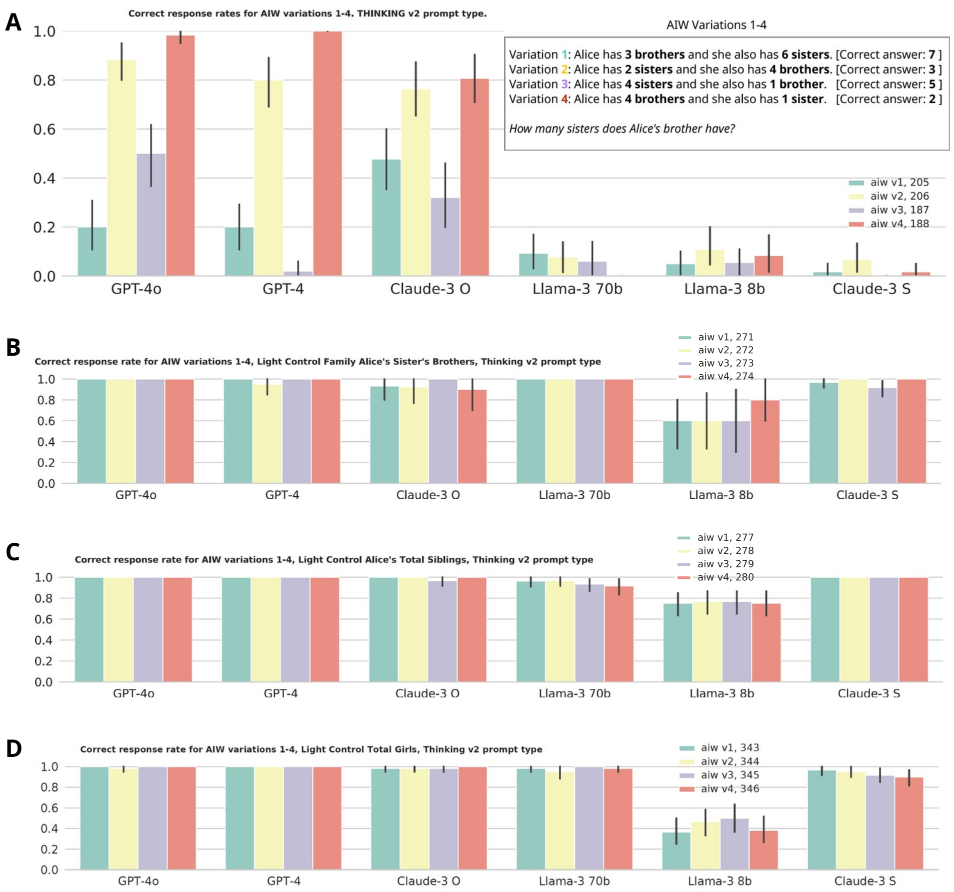
Template: Alice has *N* **brothers** and she also has *M* **sisters**.
Variations 1-4: changing *N*, *M* <= 7. Correct responses: *C* <= 7

<u>AIW Original</u> (SOTA LLM breakdown)
How many sisters does Alice's brother have? [correct: C = M + 1] (**A**)

<u>AIW Light Control</u> (SOTA LLM succeed)
How many brothers does Alice's sister have? [correct: C = N] (**B**)
How many siblings does Alice have? [correct: C = N + M] (**C**)
How many girls are there in total? [correct: C = M + 1] (**D**)

Prompt type

THINKING v2 : Before providing answer to this problem, think carefully step by step and double check the path to the correct solution for any mistakes. Provide then the final answer in following form: "### Answer: ".



Generalization: measuring it right

- Sensitivity to problem variants: revealing training data contamination?

AIW Variations 1-4

Variation 1: Alice has **3 brothers** and she also has **6 sisters**. [Correct answer: **7**]
Variation 2: Alice has **2 sisters** and she also has **4 brothers**. [Correct answer: **3**]
Variation 3: Alice has **4 sisters** and she also has **1 brother**. [Correct answer: **5**]
Variation 4: Alice has **4 brothers** and she also has **1 sister**. [Correct answer: **2**]

How many sisters does Alice's brother have?

AIW Ext Alice and Bob, Alice's Brothers, Variations 1-4

Alice and Bob are sister and brother.

Variation 1: Alice has **3 sisters** and Bob has **6 brothers**. [Correct answer: **7**]
Variation 2: Alice has **2 sisters** and Bob has **2 brothers**. [Correct answer: **3**]
Variation 3: Alice has **1 sister** and Bob has **4 brothers**. [Correct answer: **5**]
Variation 4: Alice has **3 sisters** and Bob has **1 brother**. [Correct answer: **2**]

How many brothers does Alice have?

Nezhurina et al, ArXiv:2406.02061, NeurIPS SciForDL, 2024

Generalization: measuring it right

- Hints on training contamination and generalization deficit: strong performance difference on similar problems

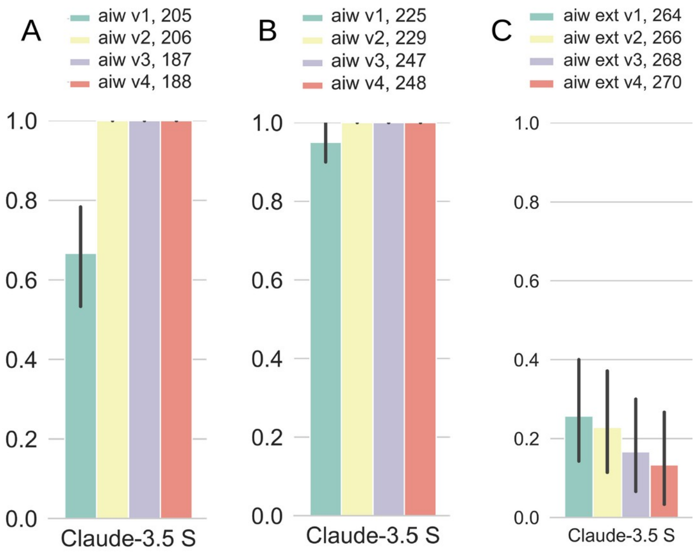
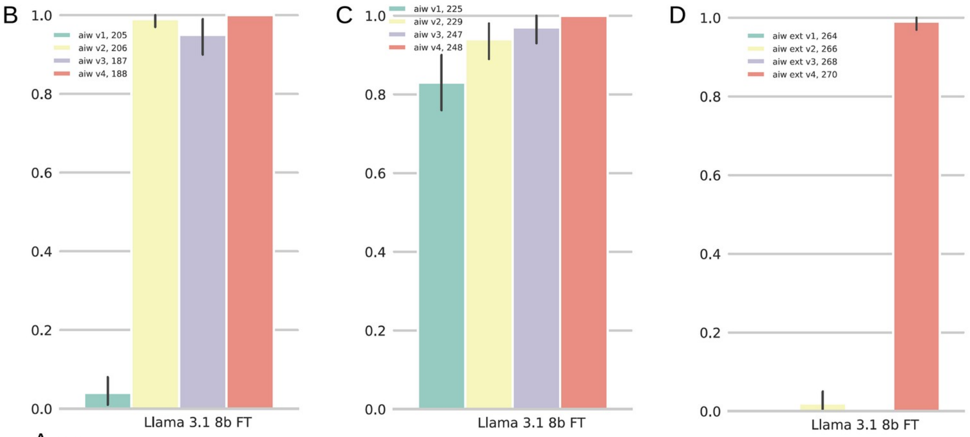
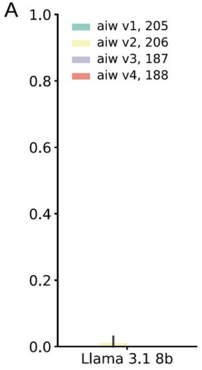


Figure 18: A Tale of Rise and Fall of Claude 3.5 Sonnet. While correct response rates go up close to 1 on (**A**) AIW original and also (**B**) AIW Original Bob version, strong breakdown of correct response rates is observed on AIW extension (**C**) (AIW Ext), accompanied with fluctuations across variations 1-4. Strongly elevated correct respones rates on AIW original might hint on exposure of Claude 3.5 Sonnet to AIW problem data for tuning. Collapse on AIW Ext, which has same problem structure as AIW original, shows though again clearly lack of robustness and hints on same basic reasoning deficits as suspected for other tested models.

Nezhurina et al, ArXiv:2406.02061, NeurIPS SciForDL, 2024

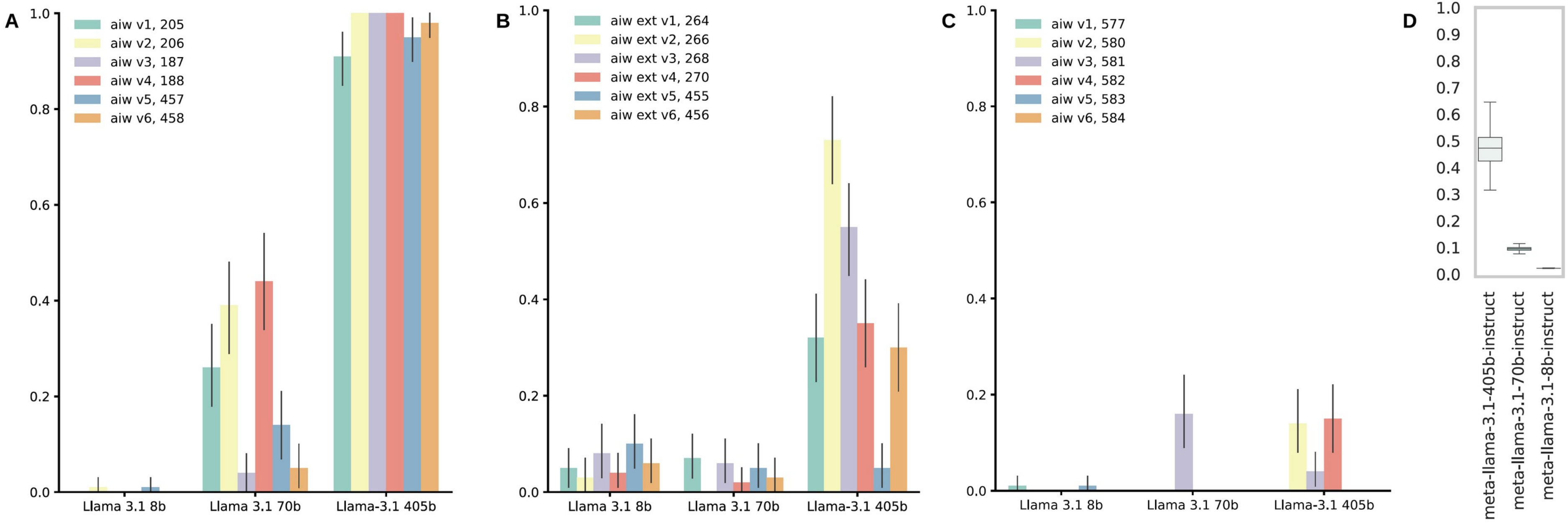




- Fine-tuning of Llama 3.1 8B
- **A**: model fails on AIW original
- **B, C**: model fine-tuned on AIW original performs well on variations
- **D**: fine-tuned model collapses and fluctuates on AIW Ext

Generalization: measuring it right

- Effect of scale: small scale models undergo severe collapse. Larger scale models exhibit strong fluctuations.



Nezhurina et al, ArXiv:2406.02061, NeurIPS SciForDL, 2024

Generalization: measuring it right

- Reasoning models: solve AIW original and AIW ext. How about further AIW versions?

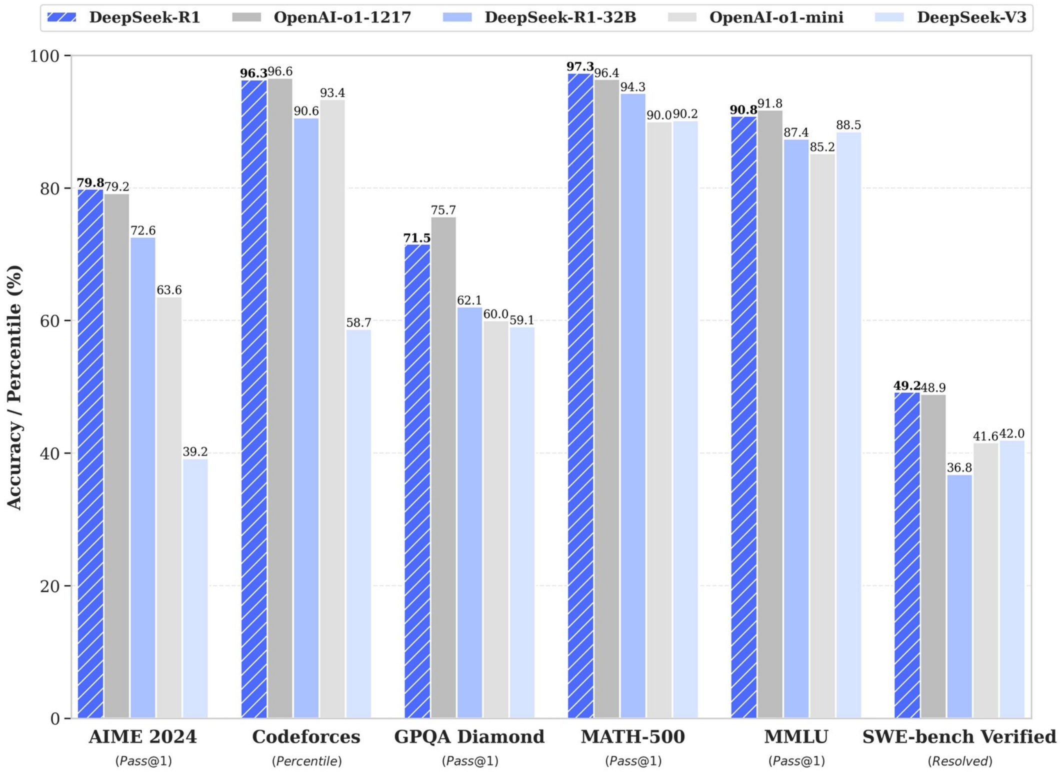
AIW Friends, Variations 1-6, Prompt IDs: 577 580 581 582 583 584

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?

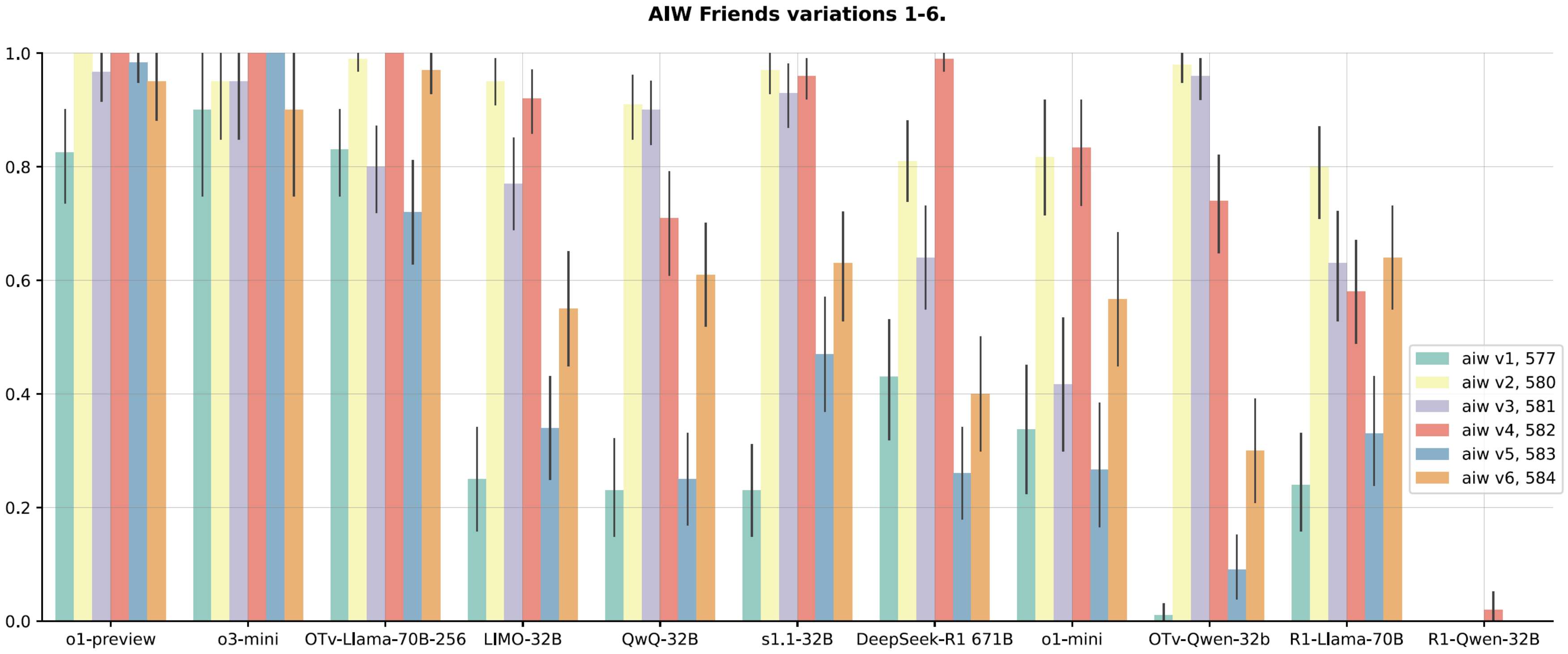
Scaling laws: predicting generalization

- High scores on reasoning benchmarks suggest robust problem solving on graduate or olympiad level.



Scaling laws: predicting generalization

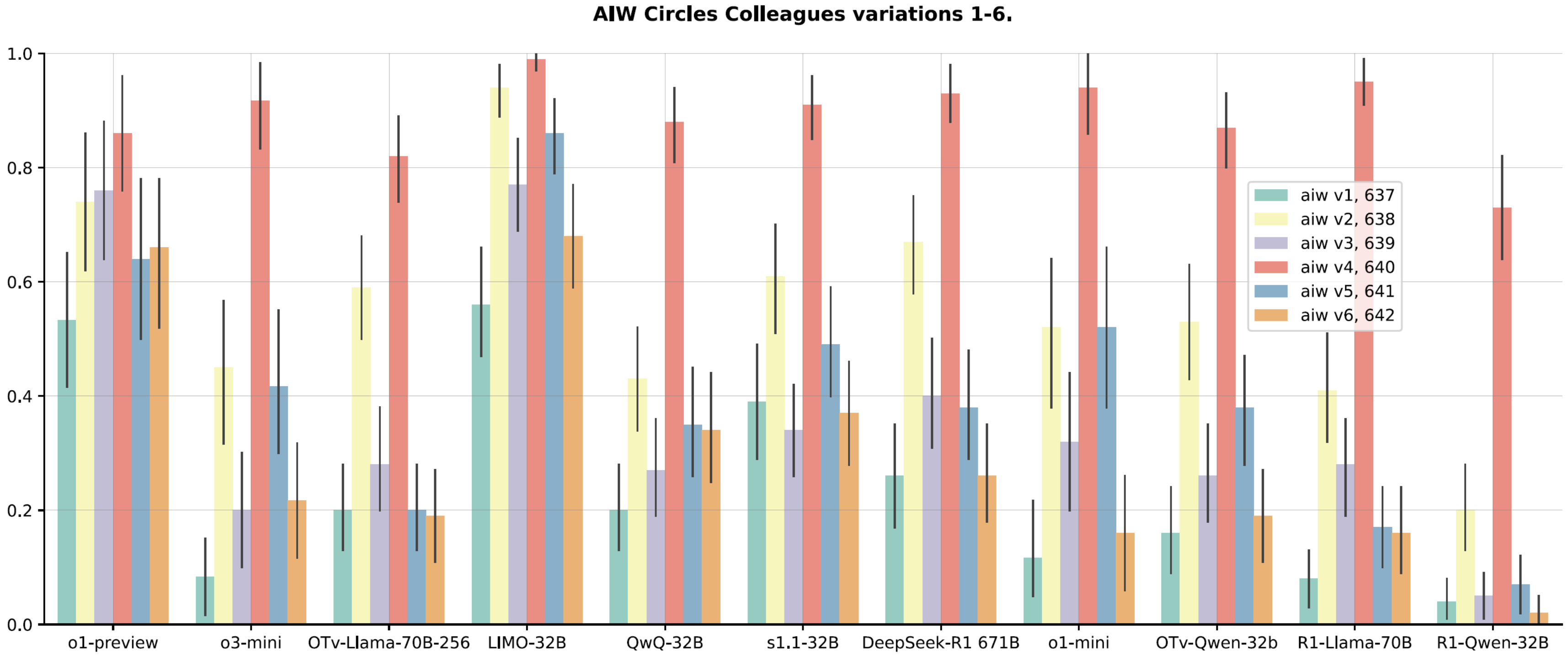
- Reasoning models: Still show strong fluctuations across variations that DO NOT CHANGE problem structure at all



Nezhurina et al, ArXiv:2406.02061, NeurIPS SciForDL, 2024

Scaling laws: predicting generalization

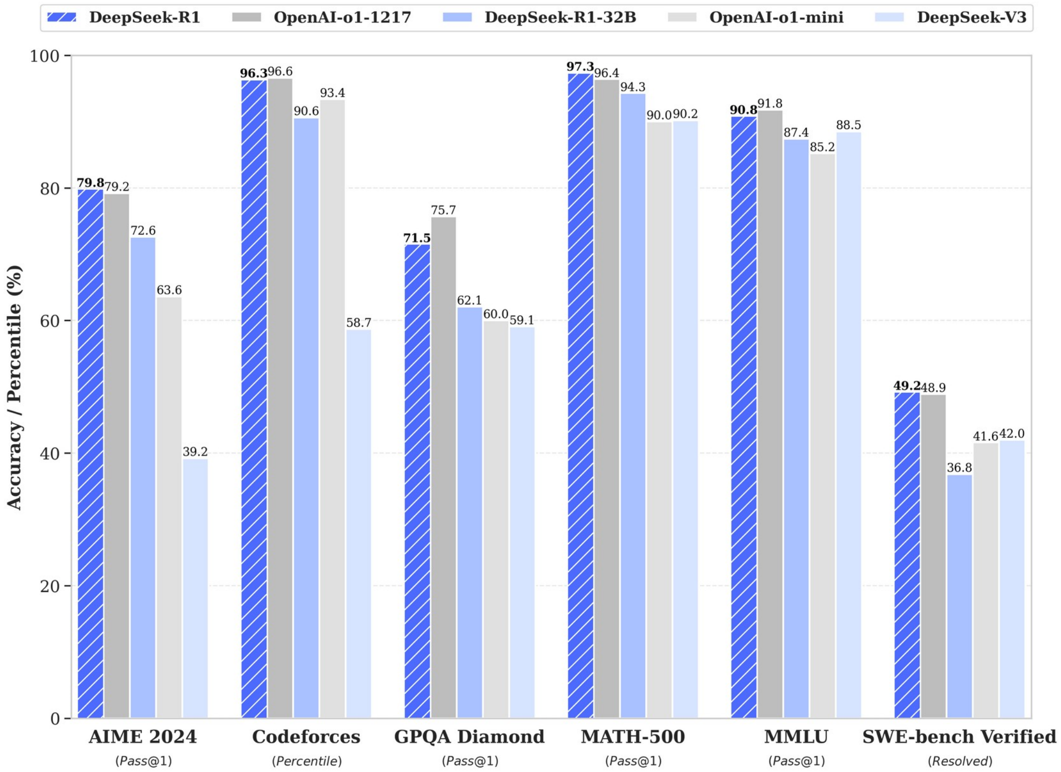
- Reasoning models: Still show strong fluctuations across variations that DO NOT CHANGE problem structure at all



Nezhurina et al, ArXiv:2406.02061, NeurIPS SciForDL, 2024

Scaling laws: predicting generalization

- AIW problems are far below graduate or olympiad level. High scores on reasoning benchmarks are misleading



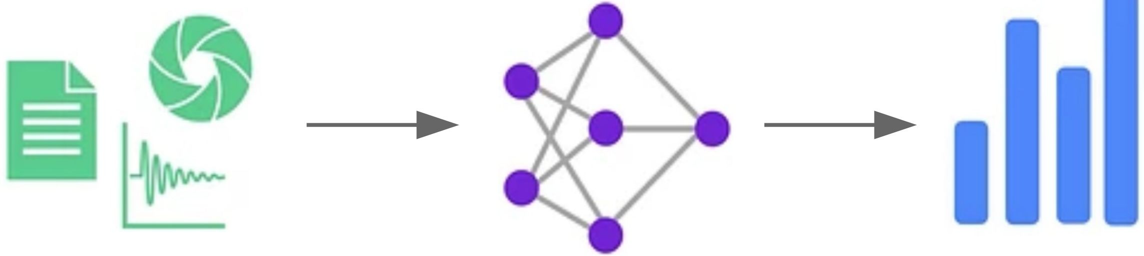
Open foundation models: improving scaling

- Long-term goal: improve open foundation models scalability, provide strongly transferable generalist models as basis for basic research

OPEN-SOURCE
Dataset &
Dataset composition

OPEN-SOURCE
Training procedure,
model weights,
checkpoints

OPEN-SOURCE
Evaluation benchmarks,
downstream transfer procedures



Supercomputers required!

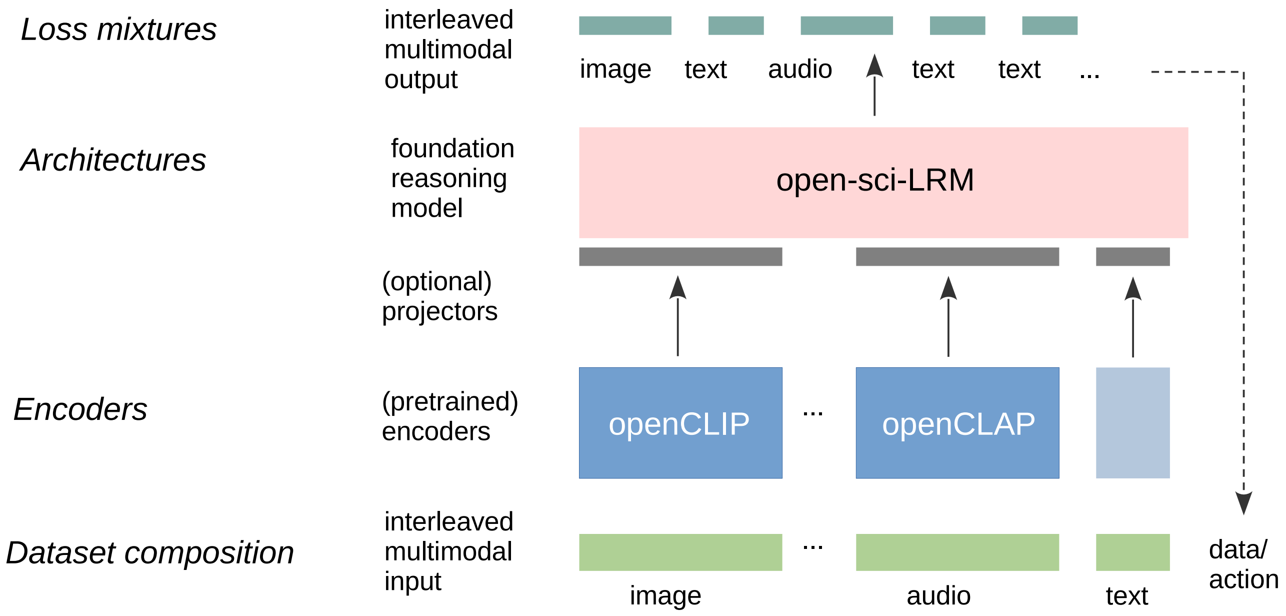
Dataset composition studies, scaling laws

Learning procedure studies, scaling laws

Novel benchmarks for model capabilities, transfer

Open multi-modal foundation models: progress

- Scaling laws for guided search of scalable open FoMos
- Comparison via reference scaling laws for established FoMo designs
  - eg MLLMs (VLMs, etc): pretrained FoMo components of various modality, 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 agentic 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

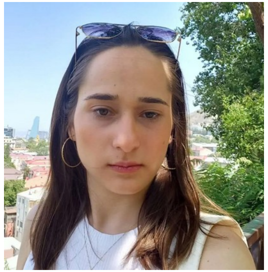
- **OSFoMo Alliance** : **Coordination of colab and resource acquisition** for open source foundation models and datasets R & D
- Build by orgas with strong track of record researching and building open FoMos
  - HuggingFace (**EU**), BlackForestLabs (**EU**), PriorLabs (**EU**), LAION (**EU**), TogetherAI, 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/Kimi/Qwen/GPT OSS level), strong reasoning and generalization
  - Open multi-modal language action models & datasets (transferable backbone for agents, open OS for robotics & automonous systems)

Open foundation models: outlook

- „Moonshot“: **open-sci-MMA - open multi-modal foundation agentic models**
  - Identifying better candidates via scaling law derivation based search

- **OpenEuroLLM, ELLIOT – LAION/ELLIS & friends** : EU consortia for building open foundation models with strongly improved generalization & reasoning
  - Will deliver the strong reasoning language models for open-sci-MMA
    → Hiring - Join us! Multiple open ML researcher (junior/senior postdoc levels), large scale machine learning engineers, science managers/administrators positions open (drop a message j.jitsev@fz-juelich.de)

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!

BigScience

EleutherAI

LAION
Large-scale Artificial Intelligence Open Network

open-sci collective

OPEN EURO LLM

MINERVA

ELLIOT

ellis INSTITUTE TÜBINGEN

ellis European Laboratory for Learning and Intelligent Systems

Tübingen AI Center

JÜLICH Forschungszentrum

JÜLICH SUPERCOMPUTING CENTRE

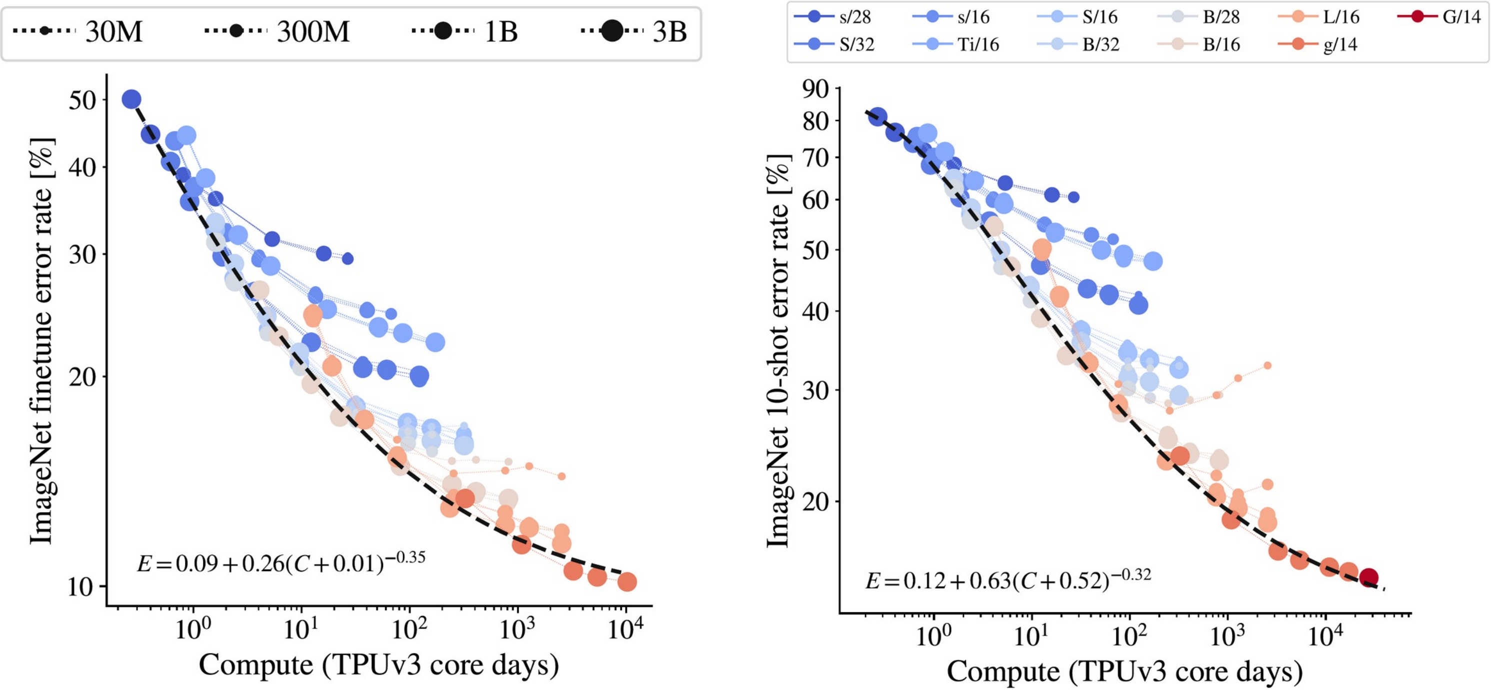
WEST AI KI-Servicezentrum

Thanks
for
your
Attention

Supplementary Material

Foundation models: scaling laws

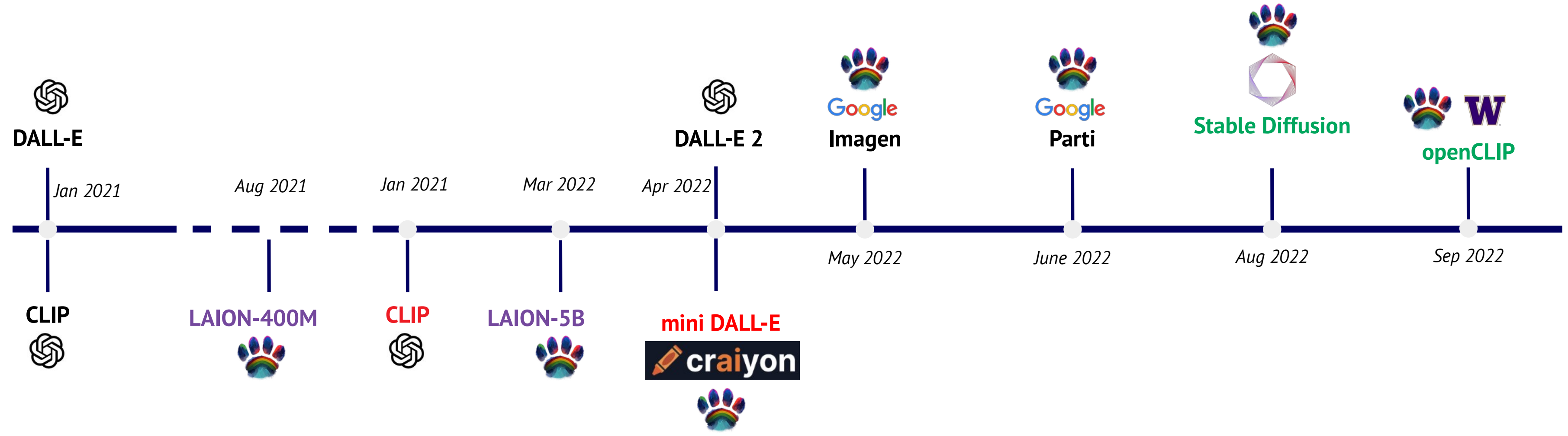
- Scaling Laws: exist for various generalist learning procedures
- Example: Supervised classification, ViT (JFT-3B dataset)



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“

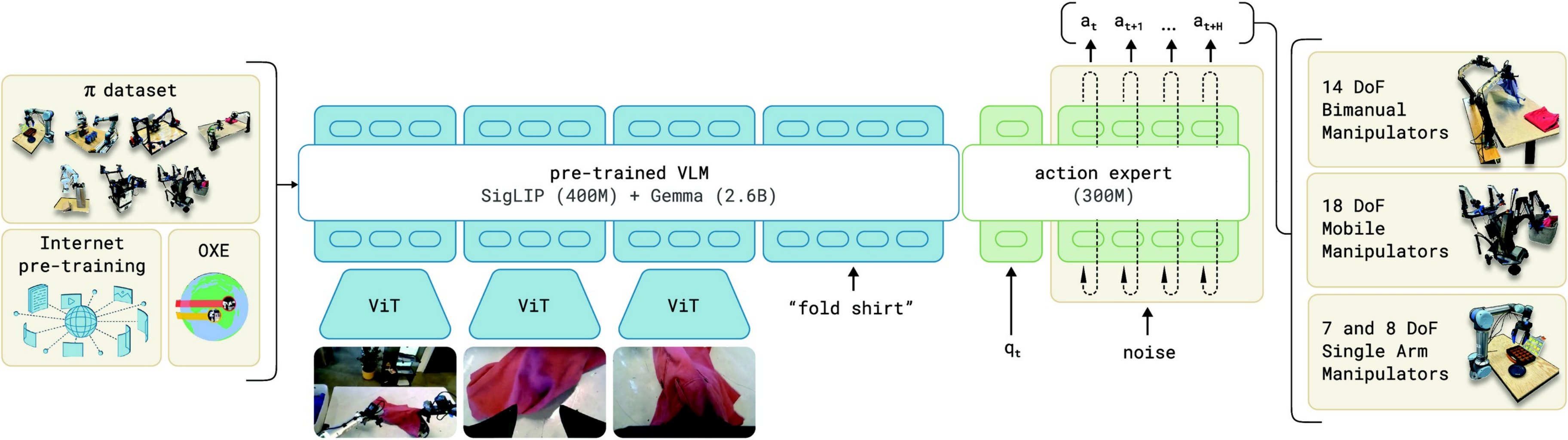
Open science for large-scale foundation models

- 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

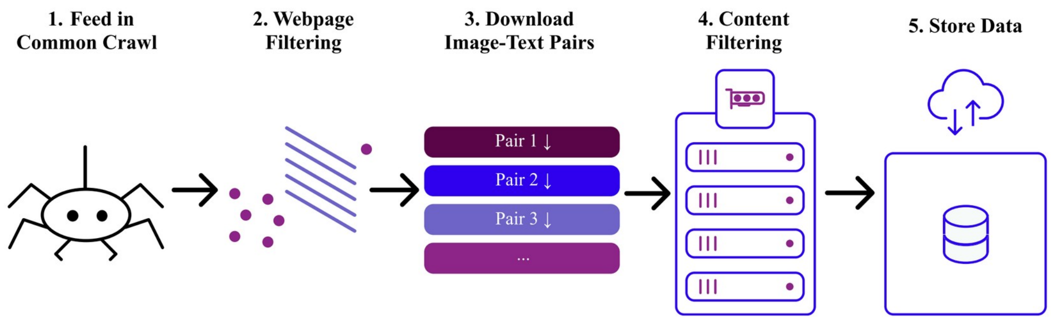
- Combining pre-trained models into multi-modal generalist foundation models (no or little adaptation required): Flamingo, BLIP-2, ImageBind, LENS, LlaVA, EMU, MM-1, PaliGemma, ...



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

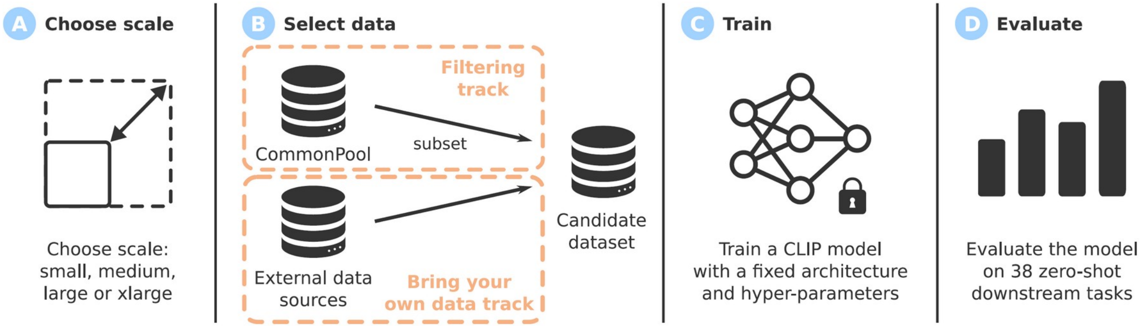
Dataset	# English Img-Txt Pairs
Public Datasets	
MS-COCO	330K
CC3M	3M
Visual Genome	5.4M
WIT	5.5M
CC12M	12M
RedCaps	12M
YFCC100M	100M[2]
LAION-5B (Ours)	**2.3B**
Private Datasets	
CLIP WIT (OpenAI)	400M
ALIGN	1.8B
BASIC	6.6B

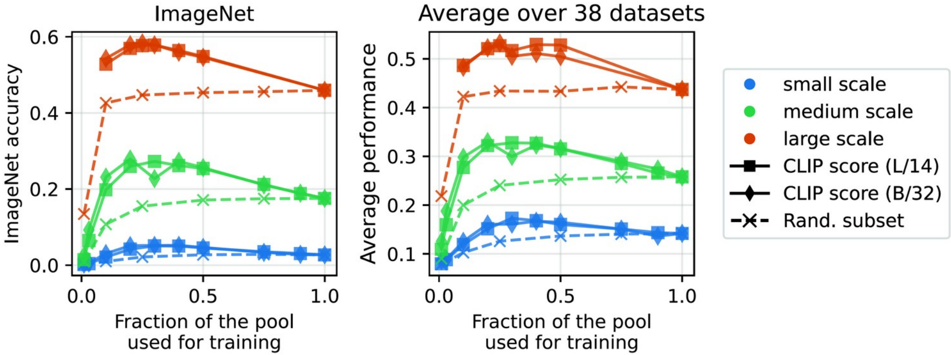
- Follow-ups: DataComp-1B; Re-LAION (safety revision update, Aug 2024)

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	1.1×10^{21}	75.5
LAION-400M [128, 28]	0.4B	13B	ViT-L/14	1.1×10^{21}	72.8
LAION-2B [129, 28]	2.3B	13B	ViT-L/14	1.1×10^{21}	73.1
LAION-2B [129, 28]	2.3B	34B	ViT-H/14	6.5×10^{21}	78.0
LAION-2B [129, 28]	2.3B	34B	ViT-g/14	9.9×10^{21}	78.5
DataComp-1B (ours)	1.4B	13B	ViT-L/14	1.1×10^{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

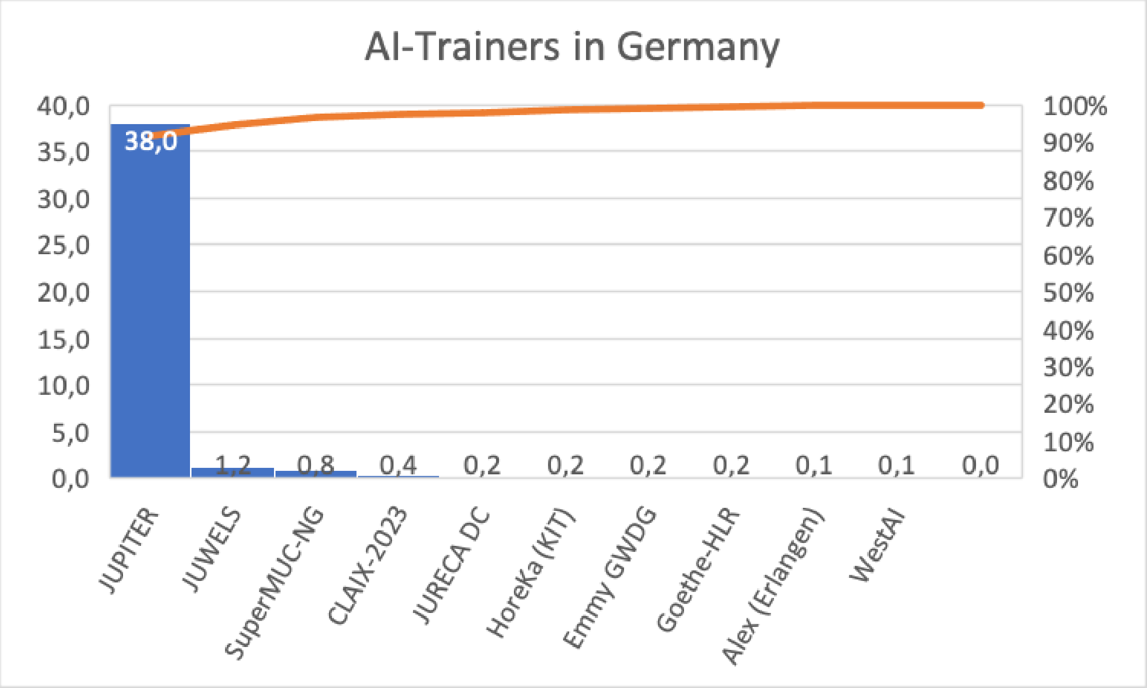
- Perform learning procedure comparison
- Guide search towards stronger scalable learning procedures

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)

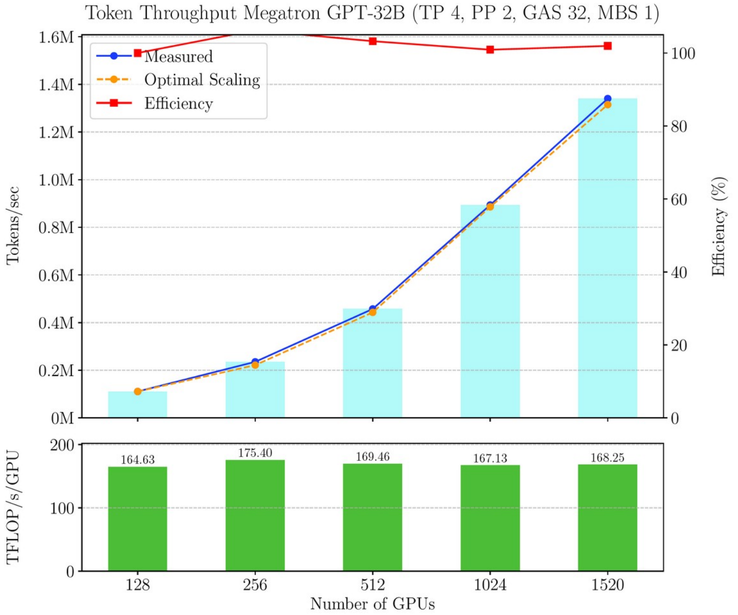
Open science for large-scale foundation models

- **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



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

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.