



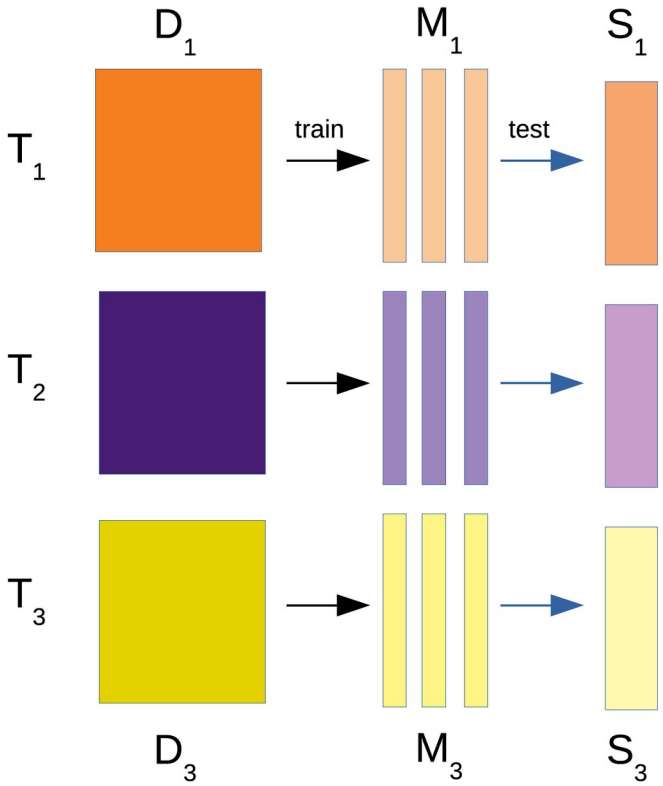
Open foundation models: reproducible scaling laws and generalization

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)

07. May 2025 | Marianna Nezhurina, Mehdi Cherti, Jenia Jitsev

Foundation models: generic transferable learning

- Machine learning before (< 2012): **poor generalization, poor transfer**
- Relying on **accurately labeled data for each task**, **specialized models (no re-use)**

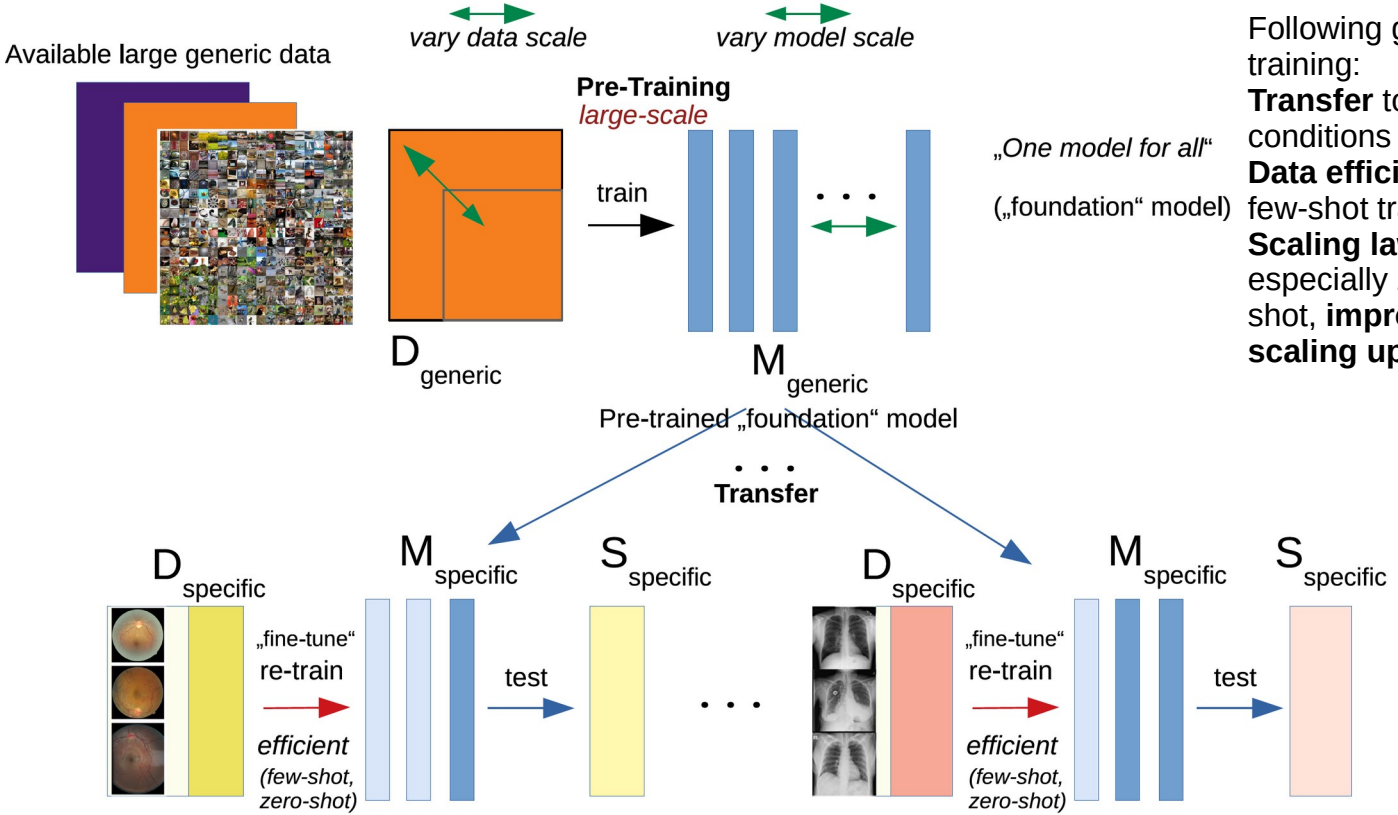


Long training

Slow
Large (Labeled) Data

Foundation models: generic transferable learning

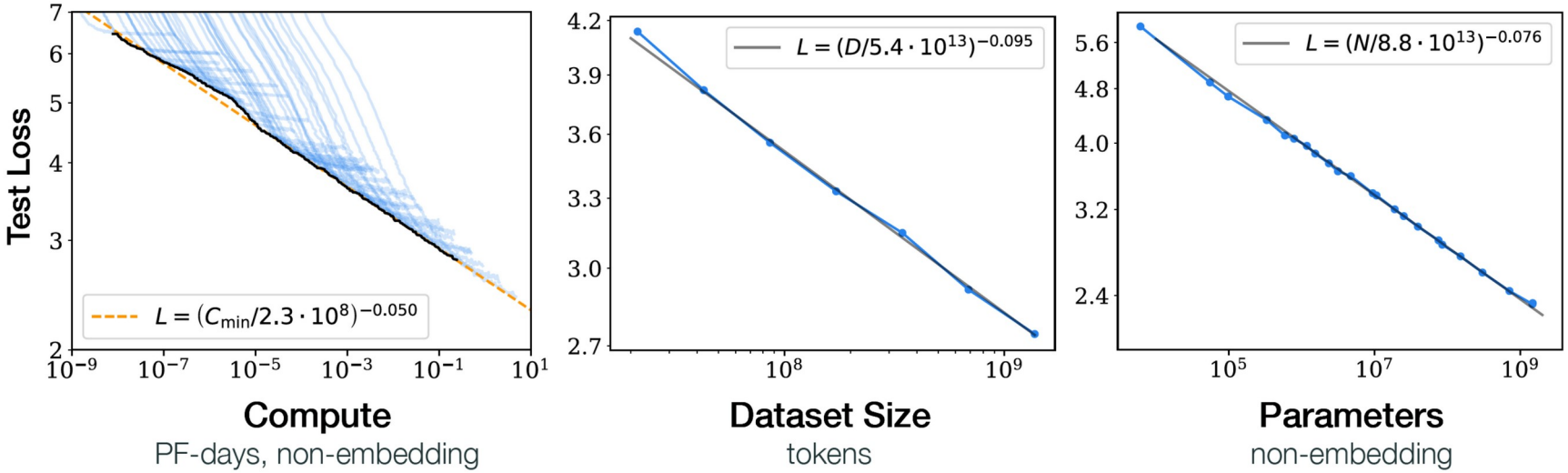
- Core breakthroughs (since ca. 2012): **scalable learning that transfers across tasks**



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

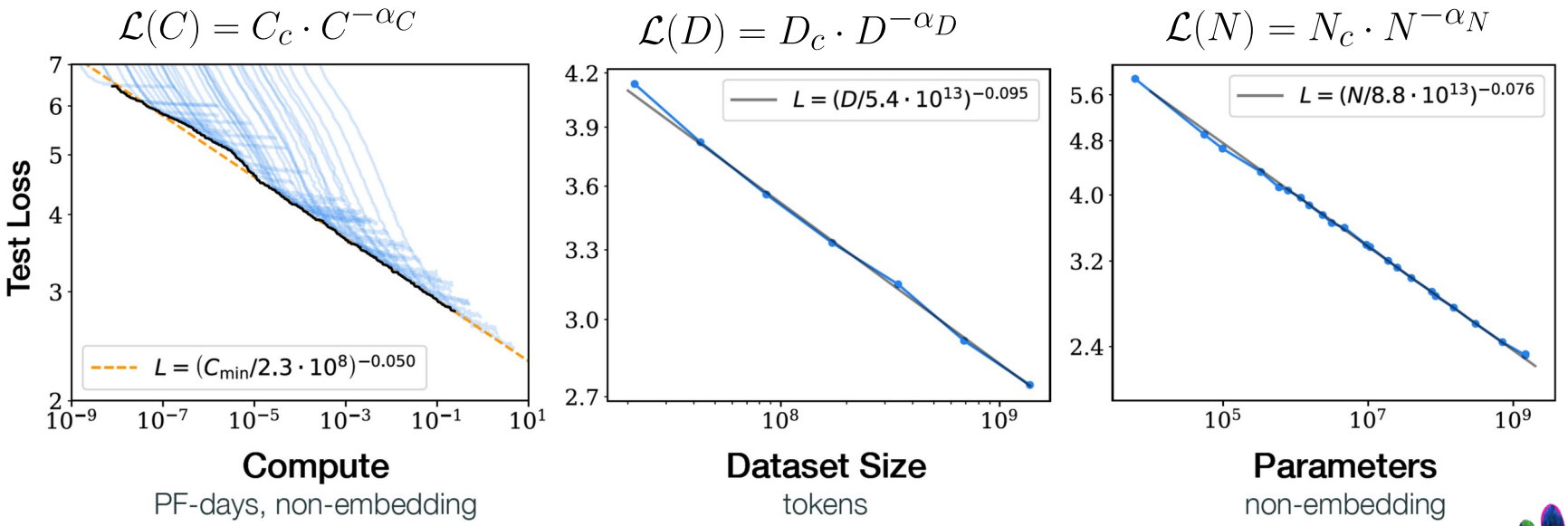
- **Scaling Laws**: larger model, data and compute scale during pre-training – **stronger generalization & transferability**
- **No change** in core algorithmic procedure required! Scaling up alone improves important core functions



Kaplan et al, 2020

Foundation models: scaling laws

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

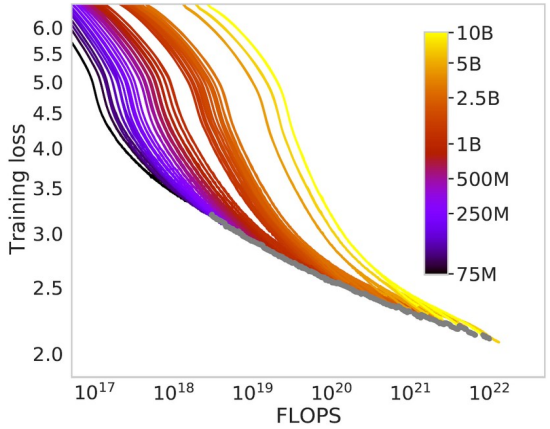


Kaplan et al, 2020

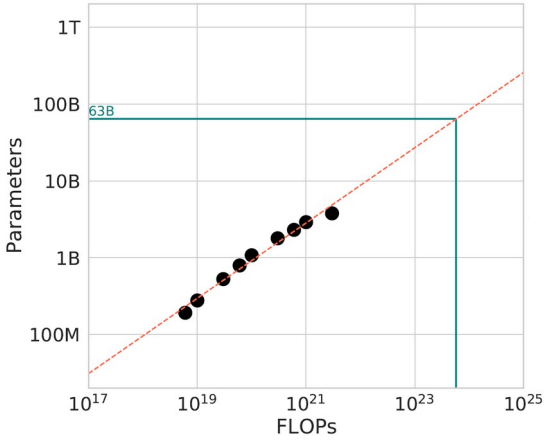
Foundation models: scaling laws

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

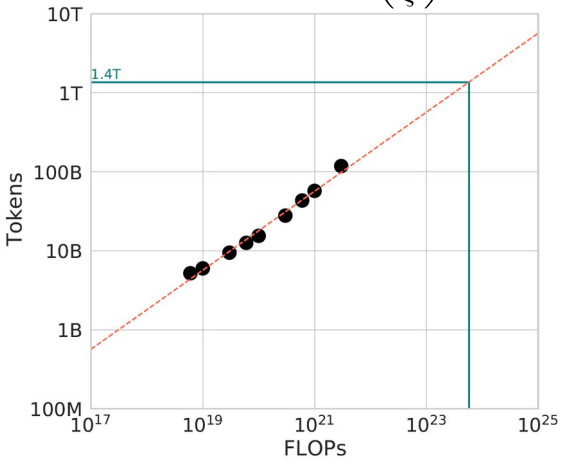
$$\mathcal{L}(C) = C_c \cdot C^{-\alpha_C} + L_\epsilon$$



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



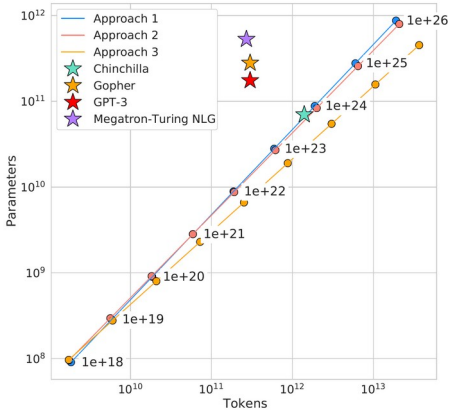
$$D^*(C) = G^{-1}\left(\frac{C}{\xi}\right)^b$$



Hoffmann et al, arXiv:2203.15556, NeurIPS, 2022

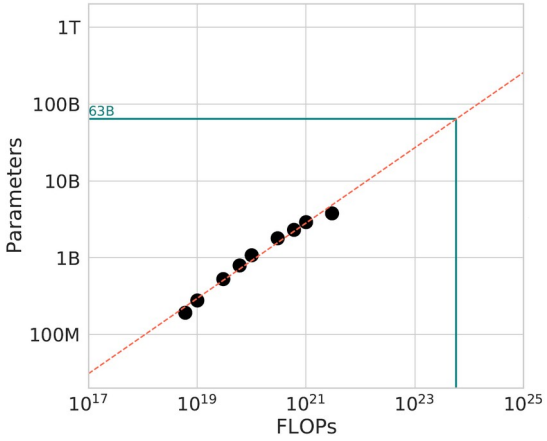
Foundation models: scaling laws

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

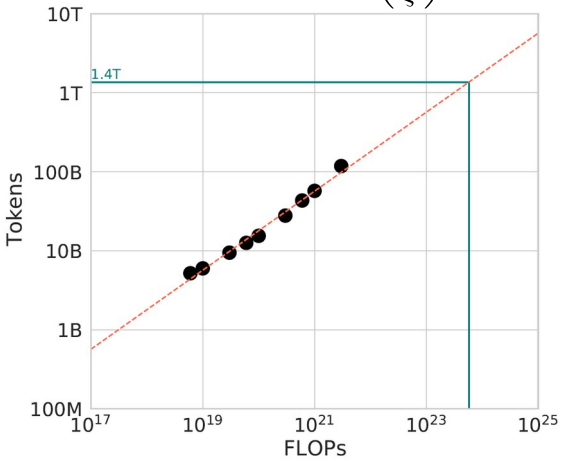


Parameters	FLOPs	FLOPs (in *Gopher* unit)	Tokens
400 Million	1.92e+19	1/29,968	8.0 Billion
1 Billion	1.21e+20	1/4,761	20.2 Billion
10 Billion	1.23e+22	1/46	205.1 Billion
67 Billion	5.76e+23	1	1.5 Trillion
175 Billion	3.85e+24	6.7	3.7 Trillion
280 Billion	9.90e+24	17.2	5.9 Trillion
520 Billion	3.43e+25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

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



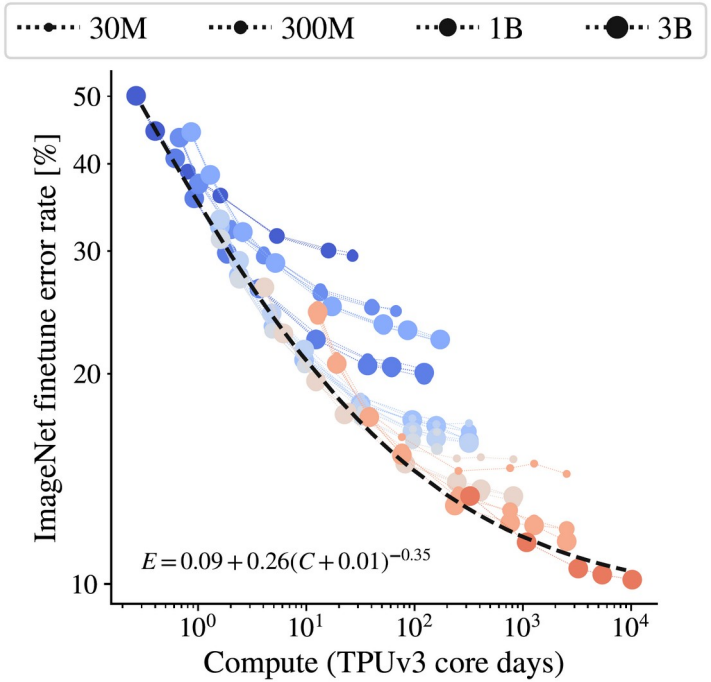
$$D^*(C) = G^{-1}\left(\frac{C}{\xi}\right)^b$$

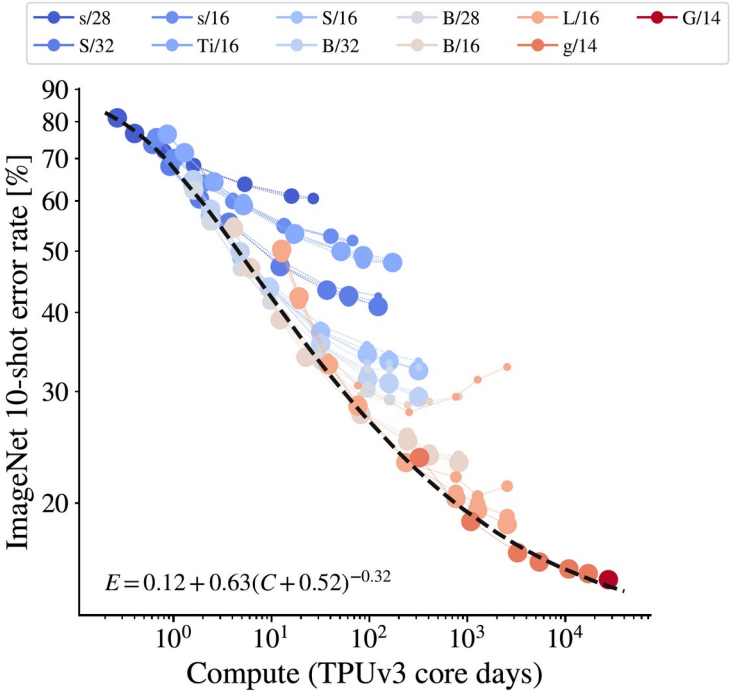


Hoffmann et al, arXiv:2203.15556, 2022

Foundation models: scaling laws

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

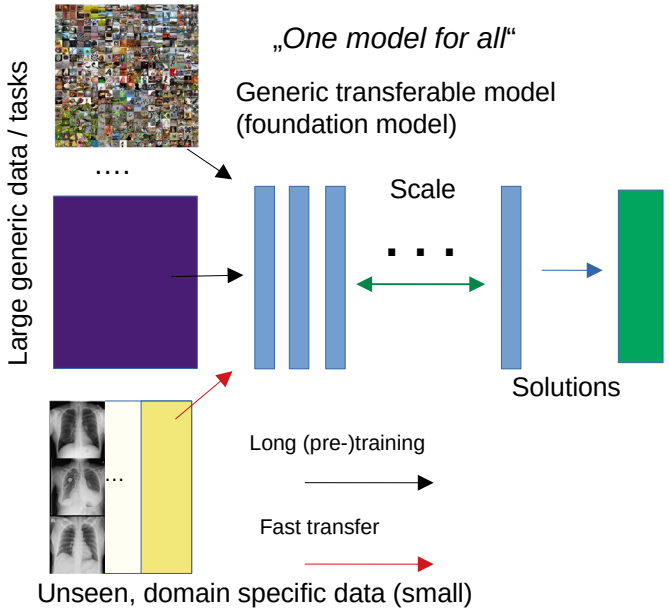




Zhai et al, arXiv:2106.04560, CVPR, 2022

Foundation models: reproducibility & progress

- **Problem**: research on foundation models, datasets & scaling laws reproducible only by few large industry labs (Google; openAI; Microsoft; Facebook; NVIDIA; …)
- **Important large foundation models:** GPT-3/4, DALL-E 2/3 Flamingo, CLIP **- closed to public research**
- **Datasets** used to train those models**: REQUIRED! closed**
- **Non-reproducible, 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, Llema (math)
- **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
- **Open large datasets and foundation models: reproducibility !**
  - **joint efforts accross institutions/organisations boundaries**

EleutherAI

BigScience

LAION

Large-scale Artificial Intelligence Open Network

JÜLICH Forschungszentrum

JÜLICH SUPERCOMPUTING CENTRE

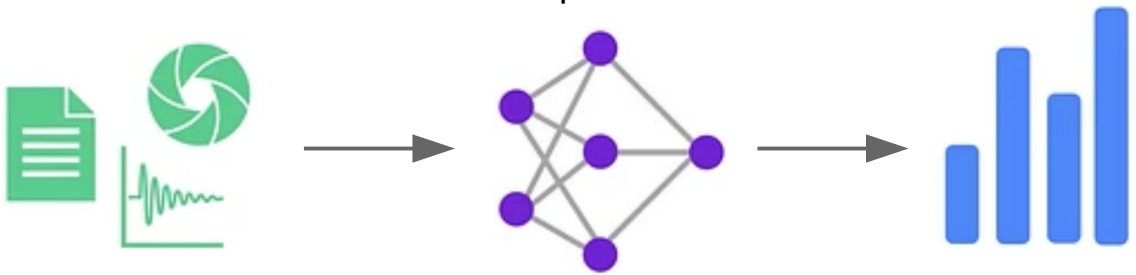
Open-source foundation models & datasets

- 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!

LAION-400M,
LAION-5B,
DataComp-1B

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

OpenCLIP,
openFlamingo

https://github.com/mlfoundations/open_clip

openCLIP
Benchmarks

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

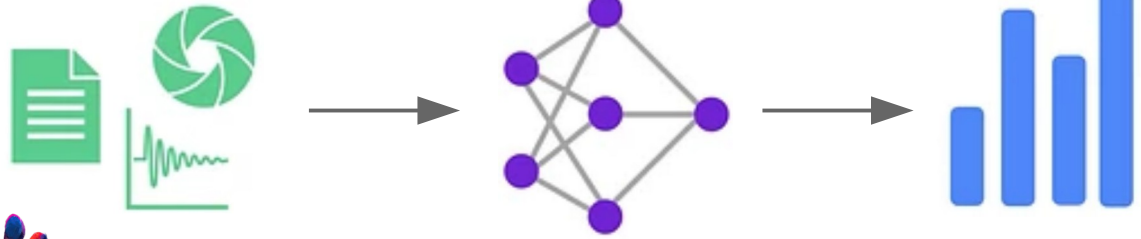
Open-source foundation models & datasets

- 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, model stress-test downstream transfer procedures



BigScience
EleutherAI
together.ai
AI2

Pile,
RedPajama,
Dolma.
DCLM-Baseline

BigScience
AI2
EleutherAI

Pythia, Together-INCITE, Olmo.
open-LM/DCLM

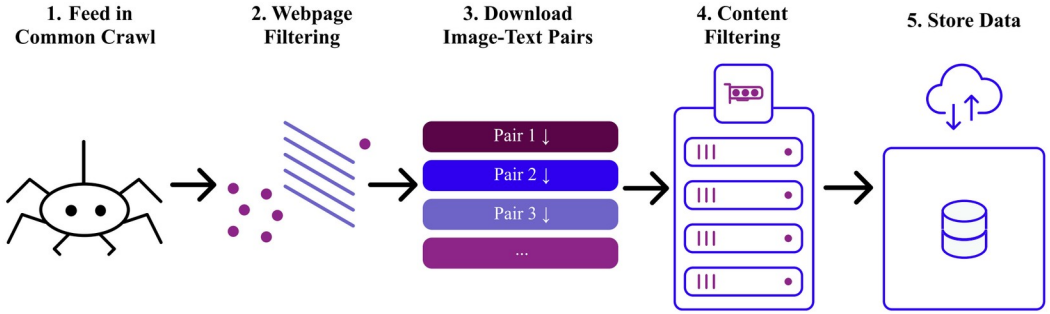
EleutherAI
AI2
BigScience

Lm-eval-harness, bigcode-evaluation-harness, LAION-AIW, EvalChemy

https://github.com/EleutherAI/lm-evaluation-harness

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

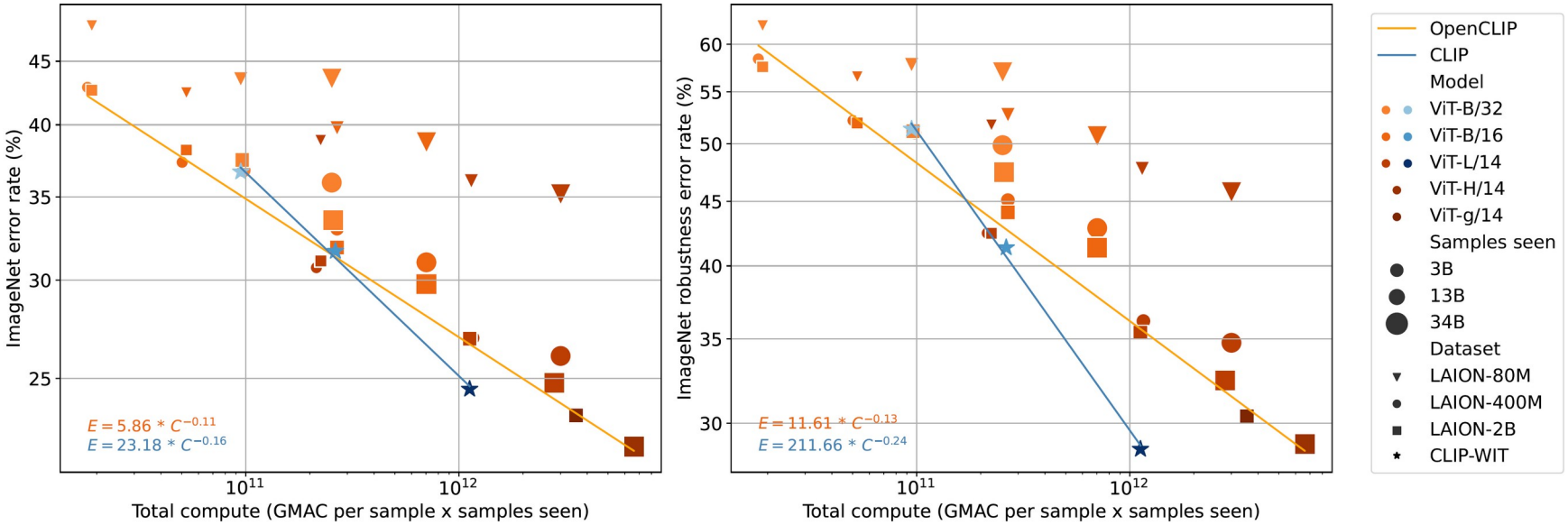
Dataset	# English Img-Txt Pairs
Public Datasets	
MS-COCO	330K
CC3M	3M
Visual Genome	5.4M
WIT	5.5M
CC12M	12M
RedCaps	12M
YFCC100M	100M[2]
LAION-5B (Ours)	**2.3B**
Private Datasets	
CLIP WIT (OpenAI)	400M
ALIGN	1.8B
BASIC	6.6B

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

Schuhmann et al, NeurIPS, 2022

Reproducible scaling laws for foundation models

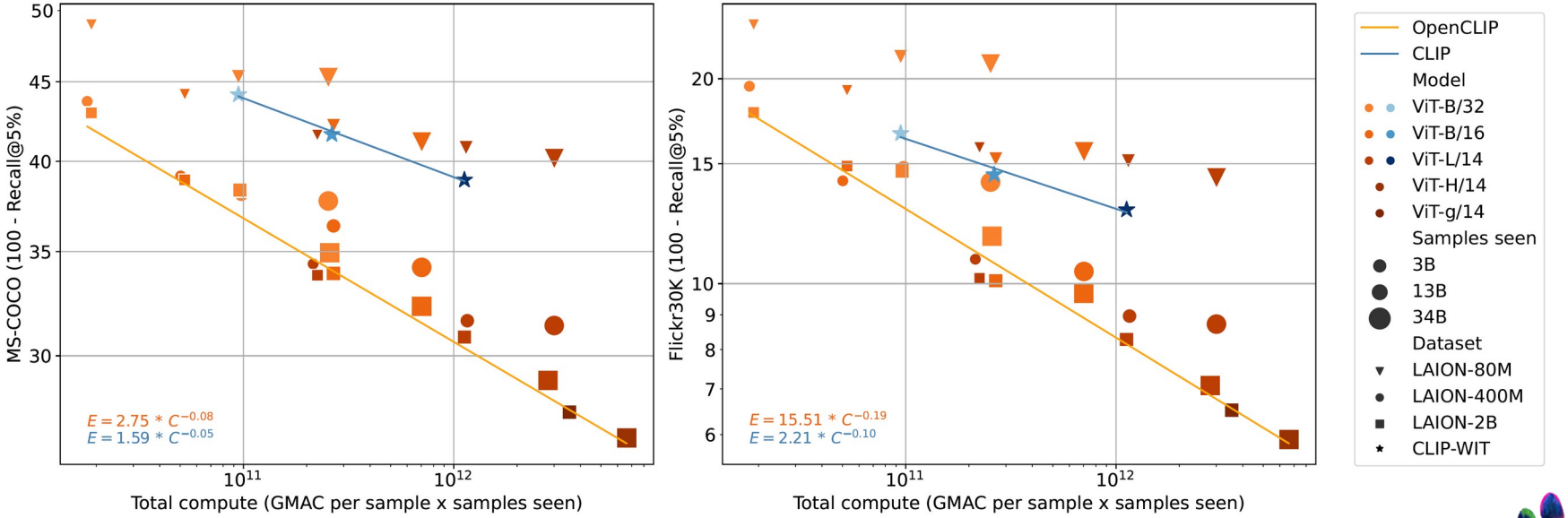
- Scaling laws with LAION-400M/2B and openCLIP: open-source data, models and code - **reproducible** science of foundation models
- Below: zero-shot image classification, ImageNet-1k & robustness sets



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

Scaling laws: pre-training procedure comparison

- Comparing LAION-400M/2B (LAION) and WIT (openAI)
- Zero-shot image retrieval, MS-COCO & Flickr30K
- Task dependency (classification vs retrieval) revealed
- **Matching or beating strong closed models** by using open data



Cherti et al, CVPR 2023

Open foundation models: comparison

- DataComp-LM: fully open, reproducible pipeline for language modelling; rivaling Llama-3-8B & Mistral-7B; fully open data (DCLM-Baseline, 2.6T training tokens; 4.4T tokens in total) & models (DCLM-1B/3B/7B)

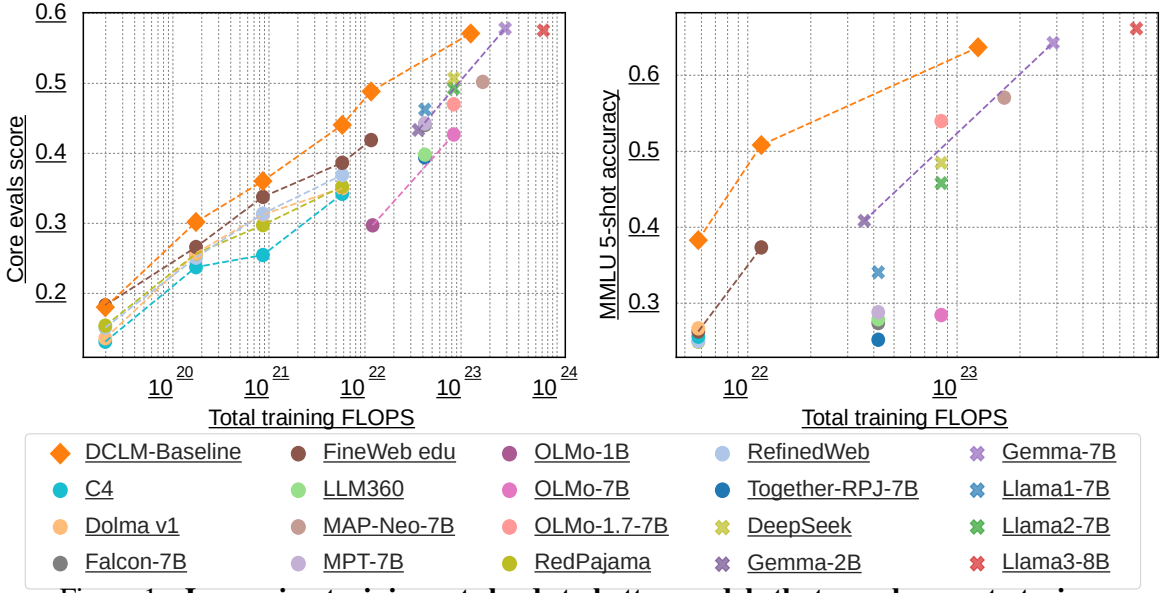


Figure 1: **Improving training sets leads to better models that are cheaper to train.**

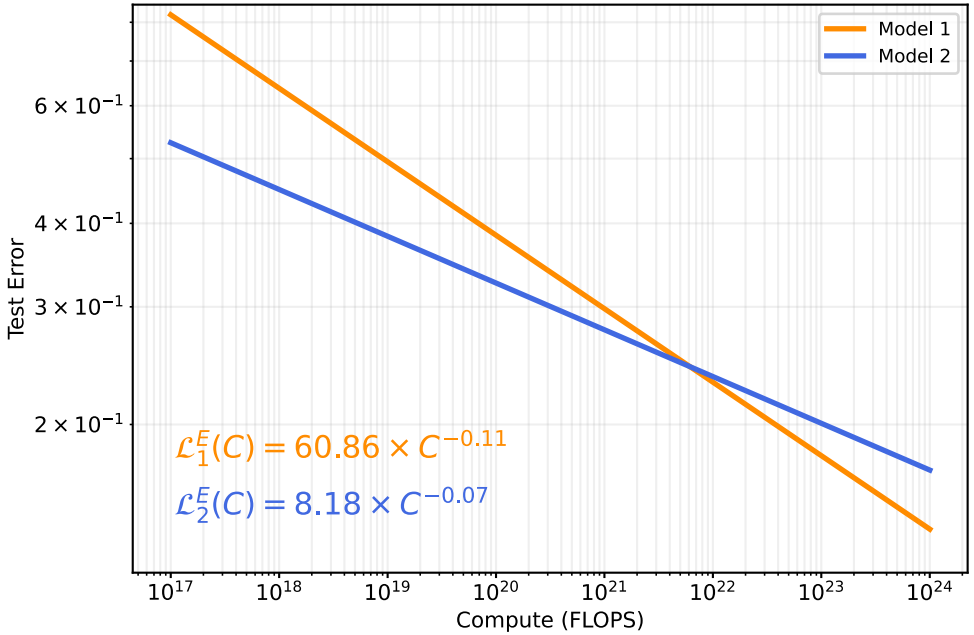
Li et al, ArXiv:2406.11794, NeurIPS, 2024

https://github.com/mlfoundations/dclm

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

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

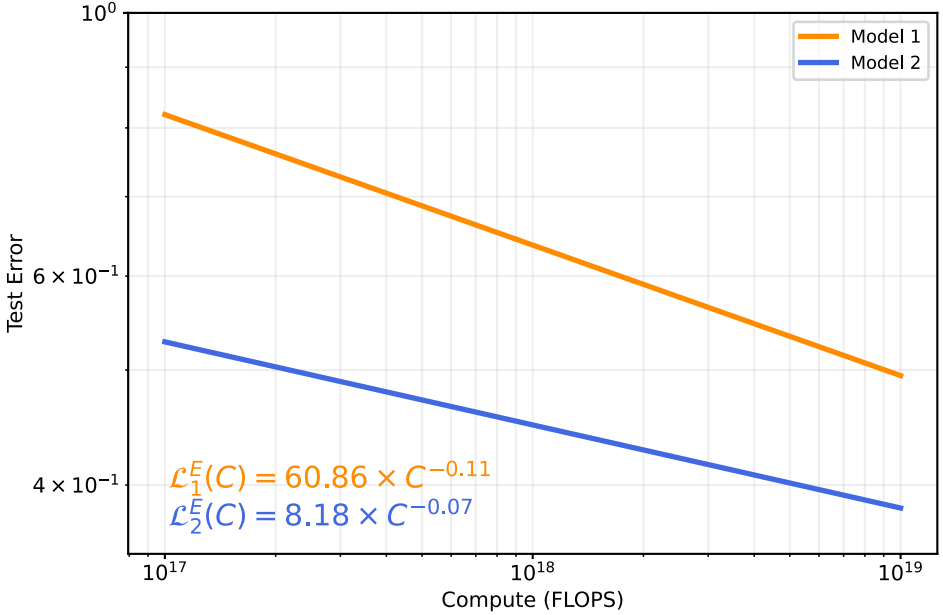


- Learning procedure 1 vs Learning procedure 2
- Scenarios:
  - Comparing Model 1 vs Model 1 while fixing same open data
  - Comparing open Dataset 1 vs Dataset 2 while fixing same open training/model
  - ...

Scaling laws: pre-training procedure comparison

- Comparison done properly requires proper scaling law derivation
  - measuring sufficient scaling span (comparison using single points not possible)
  - measuring at scales with sufficient signal (low performance range not predictive)

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

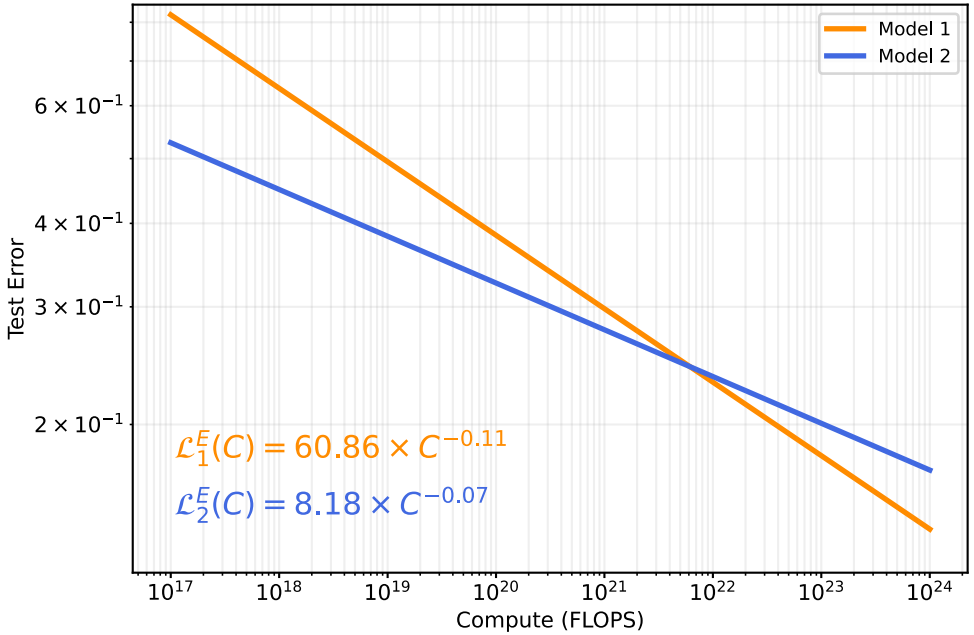


- Measuring narrow span at smaller scales, Model 2 may look much better than Model 1
- Advantage at poor performance levels is often deceptive
- Common issue of most „toy problem“ study designs: seemingly large advantage of method X turns out to vanish at relevant performance levels

Scaling laws: pre-training procedure comparison

- Comparison: offers measure to make systematic progress towards stronger learning procedures (datasets, training procedures, models, ...)

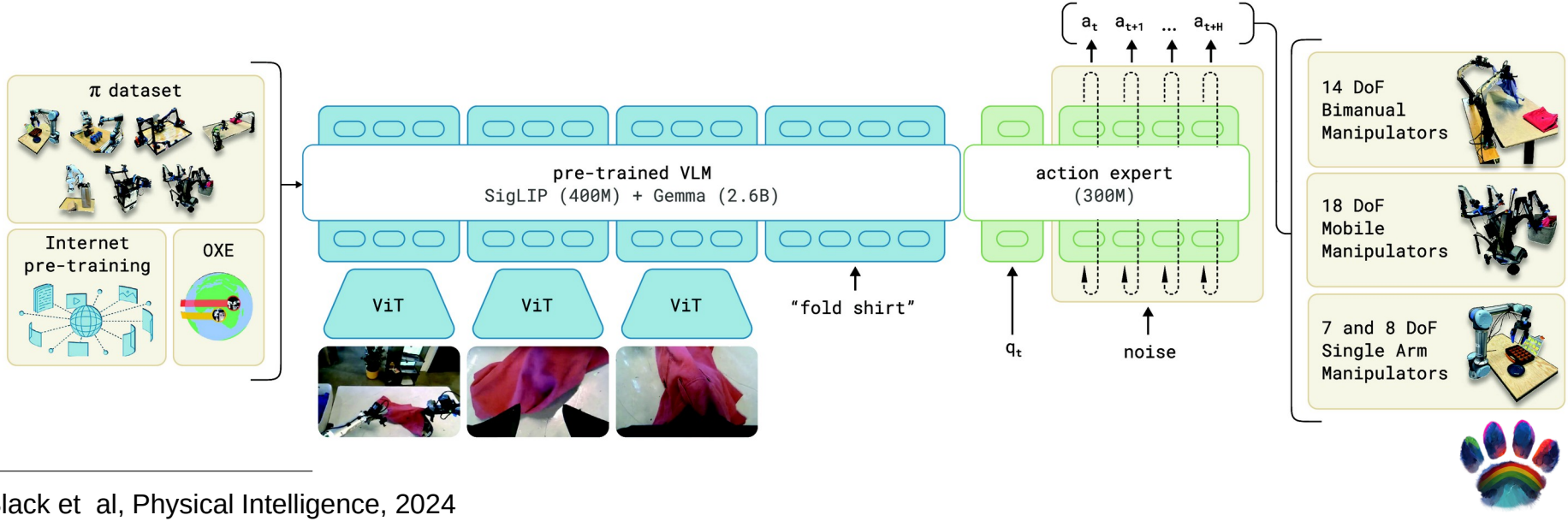
$$\mathcal{L}(C) = C_c \cdot C^{-\alpha_C} + L_\epsilon$$



- Without open data/training : impossible to differentiate between various factors contributing to the observed model performance (Was it data? Was it algorithm? Both?)
- Open standardized evals: making sure comparison done on same sets of measures

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, …
  - open foundation models as components: controlling whole pipeline, properties/function prediction and comparison via scaling laws



Black et al, Physical Intelligence, 2024

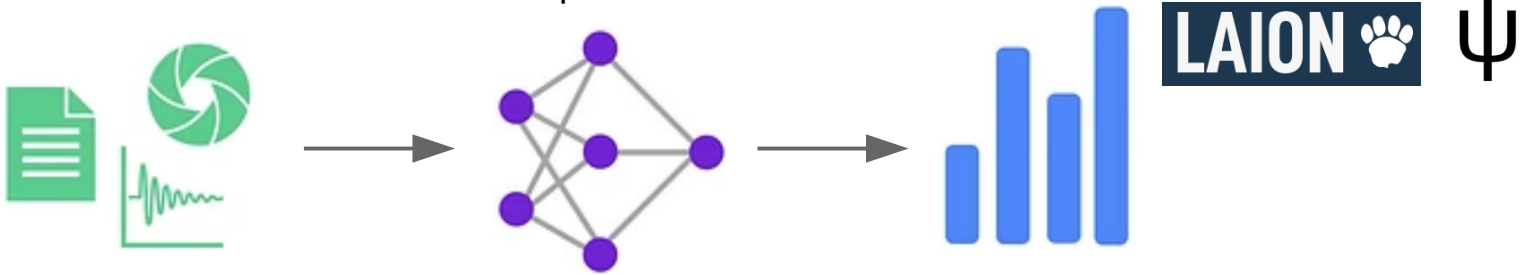
Open foundation models: improving scaling

- 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!

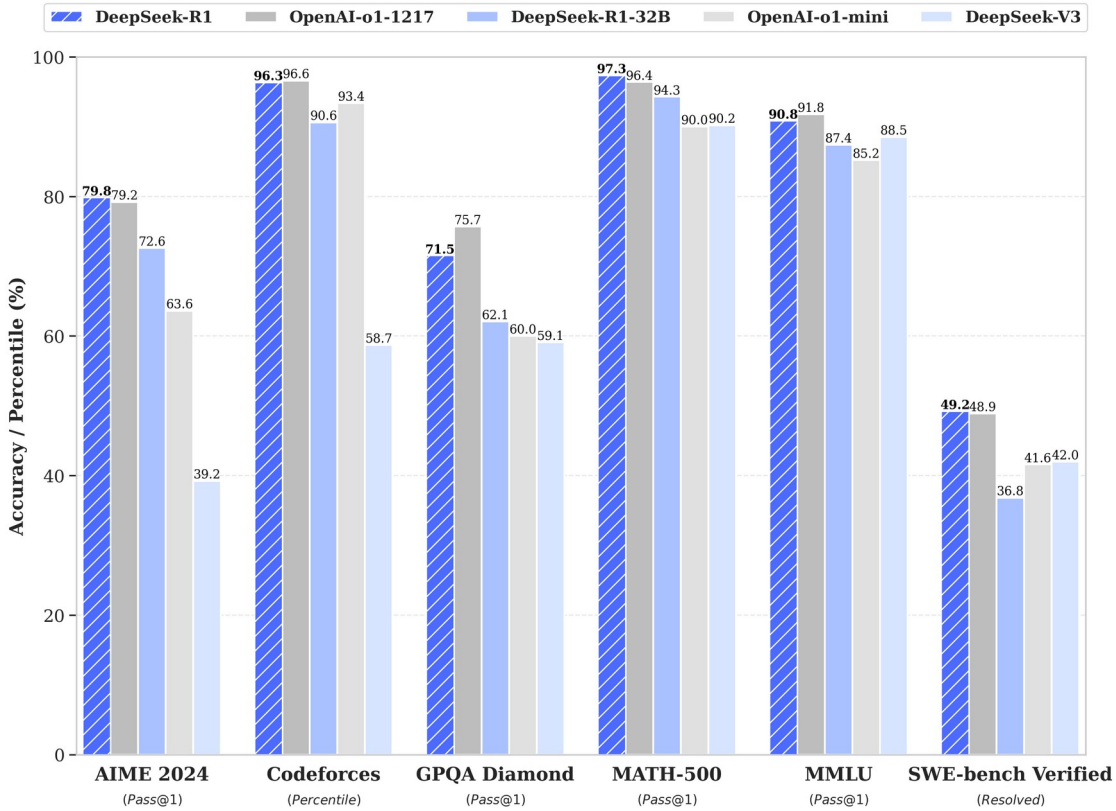
Dataset composition studies, scaling laws

Learning procedure studies, scaling laws

Novel benchmarks for model capabilities, transfer

Scaling laws: predicting generalization

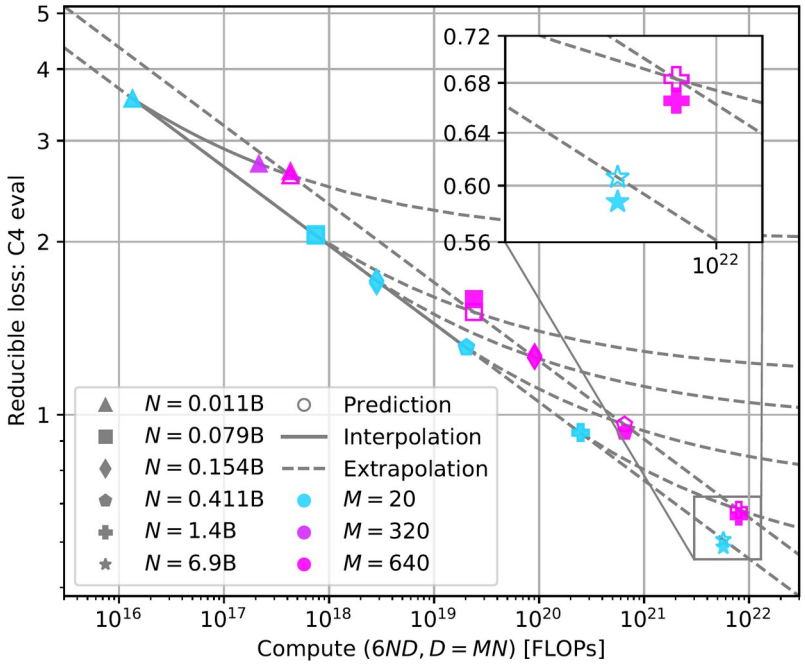
- Do benchmark downstream tasks reflect generalization properly?

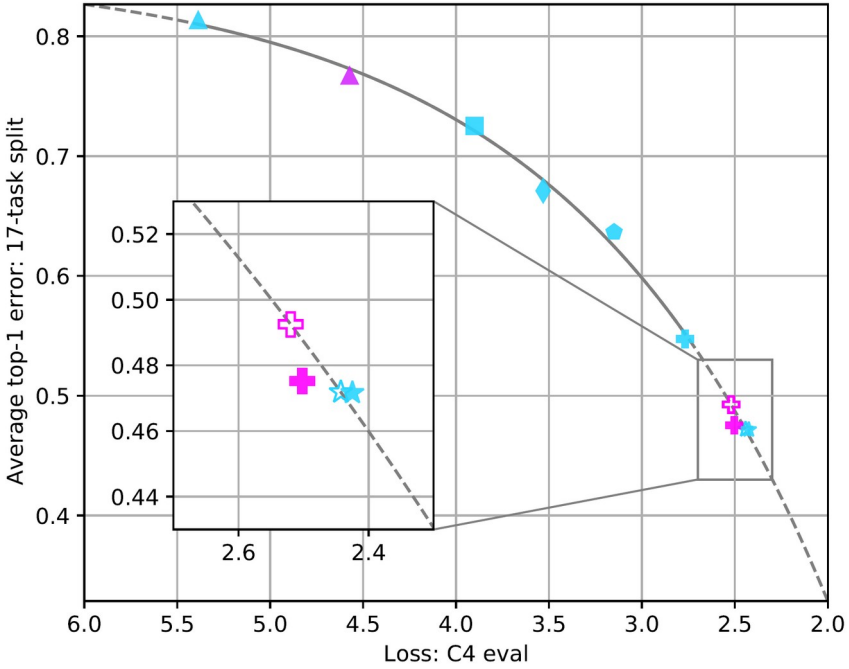


DeepSeek R1 Technical Report, DeepSeek, 2025

Scaling laws: predicting generalization

- Upstream loss might be predictive for benchmark downstream tasks performance

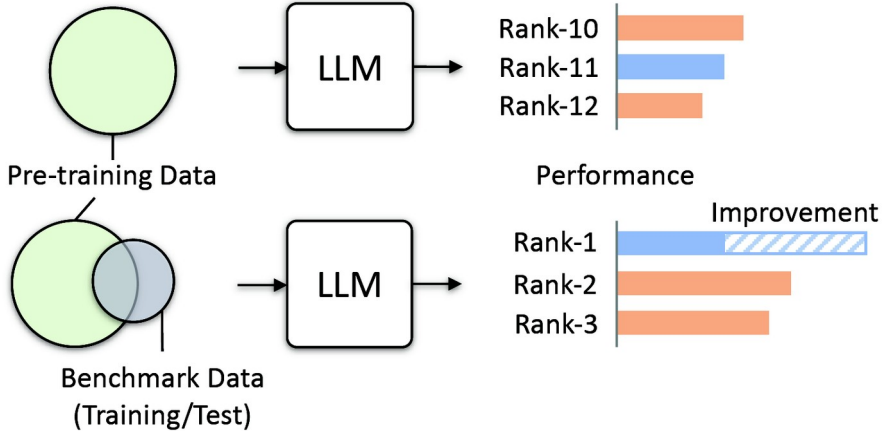


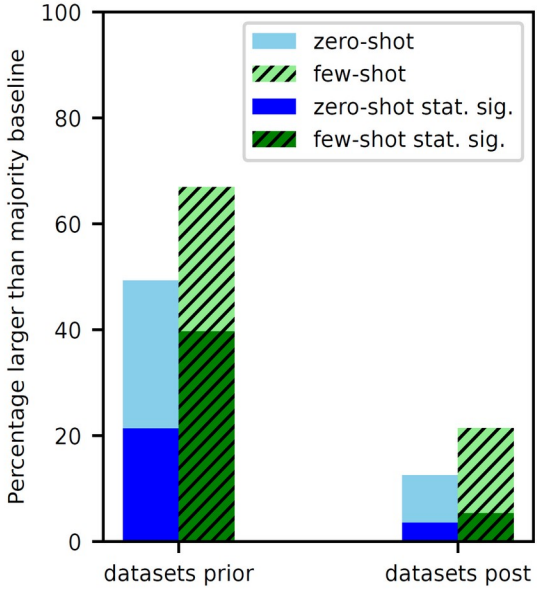


Gadre et al, arXiv:2403.08540, 2024; ICLR, 2025

Scaling laws: predicting generalization

- Do benchmark downstream tasks reflect generalization properly?
- Big issue: test set leakage, training data contamination
  - might **strongly overestimate** generalization capability





Zhou et al, arXiV:2311.01964, 2023; Li et al, arXiv:2312.16337, AAAI, 2024

Scaling laws: predicting generalization

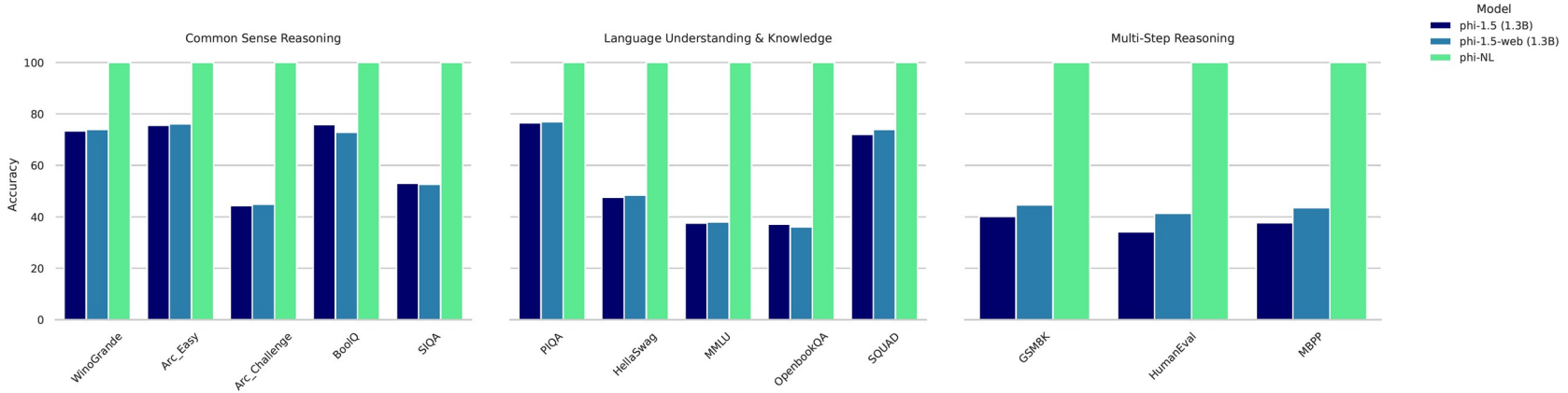
Pretraining on the Test Set Is All You Need

Rylan Schaeffer

September 19, 2023

Abstract

Inspired by recent work demonstrating the promise of smaller Transformer-based language models pretrained on carefully curated data, we supercharge such approaches by investing heavily in curating a novel, high quality, non-synthetic data mixture based solely on evaluation benchmarks. Using our novel dataset mixture consisting of less than 100 thousand tokens, we pretrain a 1 million parameter transformer-based LLM **phi-CTNL** (pronounced "fictional") that achieves perfect results across diverse academic benchmarks, strictly outperforming all known foundation models. **phi-CTNL** also beats power-law scaling and exhibits a never-before-seen grokking-like ability to accurately predict downstream evaluation benchmarks' canaries.



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 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?

LAION

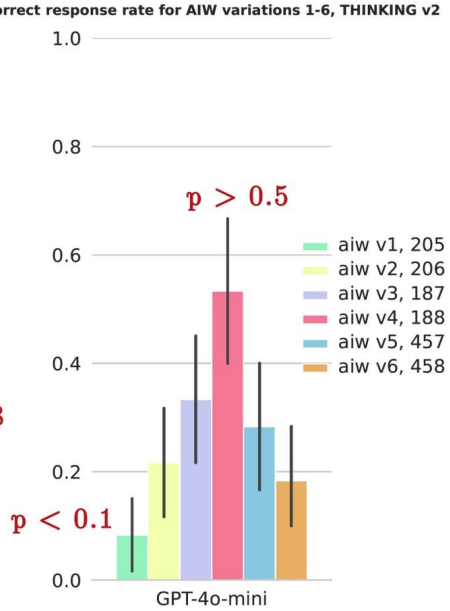
open-sci collective

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

Scaling laws: predicting generalization

- 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

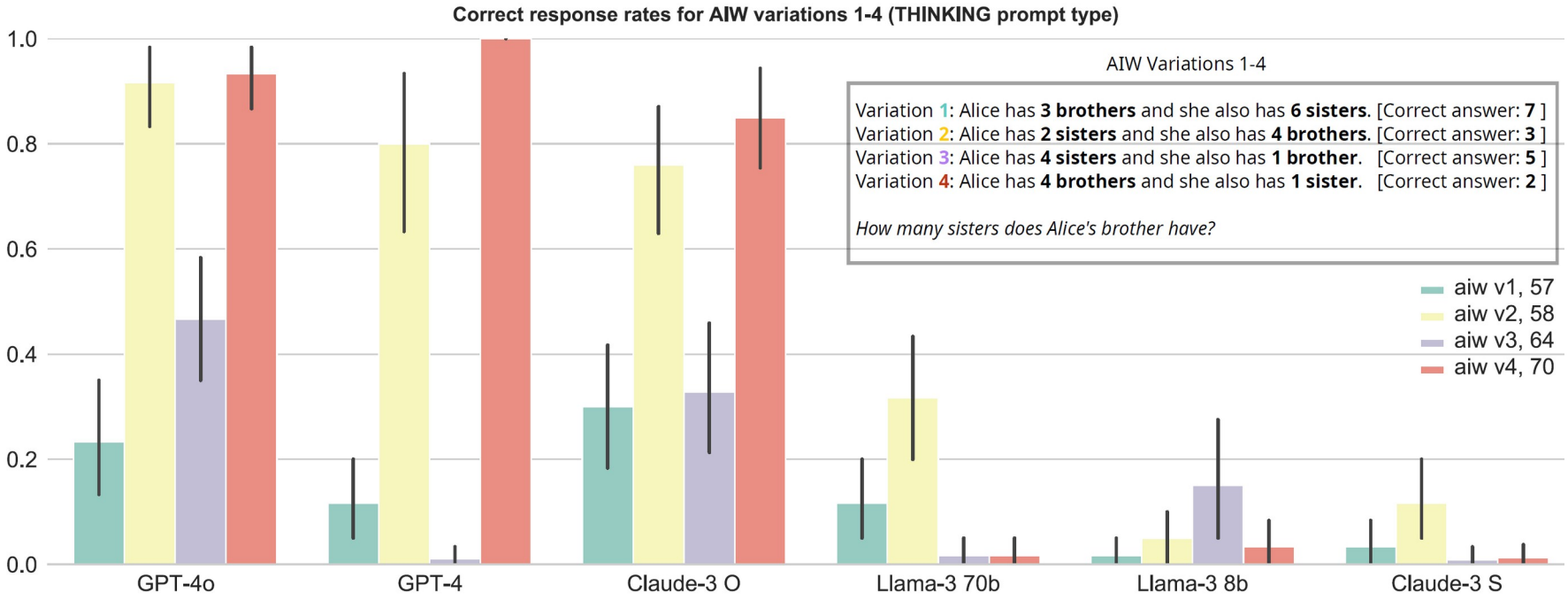
Variation 1: Alice has **3 brothers** and she also has **6 sisters**. [Correct answer: **7**]
Variation 2: Alice has **2 sisters** and she also has **4 brothers**. [Correct answer: **3**]
Variation 3: Alice has **4 sisters** and she also has **1 brother**. [Correct answer: **5**]
Variation 4: Alice has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

Scaling laws: predicting generalization

- SOTA LLMs 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

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

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**]

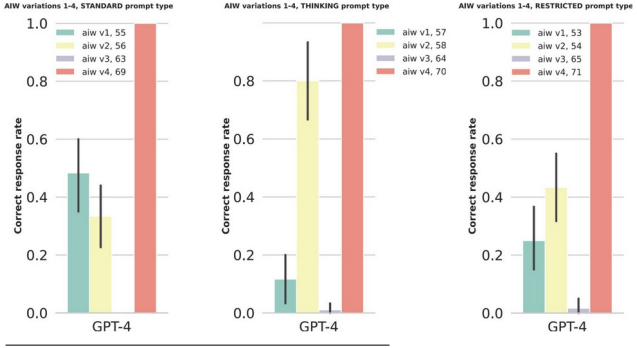
How many sisters does Alice's brother have?

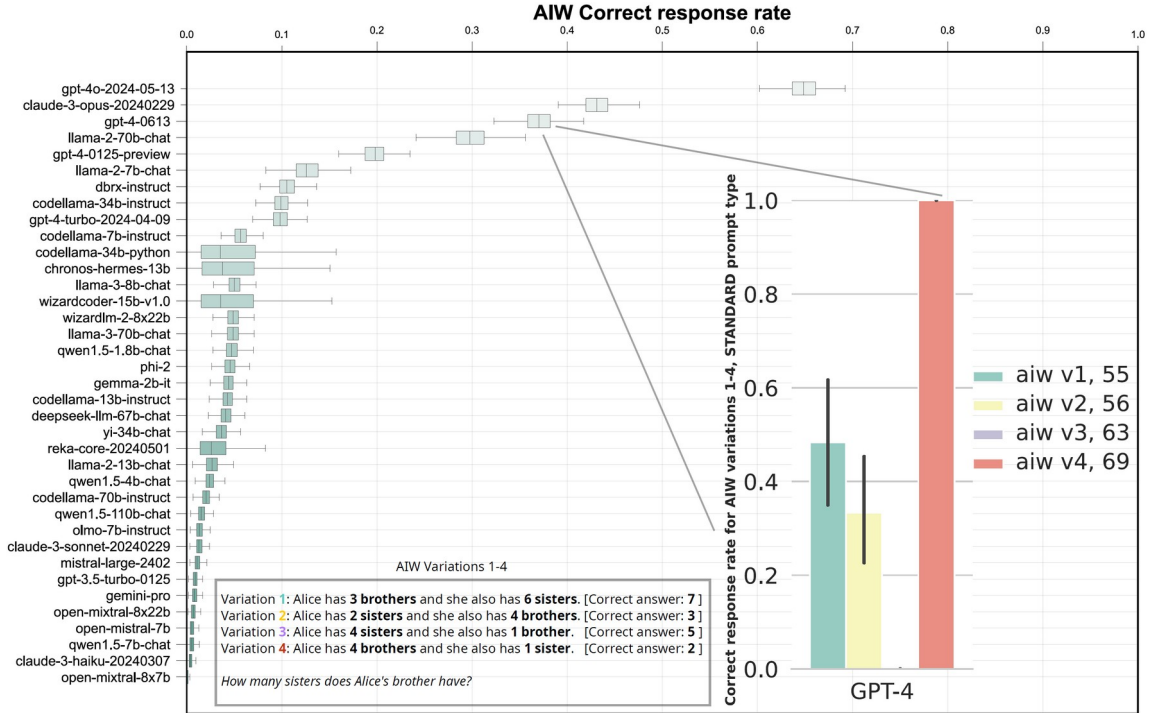
Prompt types (following after main problem description and question above):

STANDARD : Solve this problem and provide the final answer in following form: "### Answer: ".

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: ".

RESTRICTED : To answer the question, DO NOT OUTPUT ANY TEXT EXCEPT following format that contains final answer: "### Answer: ".





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

Scaling laws: predicting generalization

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

AIW Variations, Original and AIW Light Control

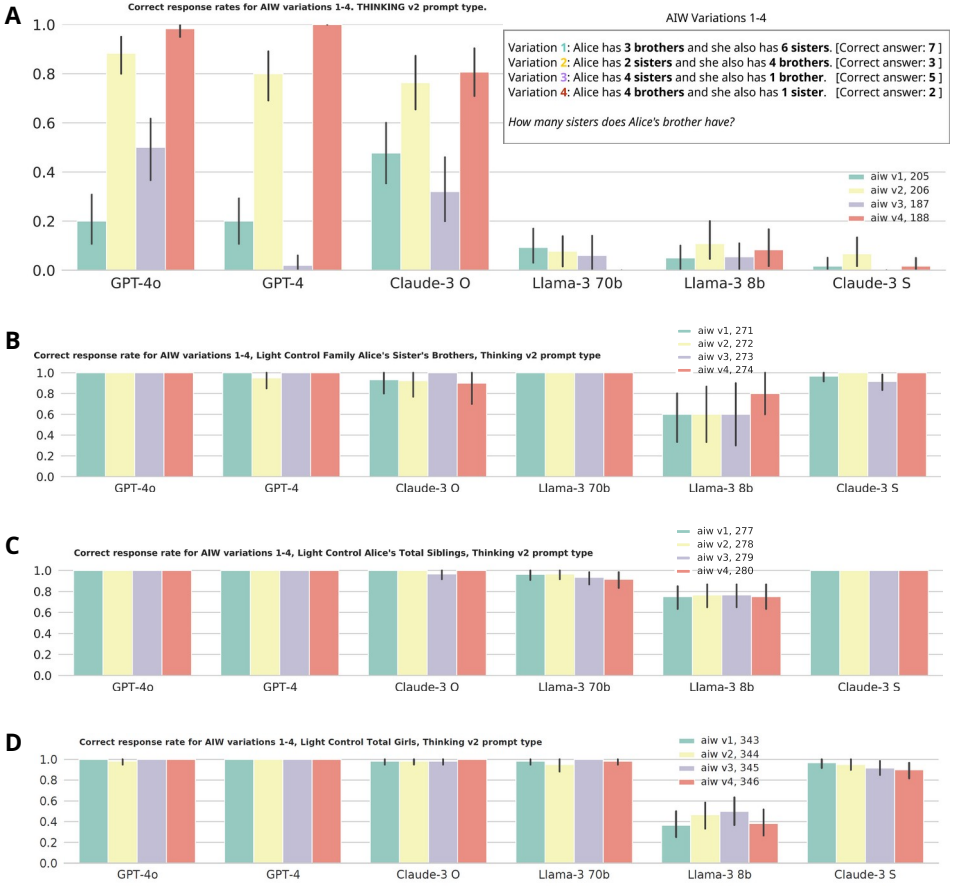
Template: Alice has *N* **brothers** and she also has *M* **sisters**.
Variations 1-4: changing *N*, *M* <= 7. Correct responses: *C* <= 7

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

AIW Light Control (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: ".



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?

Scaling laws: predicting generalization

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

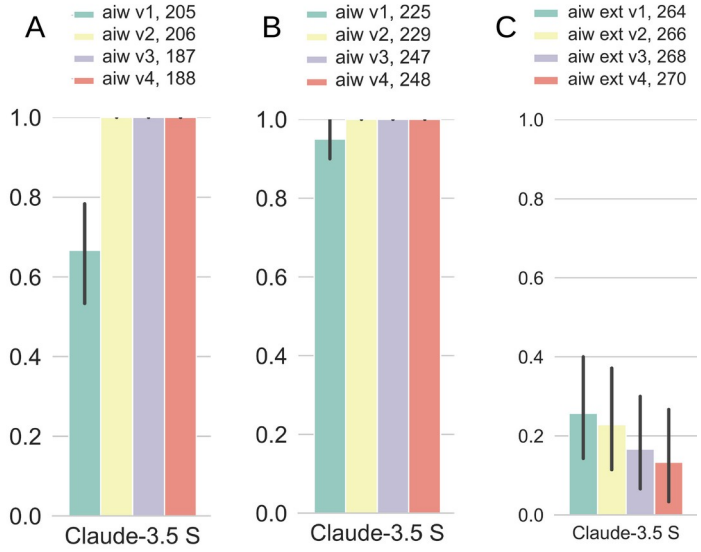
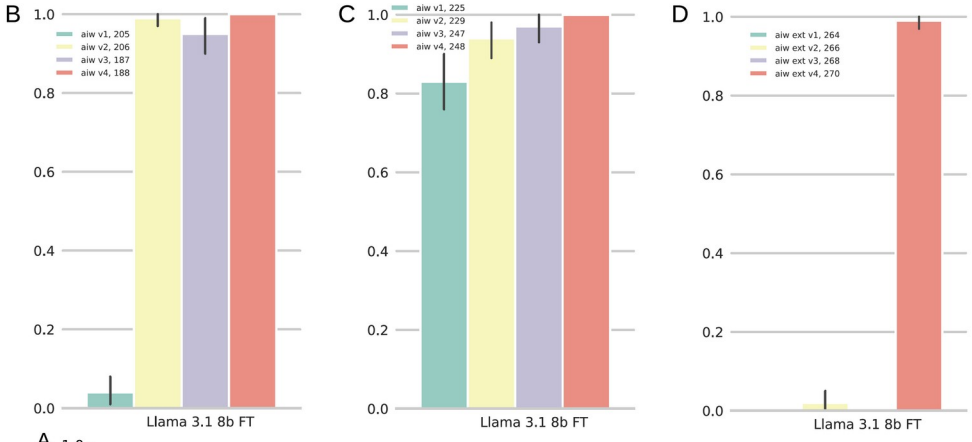
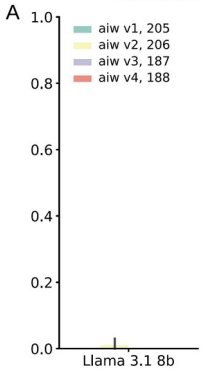


Figure 18: A Tale of Rise and Fall of Claude 3.5 Sonnet. While correct response rates go up close to 1 on **(A)** AIW original and also **(B)** AIW Original Bob version, strong breakdown of corrrect response rates is observed on AIW extension **(C)** (AIW Ext), accompanied with fluctuations across variations 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

Scaling laws: predicting generalization

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

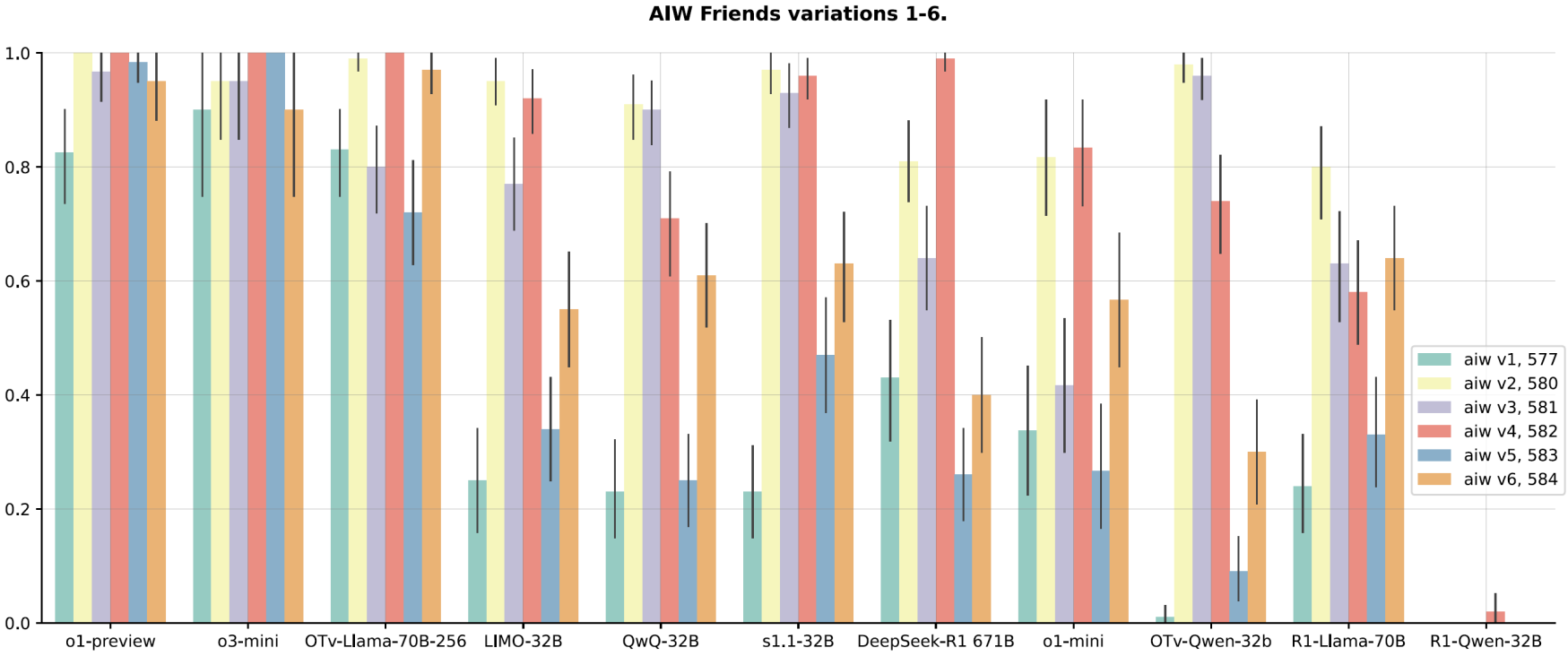
AIW Friends, Variations 1-6, Prompt IDs: 577 580 581 582 583 584

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

- 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

AIW Circles Colleagues, Variations 1-6. THINKING v2, Prompt IDs: 597 599 600 601 602 596

Variation 1: Alice has 3 male colleagues and she also has ***6*** female colleagues. These are all colleagues that Alice has. All these mentioned persons around Alice are colleagues of each other. Bob has 2 female colleagues and 1 male colleague in total. All these mentioned persons around Bob are colleagues of each other. The people in the circle around Bob do not have other colleagues aside - with the only exception of Matilda. She is colleague of Bob and she is also colleague of Alice, being part of Alice's circle. [Correct answer: ***7***]

Variation 2: … [Correct answer: ***3***]

Variation 3: … [Correct answer: ***5***]

Variation 4: … [Correct answer: ***2***]

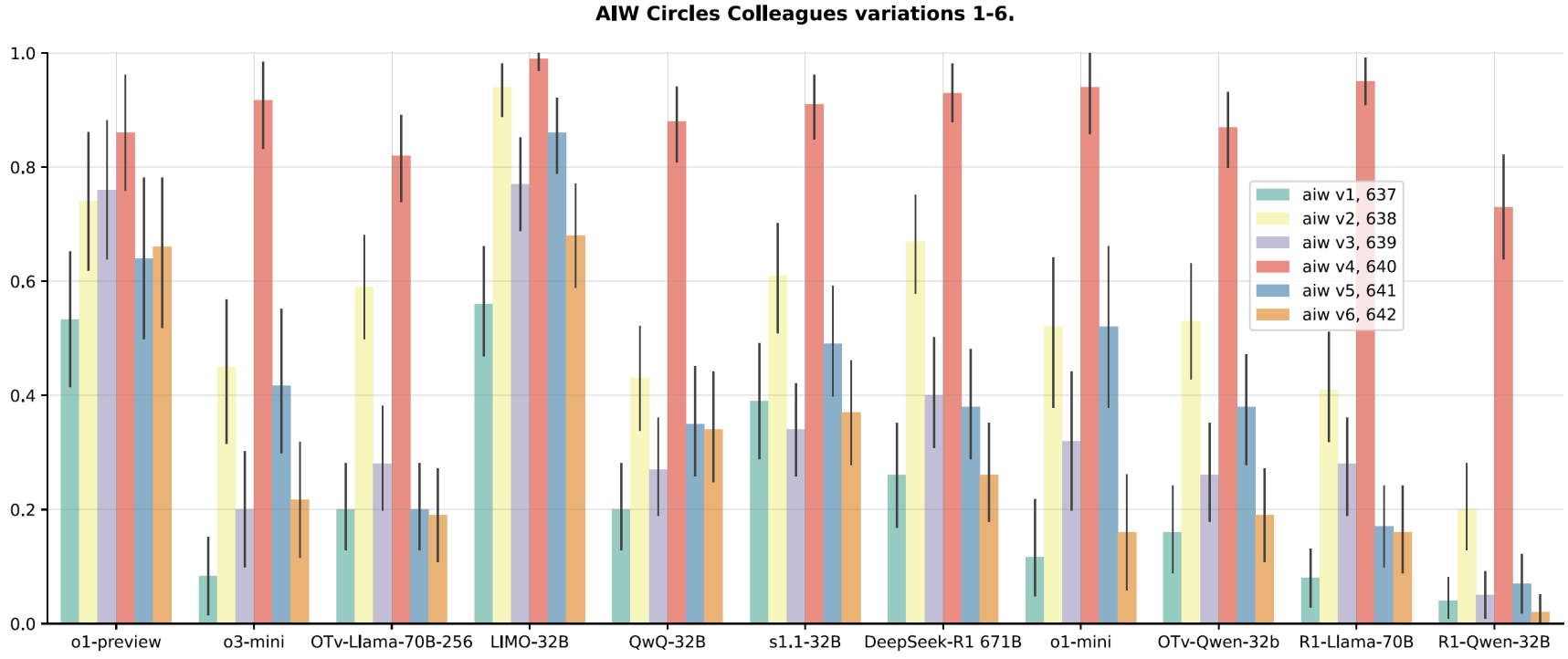
Variation 5: … [Correct answer: ***4***]

Variation 6: Alice has 3 male colleagues and she also has ***5*** female colleagues. These are all colleagues that Alice has. All these mentioned persons around Alice are colleagues of each other. Bob has 2 female colleagues and 1 male colleague in total. All these mentioned persons around Bob are colleagues of each other. The people in the circle around Bob do not have other colleagues aside - with the only exception of Matilda. She is colleague of Bob and she is also colleague of Alice, being part of Alice's circle. [Correct answer: ***6***]

How many female colleagues does Matilda have?

Scaling laws: predicting generalization

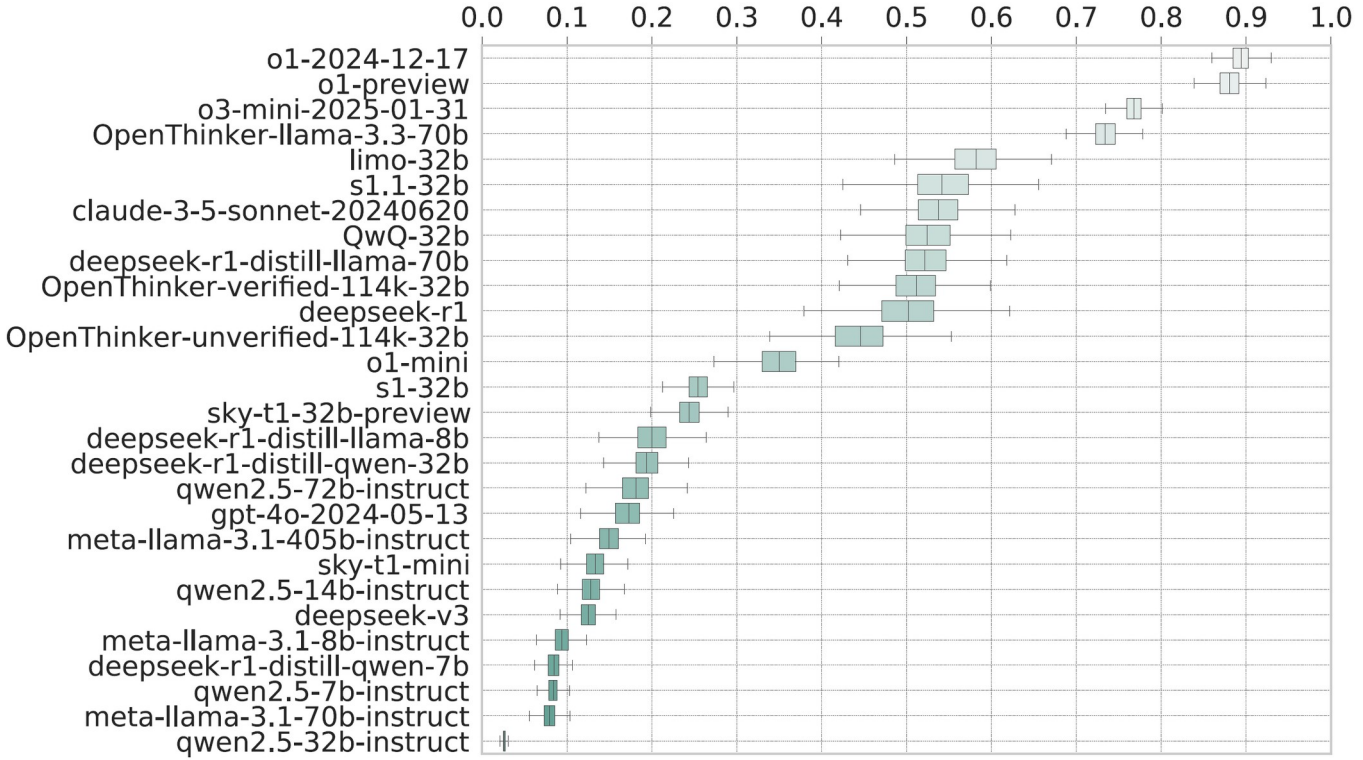
- Reasoning models: Still show strong fluctuations across variations that DO NOT CHANGE problem structure at all



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

Scaling laws: predicting generalization

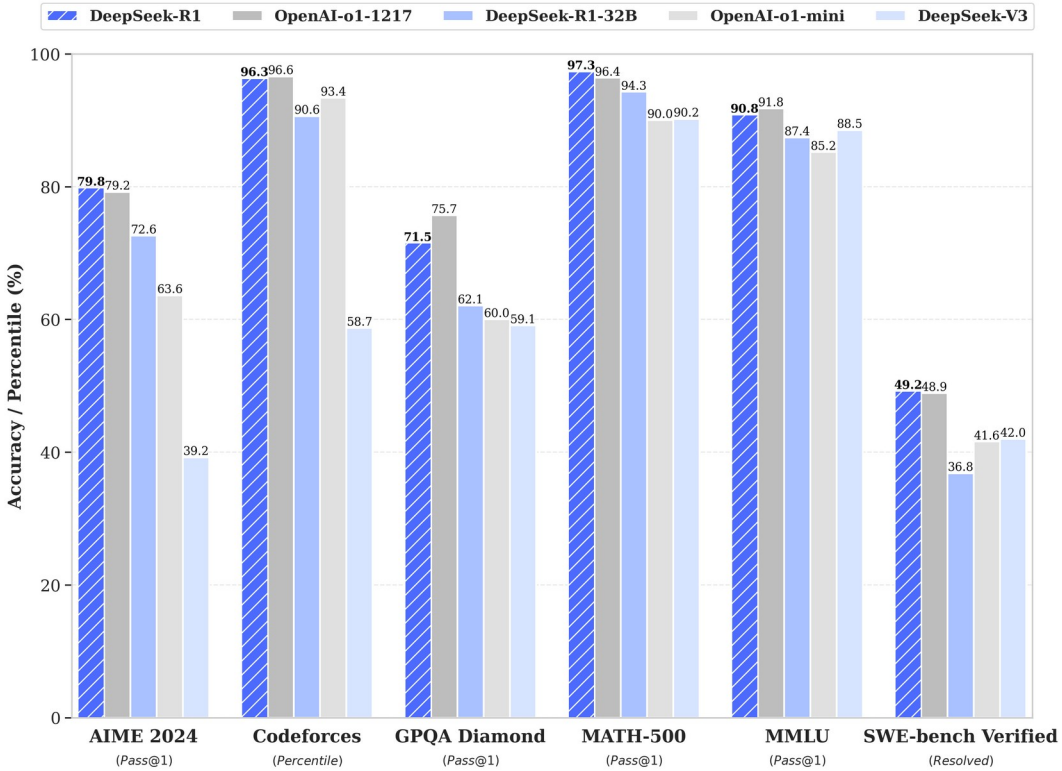
- Reasoning models: clearly better than SOTA LLMs. Still generalization deficits



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



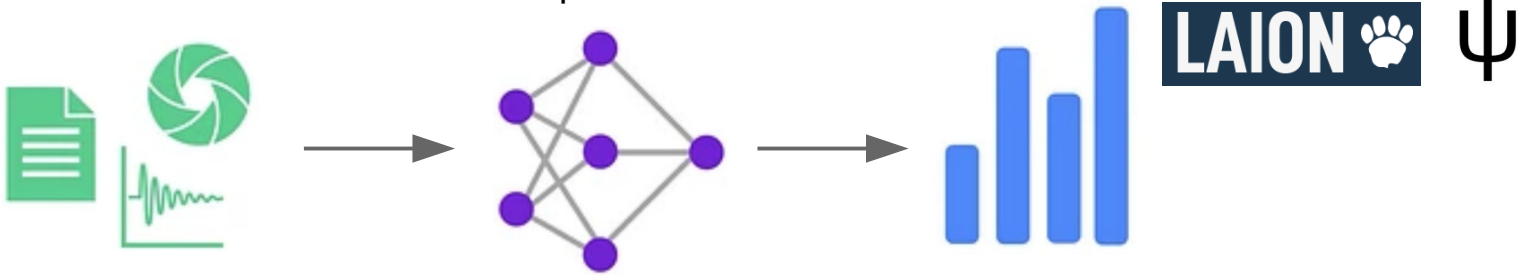
Open foundation models: improving scaling

- 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!

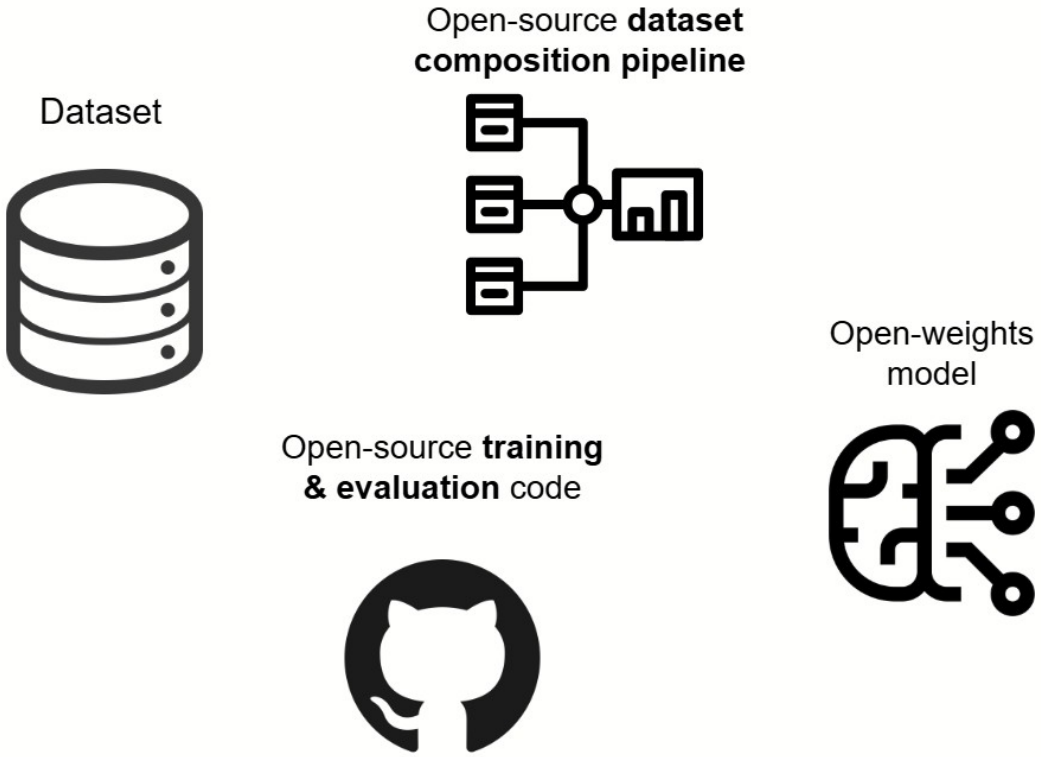
Dataset composition studies, scaling laws

Learning procedure studies, scaling laws

Novel benchmarks for model capabilities, transfer

Open science for large-scale foundation models

- All compents being open is crucial for reproducible science of foundation models



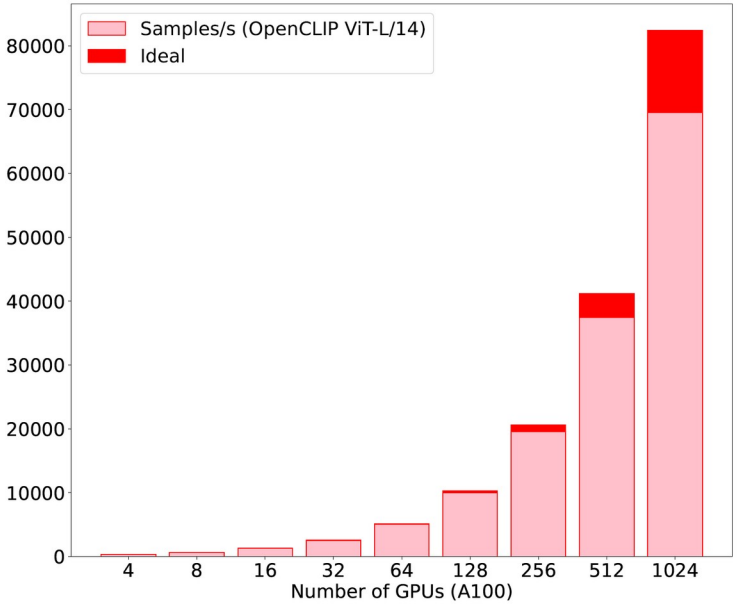
Open science for large-scale foundation models

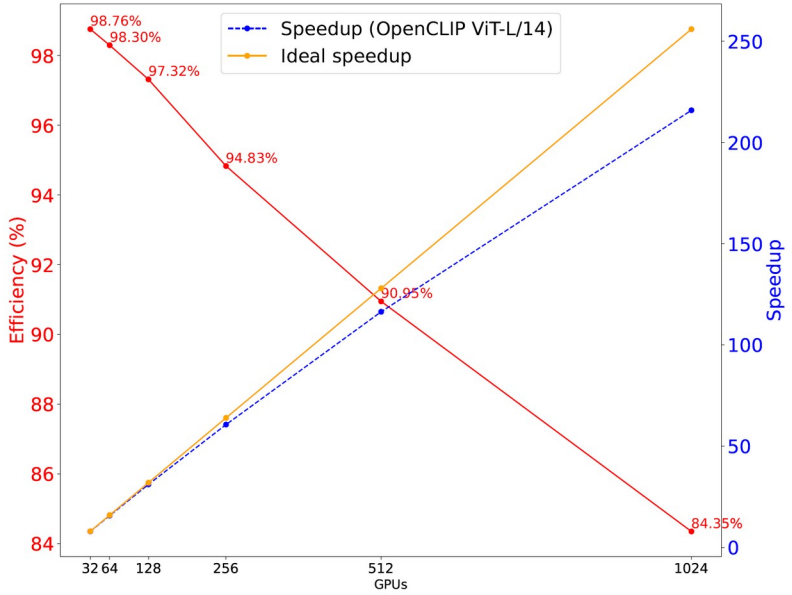
- Open-sourcing whole foundation model research pipeline, case LAION-openCLIP studies

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

Supercomputers for foundation model research

- Supercomputers: scaling law derivation, prediction testing (eg openCLIP ViT L/14: 122 hours with 1024 A100 - total of 124K GPU hours)
- Common effort avoids replication of same expensive measurements





Cherti et al, CVPR 2023

Open science for large-scale foundation models

- **LAION**: **L**arge-scale **A**rtificial **I**ntelligence **O**pen **N**etwork
  - **compute**: applying for publicly funded supercomputers
    - **JUWELS Booster**, Germany: Gauss Center for Supercomputing
    - **Summit**, USA: INCITE Leadership computing call
    - **LUMI** (Finland), **Leonardo** (Italy): Extreme Scale **EuroHPC** call

Open science for large-scale foundation models

- Supercomputers in EU – hubs for large-scale basic AI research
- Open science for advancing powerful, safe generic AI tools for public

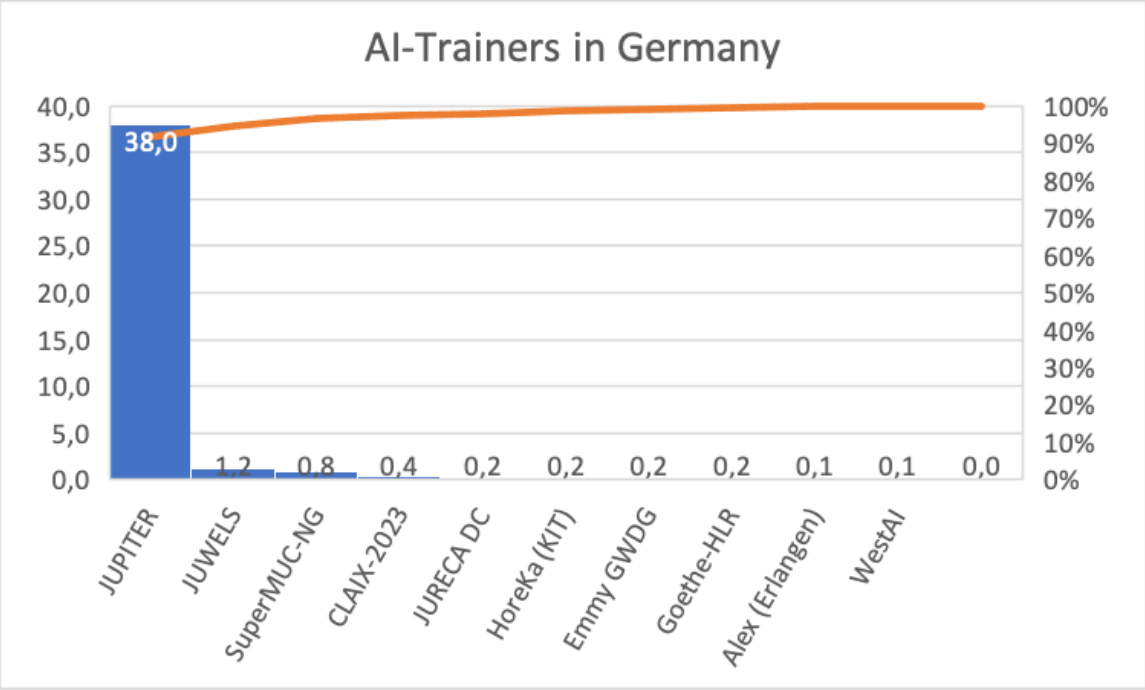
*Stable Diffusion 1.5, trained on **LAION-5B** image-text dataset.*
Prompt: "An epic scene of a supercomputing center building of the future, embedded in a rich wild green exotic blooming jungle forest, nearby a lake“

Open science for large-scale foundation models

- **Compute**: 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



Open foundation models: outlook

- „Moonshot“: **open-sci-MM – strong open multi-modal foundation model family, learning with any modality – text, code, tables, 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 resource acquisition** for open source foundation models and datasets research and development
- Should be driven by orgas with strong track of record researching and building FoMos
  - HuggingFace, TogetherAI, EleutherAI, LAION, ...
- Common grant applications for compute and fund resources
  - Securing resources for building and maintaining important artefacts
- Define important model and dataset artefacts to be maintained as open-source
- Previous examples:
  - OpenCLIP (LAION-400M/5B, openCLIP ViT-B-32, openCLIP-ViT-H-14, ...)
  - Stable Diffusion (LAION-400M/5B, SD 1.5)
  - DCLM (DCLM-baselines, DCLM-1B/3B/7B)
- Current examples: OpenThoughts, Open-R1 (creating an open source version of DeepSeek R1/o4 level reasoning models)

Open foundation models: outlook

- „Moonshot“: **open-sci-MM - open multi-modal foundation model family**
  - Identifying strongest candidates via scaling law derivation based search

- **OpenEuroLLM, MINERVA, ... – LAION/ELLIS/HPLT & friends** : EU consortia for building open foundation models with strongly improved generalization & reasoning
  - **Hiring** - **Join us**! Multiple open ML reseacher (junior/senior postdoc levels), large scale machine learning engineer, science managers/administrators positions open (drop a message j.jitsev@fz-juelich.de)

LAION
Large-scale Artificial Intelligence Open Network

MINERVA

OPEN EURO LLM

JÜLICH Forschungszentrum | JÜLICH SUPERCOMPUTING CENTRE

ellis
European Laboratory for Learning and Intelligent Systems

open-sci collective

Tübingen AI Center

ellis INSTITUTE TÜBINGEN

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

ellis
European Laboratory for Learning and Intelligent Systems

ellis INSTITUTE TÜBINGEN

Tübingen AI Center

JÜLICH Forschungszentrum

JÜLICH SUPERCOMPUTING CENTRE

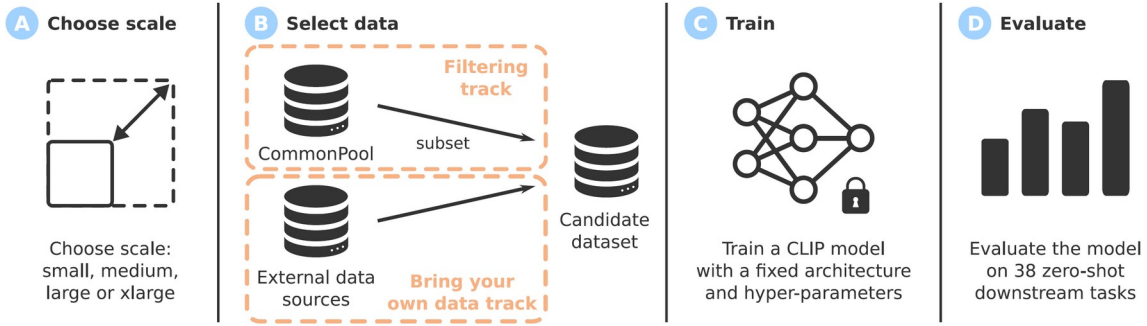
WEST AI KI-Servicezentrum

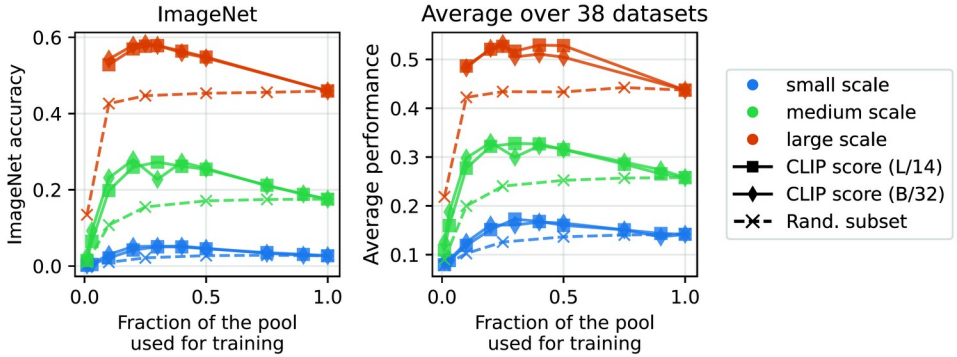

Thanks
for
your
Attention

Supplementary Material

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

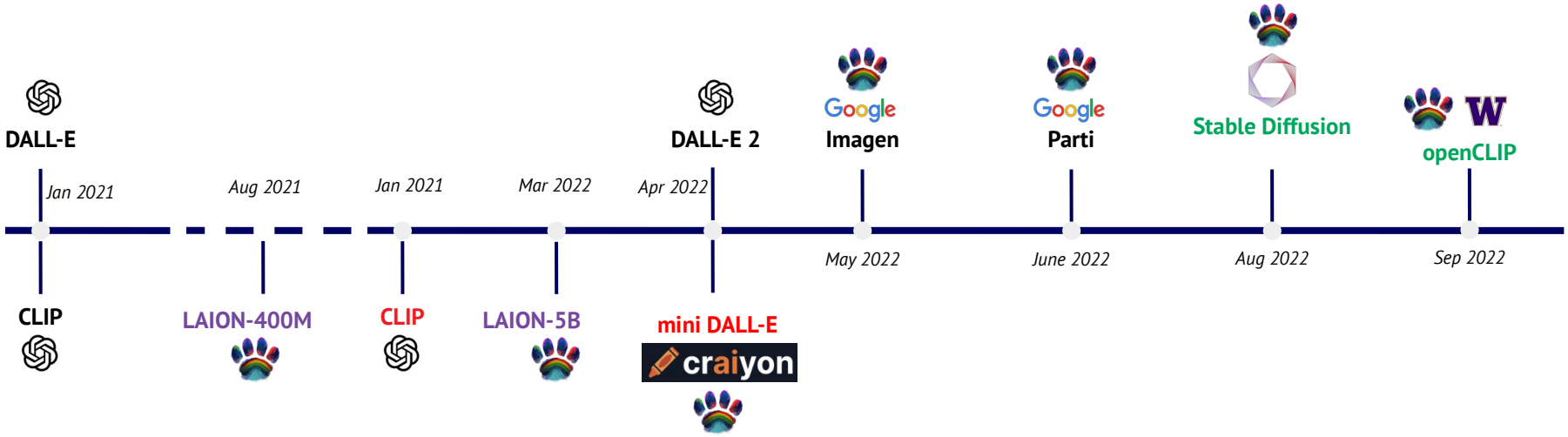
Open foundation models: reproducibility

- 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

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-source releases: millions of downloads of pre-trained models

OpenCLIP DataComp

OpenCLIP LAION-2B

CLAP: Contrastive Language-Audio Pretraining

OpenCLIP LAION-2B

OpenCLIP models trained on LAION-2B

laion/CLIP-ViT-bigG-14-laion2B-39B-b160k
Zero-Shot Image Classification • Updated Jan 16 • ↓ 415k • ♡ 226

laion/CLIP-ViT-g-14-laion2B-s34B-b88K
Zero-Shot Image Classification • Updated Mar 22 • ↓ 13.7k • ♡ 18

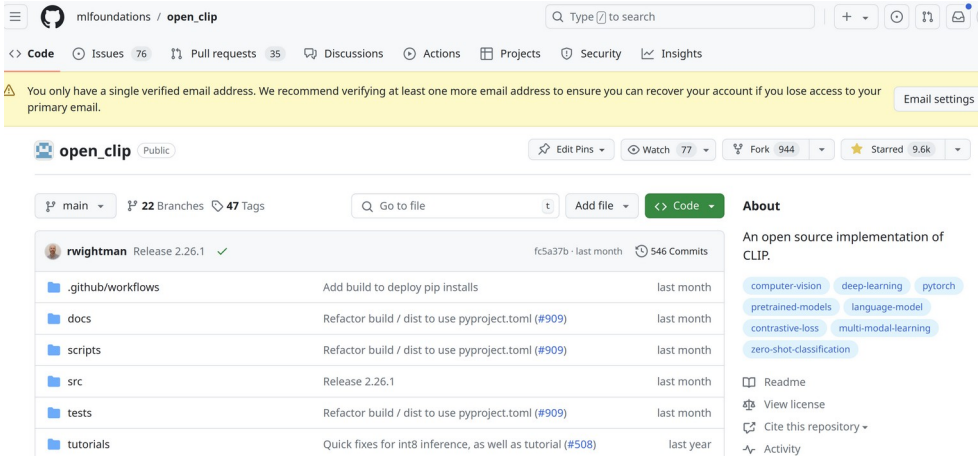
laion/CLIP-ViT-g-14-laion2B-s12B-b42K
Updated Feb 23 • ↓ 38.2k • ♡ 39

laion/CLIP-ViT-H-14-laion2B-s32B-b79K
Zero-Shot Image Classification • Updated Jan 16 • ↓ 973k • ♡ 305

laion/CLIP-ViT-L-14-laion2B-s32B-b82K
Zero-Shot Image Classification • Updated Jan 16 • ↓ 80k • ♡ 43

laion/CLIP-ViT-B-16-laion2B-s34B-b88K
Zero-Shot Image Classification • Updated Apr 19, 2023 • ↓ 5.81M • ♡ 27

laion/CLIP-ViT-B-32-laion2B-s34B-b79K
Zero-Shot Image Classification • Updated Jan 15 • ↓ 1.58M • ♡ 89



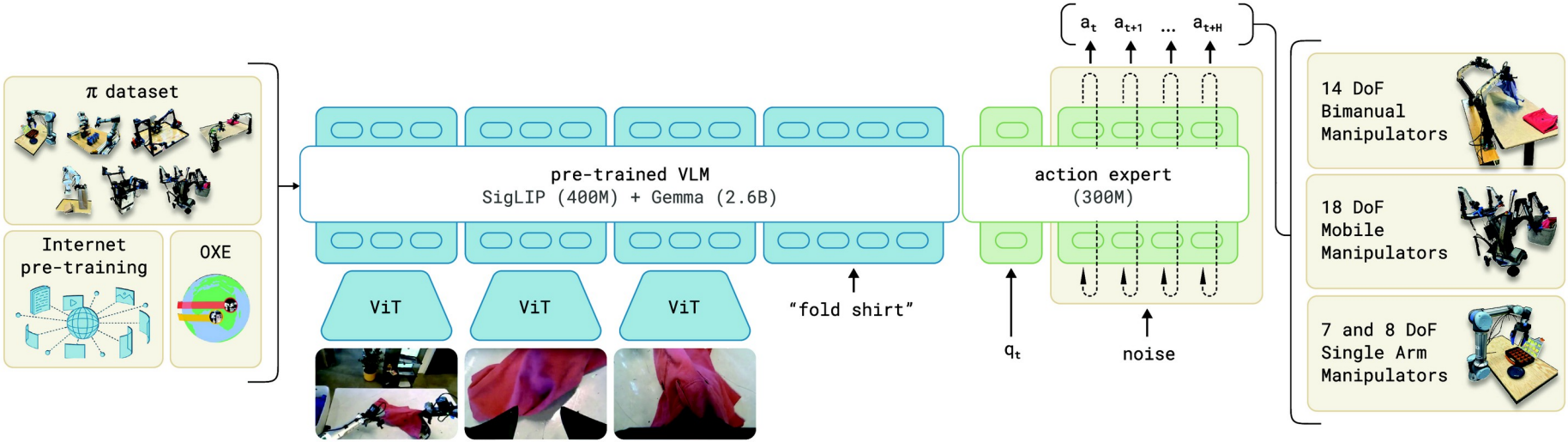
Open science for large-scale foundation models

- 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