JÜLICH
Forschungszentrum
JÜLICH
SUPERCOMPUTING
CENTRE

open-sci
collective

Tübingen AI Center

WEST AI
KI-Servicezentrum

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)

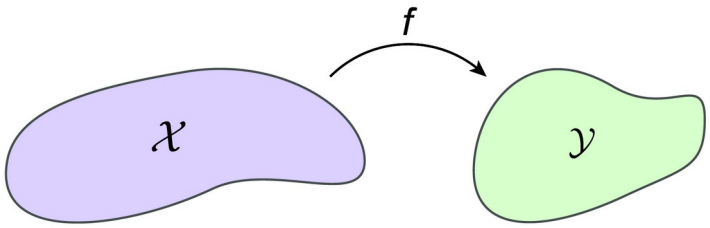
04. Februar 2025 | Marianna Nezhurina, Jenia Jitsev

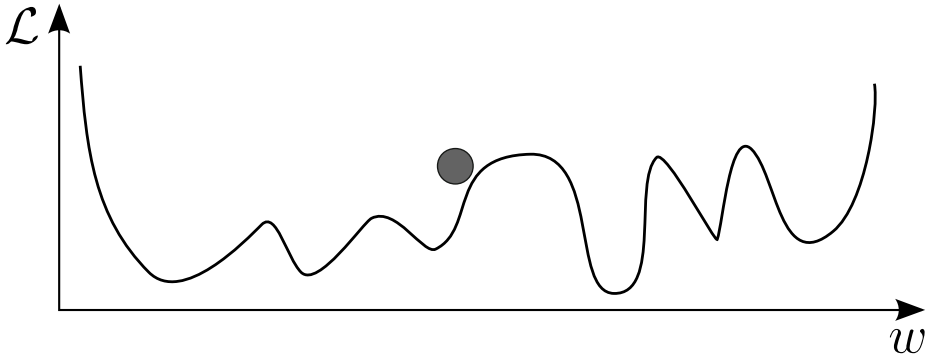
Foundation models: generic transferable learning

Modern Machine Learning

Optimizing loss (objective) of a (complex) model f using (a lot of) data $\mathcal{D}$

- a (complex) model: function (or distribution) family $f(X;\theta)$ ($p(X;\theta)$)
  - parameters θ (often **W** is used) are to adapt ("fit") given the data samples $X \in \hat{\mathcal{D}}$
- optimization:
  - defining a loss $\mathcal{L}(f(X;\theta), \hat{\mathcal{D}})$
  - loss $\mathcal{L}$: measure of quality ("fit") of $f(X;\theta)$ in terms of a task solution on $\hat{\mathcal{D}}$
  - seeking to minimize $\mathcal{L}(f, \hat{\mathcal{D}})$ with respect to all possible $\hat{\mathcal{D}} \sim P(\mathcal{D})$!





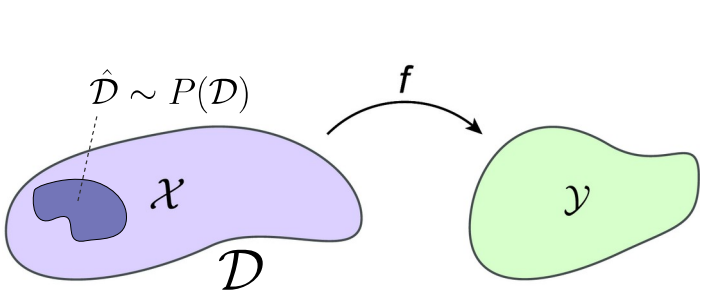
Foundation models: generic transferable learning

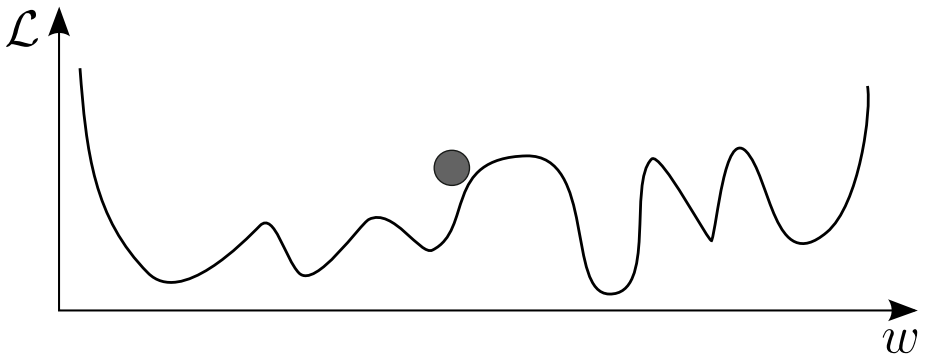
Modern Machine Learning

- General formulation: estimate unknown true data density $P(X)$ from observed data $\hat{\mathcal{D}}$
- Optimizing loss (objective) of a (complex) model f using (a lot of) generic data $\mathcal{D}$

- model $f(X;\theta)$: recipes for "solutions" to "problems" posed by $\mathcal{L}$
- optimization: looking for a "good" model f^* by minimizing $\mathcal{L}(f,\hat{\mathcal{D}})$ for all possible $\hat{\mathcal{D}} \sim P(\mathcal{D})$!
  - general principle of expected risk minimization:

$$f^* = \arg\min_f \mathbf{R}(f) = \mathbb{E}_{\hat{\mathcal{D}}\sim P(\mathcal{D})}[\mathcal{L}(f,\hat{\mathcal{D}})] = \int \mathcal{L}(f,\hat{\mathcal{D}})\, P(\hat{\mathcal{D}})\, d\hat{\mathcal{D}}$$





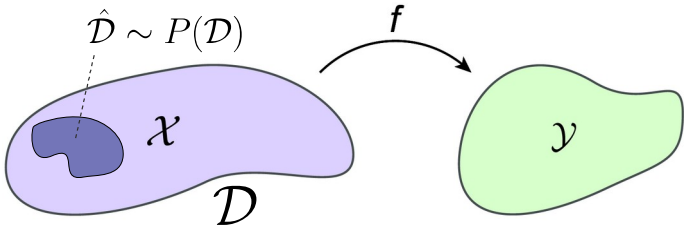
Foundation models: generic transferable learning

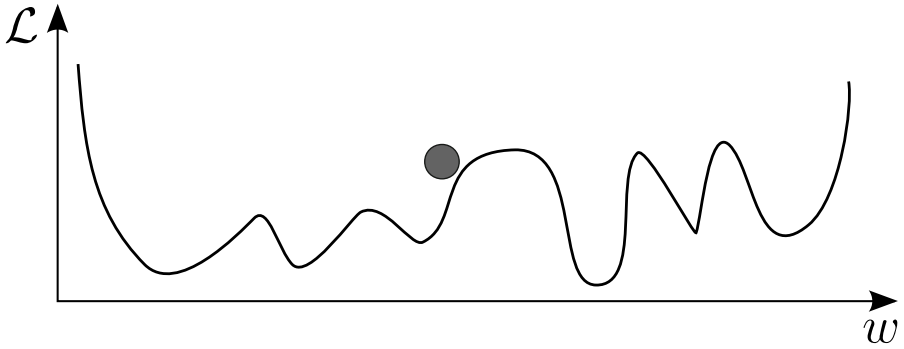
Important

Estimate loss $\mathcal{L}(f(X;\theta),\mathcal{D})$ on data $\mathcal{D}$ yet unseen!

- estimating "true" $\mathcal{L}$, expected risk: **generalization** error estimation
  - Challenge: estimate based on observed data only – "true" $\mathcal{L}$, true loss landscape, true data distribution $P(\mathcal{D})$ unknown
  - Model should perform tasks well not only on observed $\hat{\mathcal{D}}$, but on unseen data and conditions -> **Strong generalization**

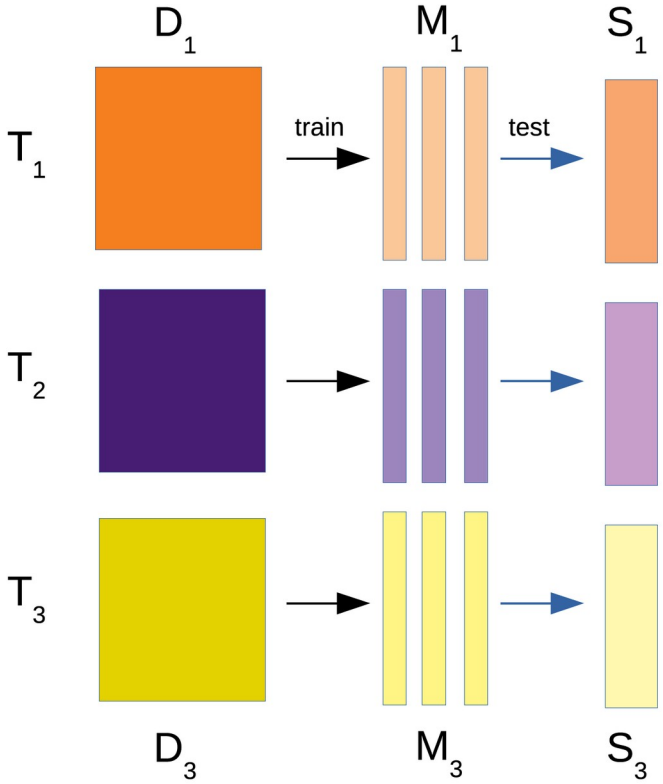
$$f^* = \arg\min_f \mathbf{R}(f) = \mathbb{E}_{\hat{\mathcal{D}}\sim P(\mathcal{D})}[\mathcal{L}(f,\hat{\mathcal{D}})] = \int \mathcal{L}(f,\hat{\mathcal{D}})\, P(\hat{\mathcal{D}})\, d\hat{\mathcal{D}}$$





Foundation models: generic transferable learning

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

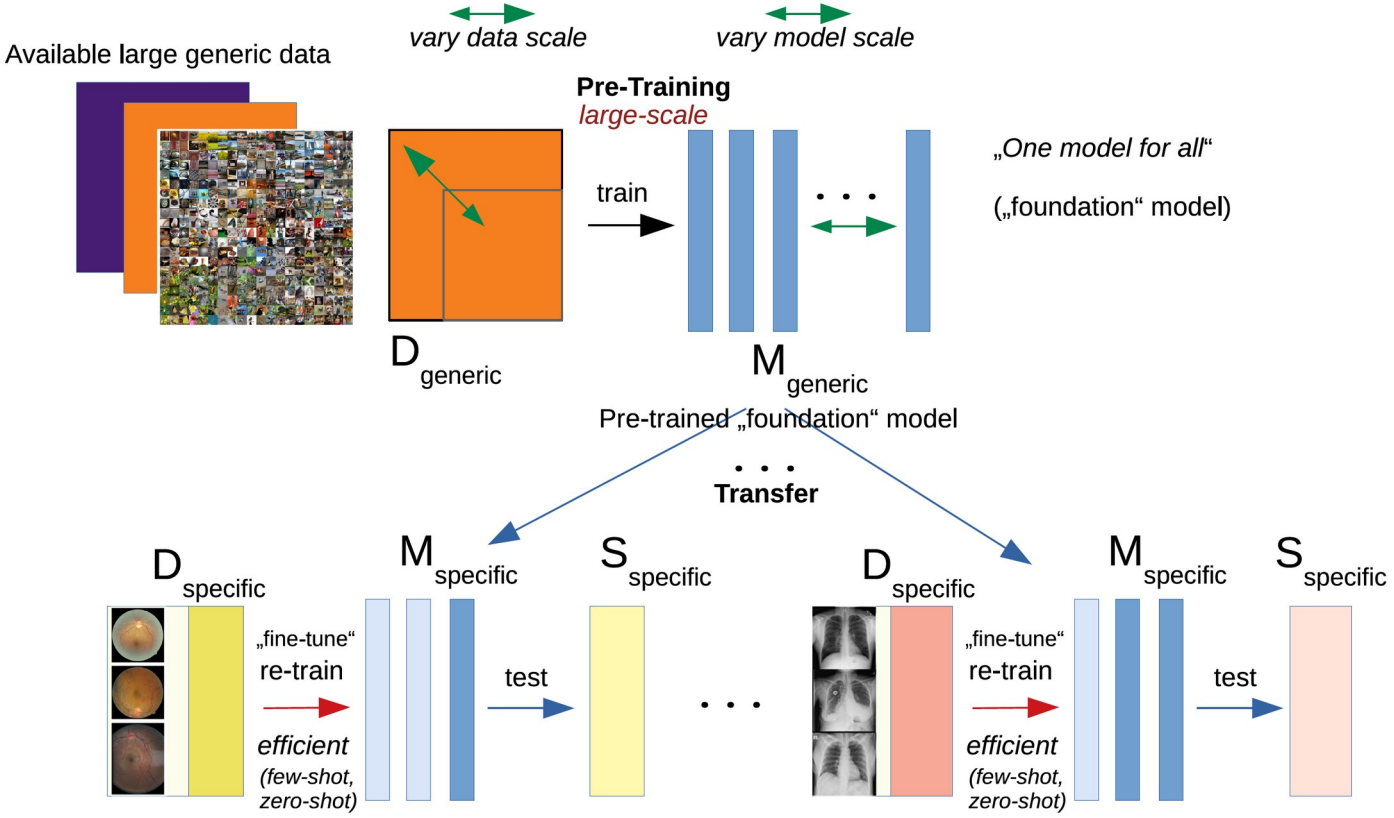


Long training

Slow
Large (Labeled) Data

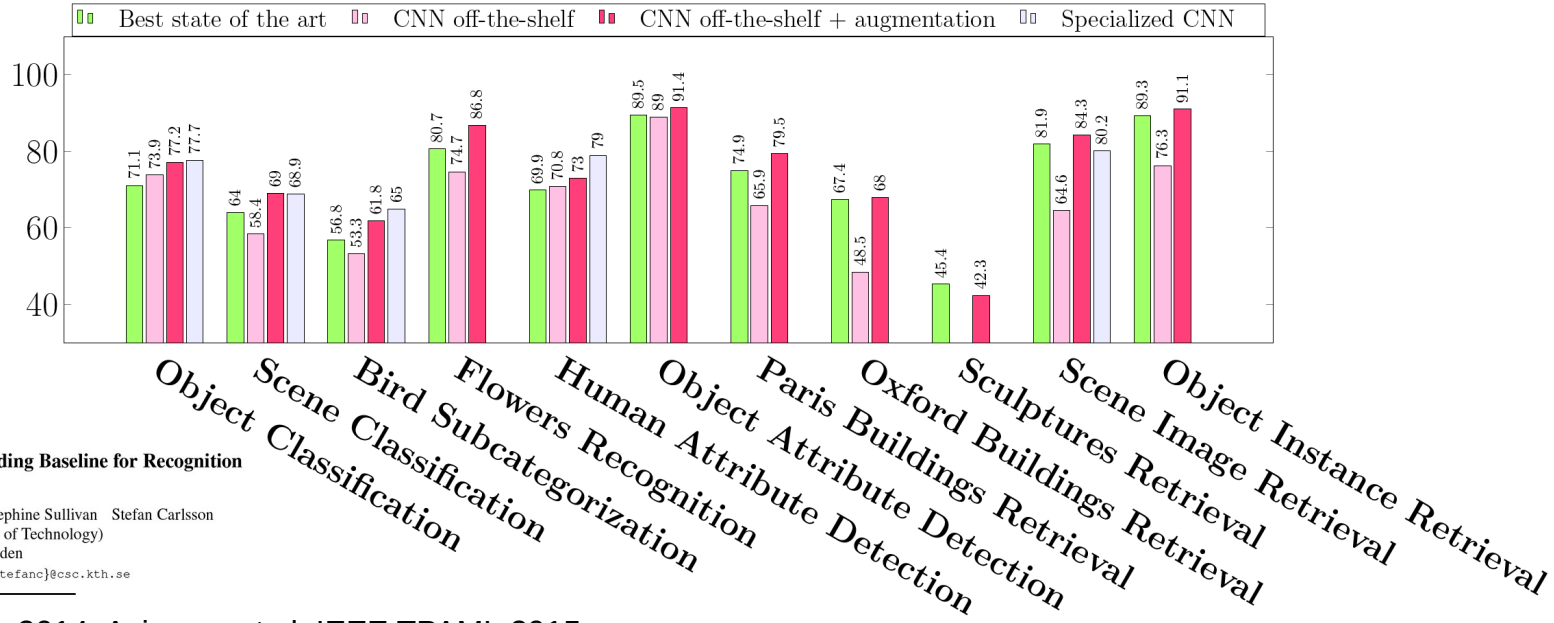
Foundation models: generic transferable learning

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



Foundation models: generic transferable learning

- **Strong transferability**: evidence for early convolutional networks (OverFeat, VGG16) dating back to 2013
- „Off-the-shelf“ **transferable models**: ConvNets (CNNs) pre-trained on ImageNet-1k (1.4M images), 2012-2017 (eg ResNet – Winner ILSVRC 2015)



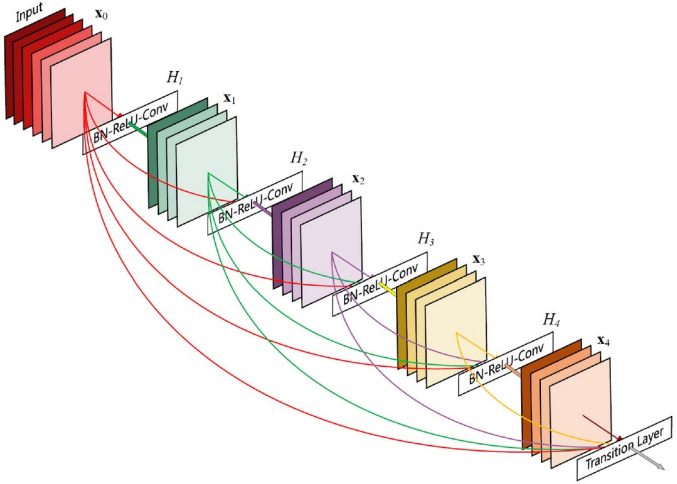
CNN Features off-the-shelf: an Astounding Baseline for Recognition

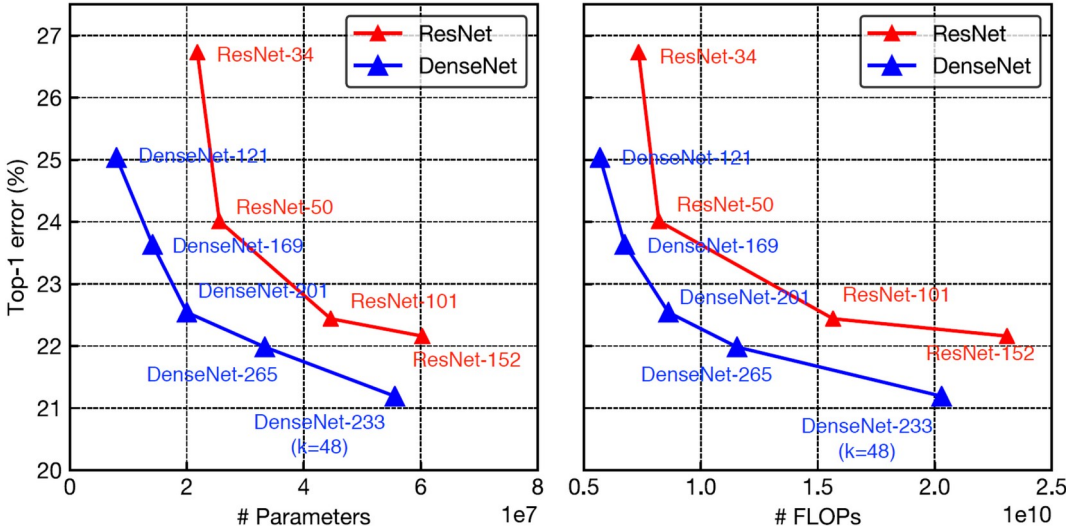
Ali Sharif Razavian Hossein Azizpour Josephine Sullivan Stefan Carlsson
CVAP, KTH (Royal Institute of Technology)
Stockholm, Sweden
{razavian,azizpour,sullivan,stefanc}@csc.kth.se

Razavian et al, CVPR, 2014; Azizpour et al, IEEE TPAMI, 2015

Foundation models: generic transferable learning

- Evidence for scaling improving model function (here: ImageNet classification)
- Model comparison (DenseNet vs ResNet) via scaling behavior (function vs FLOPs)





Huang et al, CVPR, 2017

Foundation models: generic transferable learning

- „Primordial“ scaling law observation across various learning procedures

DEEP LEARNING SCALING IS PREDICTABLE, EMPIRICALLY

Joel Hestness, Sharan Narang, Newsha Ardalani, Gregory Diamos, Heewoo Jun,
Hassan Kianinejad, Md. Mostofa Ali Patwary, Yang Yang, Yanqi Zhou

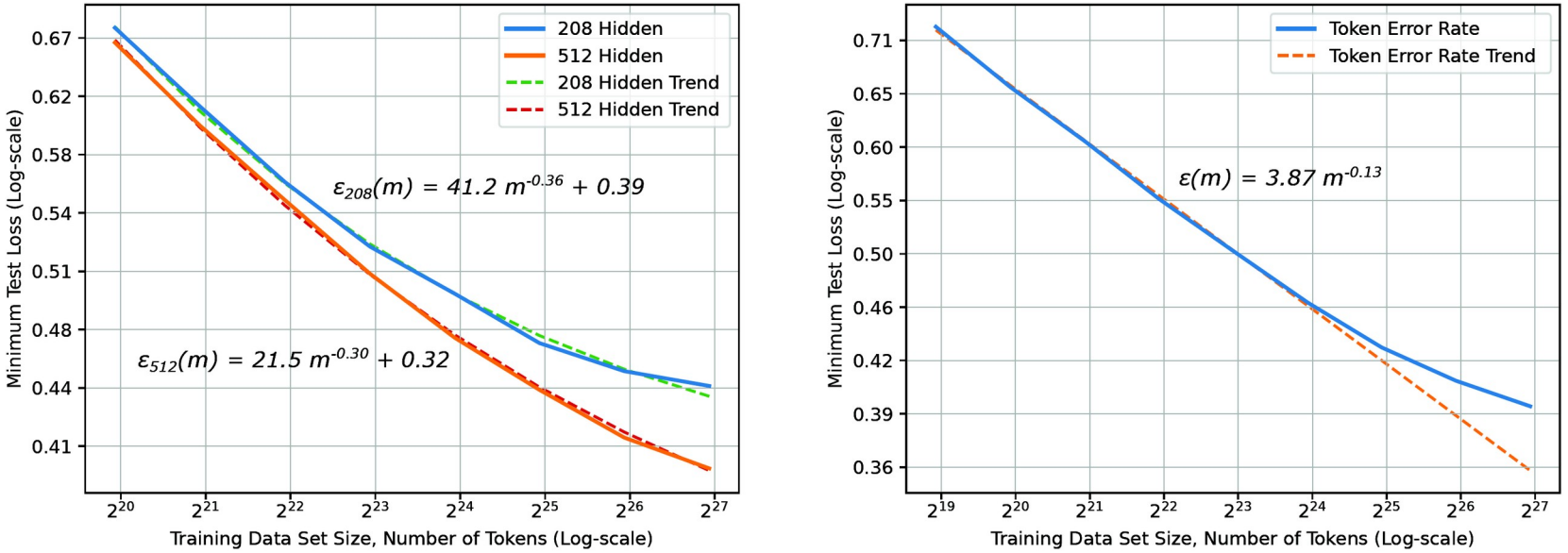
{joel,sharan,ardalaninewsha,gregdiamos,junheewoo,hassankianinejad,
patwarymostofa,yangyang62,zhouyanqi}@baidu.com

Baidu Research

Hestness et al, arXiv:1712.00409, 2017

Foundation models: generic transferable learning

- „Primordial“ scaling law observations: language translation

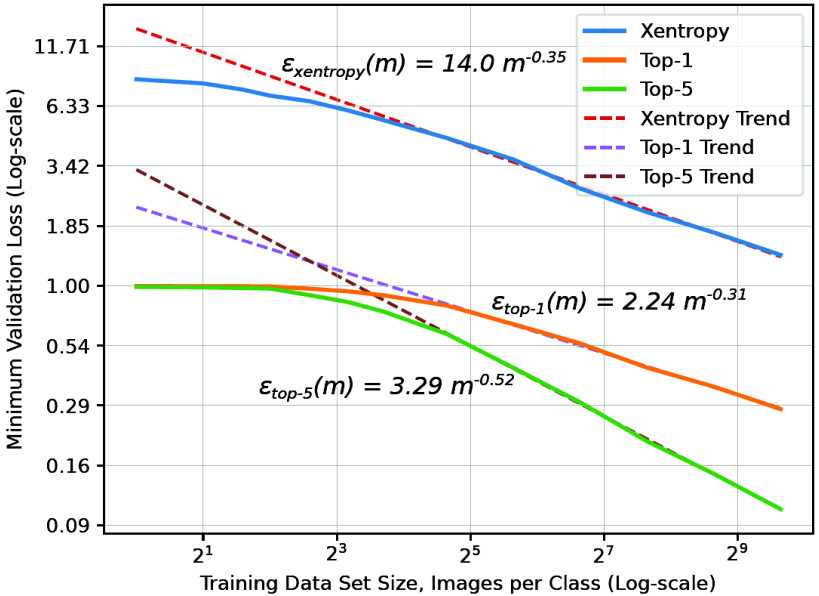


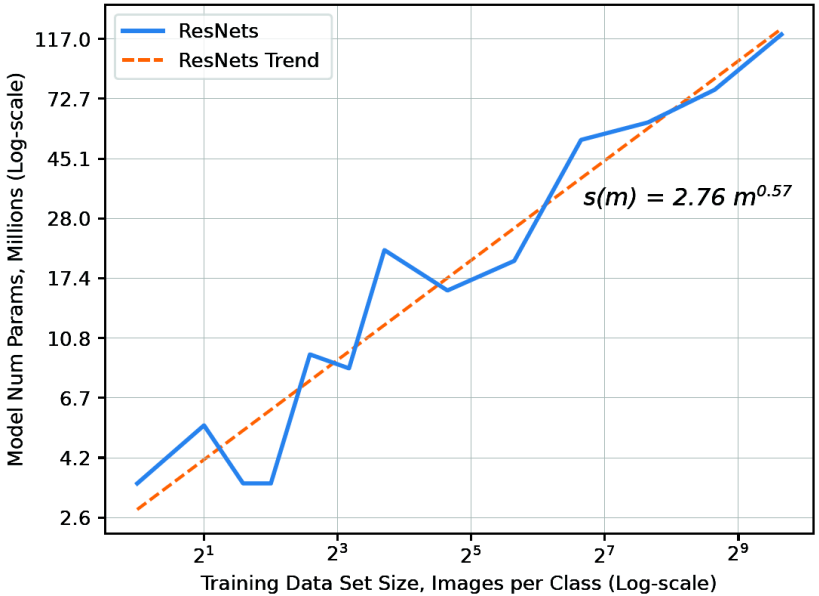
Neural machine translation learning curves. Left: the learning curves for separate models
Right: composite learning curve of best-fit model at each data set size.

Hestness et al, arXiv:1712.00409, 2017

Foundation models: generic transferable learning

- „Primordial“ scaling law observations: image classification (ResNet supervised)



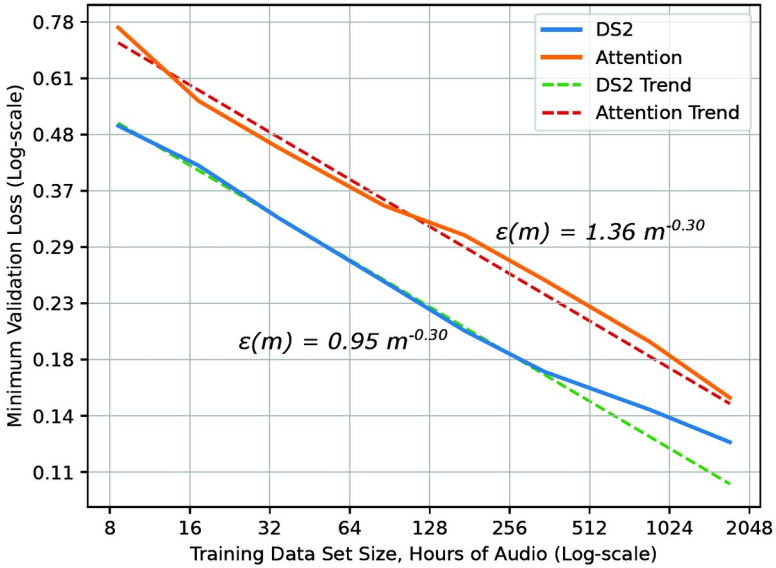


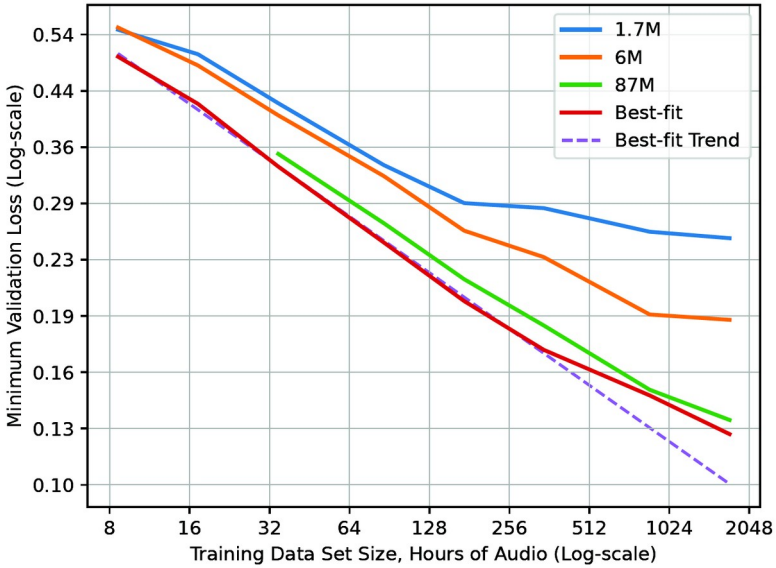
Learning curve and model size results and trends for ResNet image classification.

Hestness et al, arXiv:1712.00409, 2017

Foundation models: generic transferable learning

- „Primordial“ scaling law observations: audio speech-to-text translation





Learning curves for DS2 and attention speech models (left), and learning curves for various DS2 model sizes, 1.7M to 87M parameters (right).

Hestness et al, arXiv:1712.00409, 2017

Foundation models: generic transferable learning

- „Primordial“ scaling law observation: power law dependency of generalization error on training resources (here data samples seen)

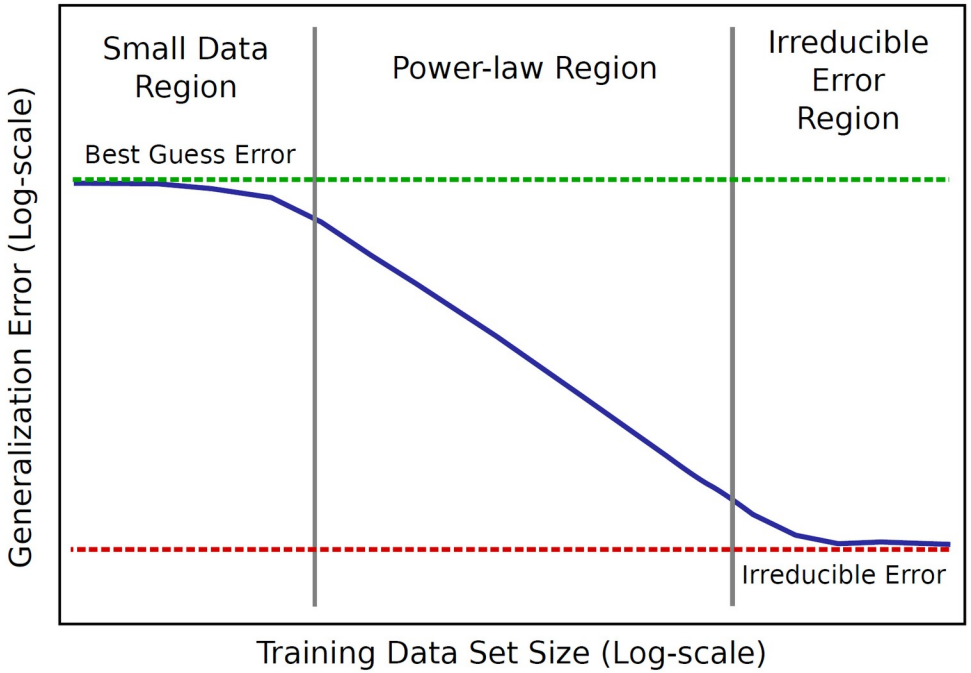
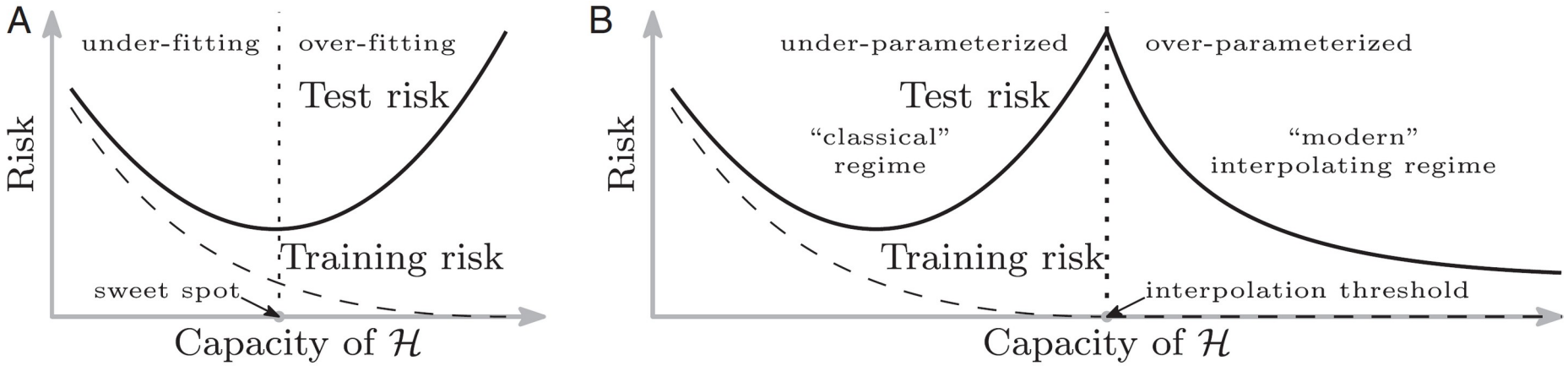


Figure 6: Sketch of power-law learning curves

Hestness et al, arXiv:1712.00409, 2017

Foundation models: generic transferable learning

- Reconciling generalization – large, overparameterized models generalize strongly



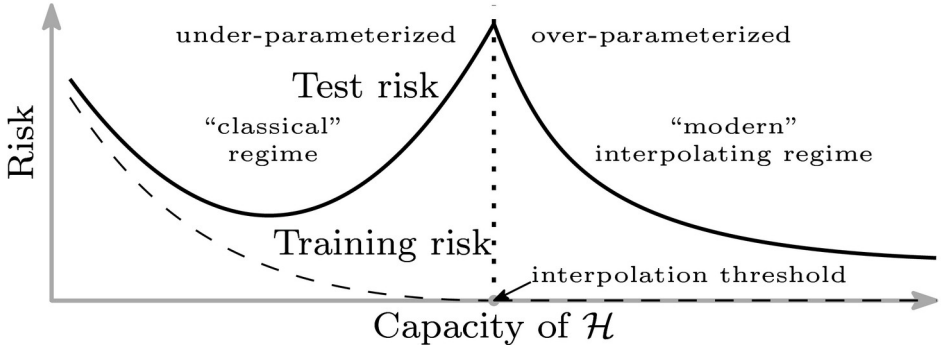
Reconciling modern machine-learning practice and the classical bias–variance trade-off

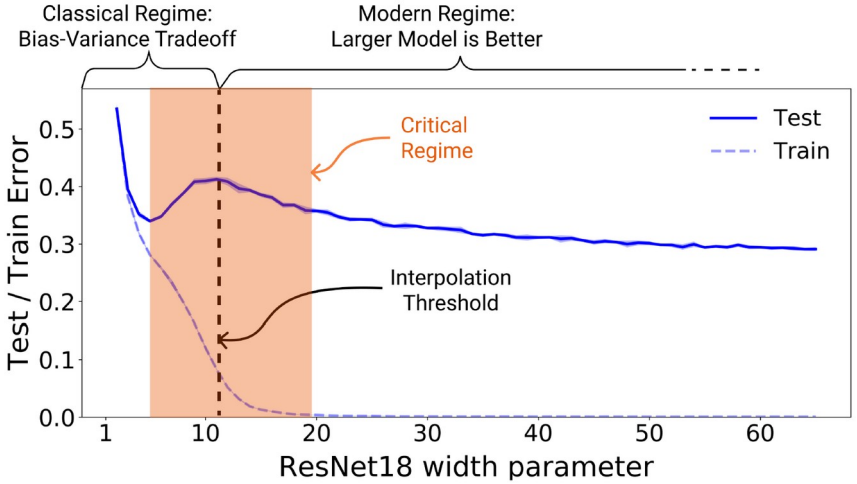
Mikhail Belkin[a,b,1], Daniel Hsu[c], Siyuan Ma[a], and Soumik Mandal[a]

Belkin et al, PNAS, 2019

Foundation models: generic transferable learning

- Reconciling generalization – large, overparameterized models generalize strongly
- Double, triple descent phenomena observed in various experiments

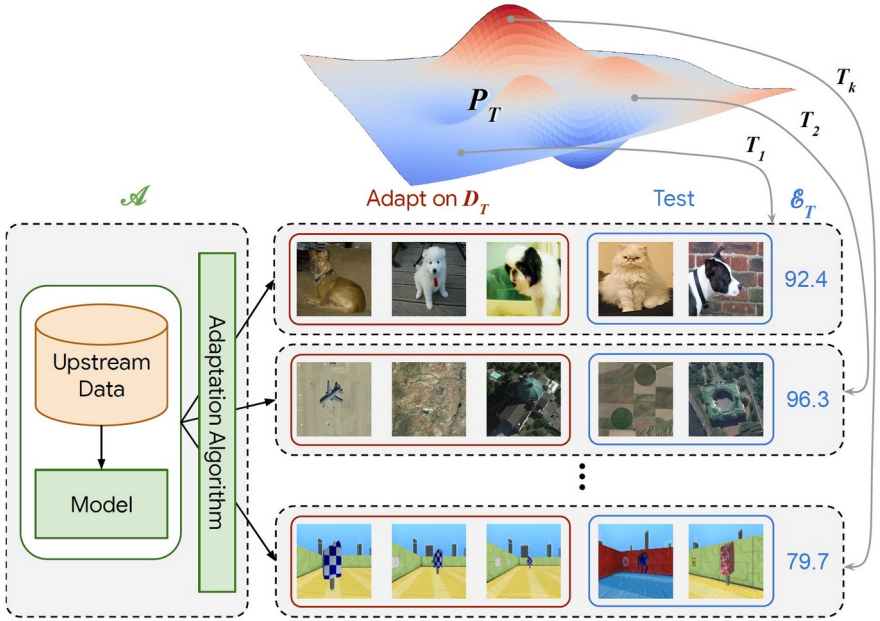




Belkin et al, PNAS, 2019; Nakkiran et al, ICLR, 2020

Foundation models: generic transferable learning

- V-TAB: testing transfer from pre-trained models vs training from scratch

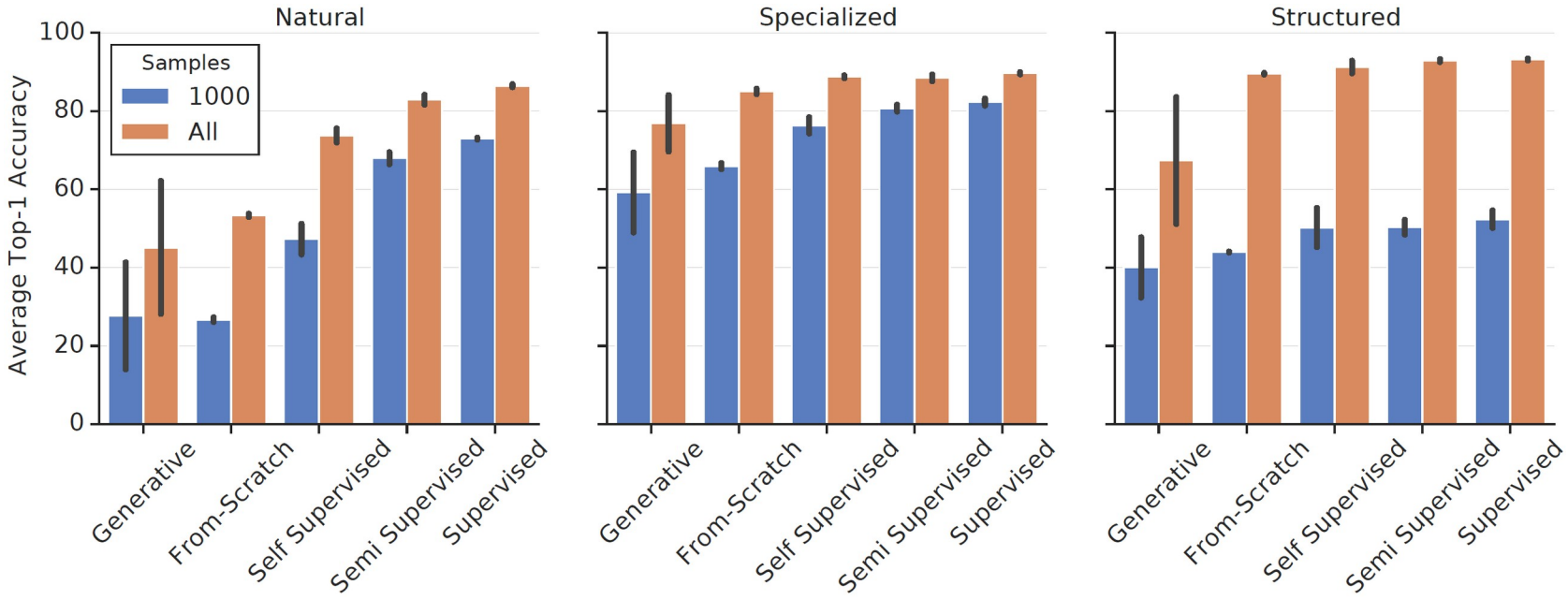


Category	Dataset	Train size	Classes	Reference
Natural	Caltech101	3,060	102	(Li et al., 2006)
Natural	CIFAR-100	50,000	100	(Krizhevsky, 2009)
Natural	DTD	3,760	47	(Cimpoi et al., 2014)
Natural	Flowers102	2,040	102	(Nilsback & Zisserman, 2008)
Natural	Pets	3,680	37	(Parkhi et al., 2012)
Natural	Sun397	87,003	397	(Xiao et al., 2010)
Natural	SVHN	73,257	10	(Netzer et al., 2011)
Specialized	EuroSAT	21,600	10	(Helber et al., 2019)
Specialized	Resisc45	25,200	45	(Cheng et al., 2017)
Specialized	Patch Camelyon	294,912	2	(Veeling et al., 2018)
Specialized	Retinopathy	46,032	5	(Kaggle & EyePacs, 2015)
Structured	Clevr/count	70,000	8	(Johnson et al., 2017)
Structured	Clevr/distance	70,000	6	(Johnson et al., 2017)
Structured	dSprites/location	663,552	16	(Matthey et al., 2017)
Structured	dSprites/orientation	663,552	16	(Matthey et al., 2017)
Structured	SmallNORB/azimuth	36,450	18	(LeCun et al., 2004)
Structured	SmallNORB/elevation	36,450	9	(LeCun et al., 2004)
Structured	DMLab	88,178	6	(Beattie et al., 2016)
Structured	KITTI/distance	5,711	4	(Geiger et al., 2013)

Zhai et al, arXiv:1910.04867, 2019

Foundation models: generic transferable learning

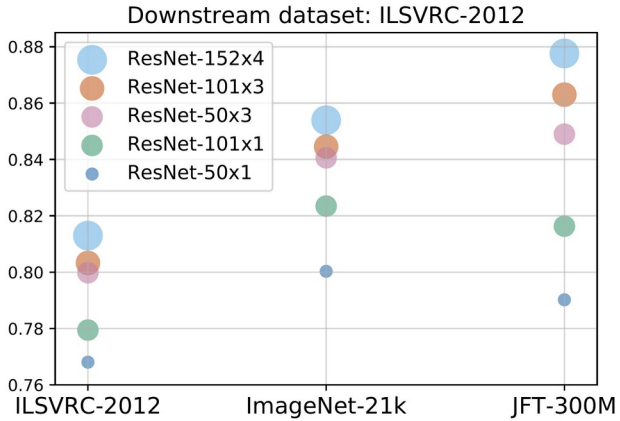
- V-TAB: testing transfer from pre-trained models vs training from scratch

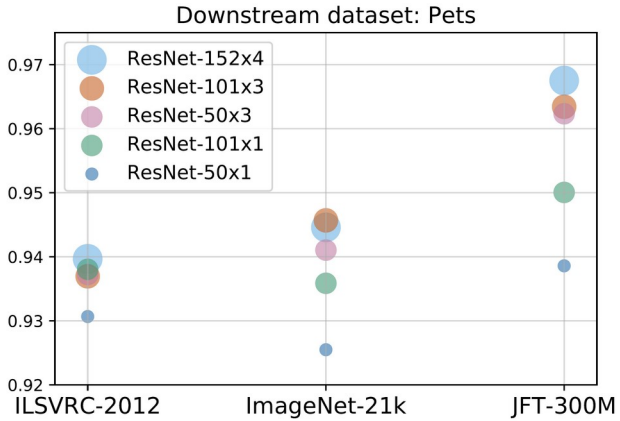


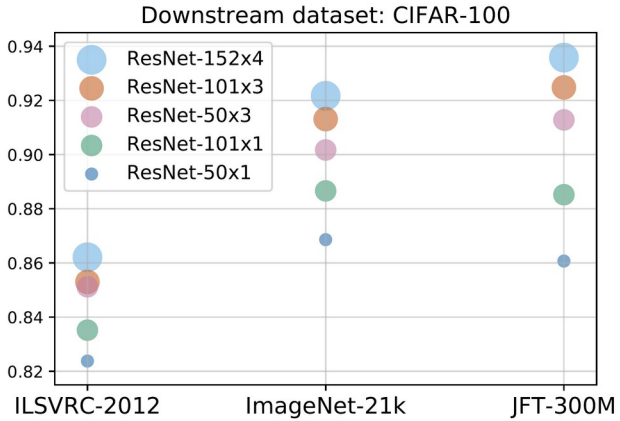
Zhai et al, arXiv:1910.04867, 2019

Foundation models: generic transferable learning

- Big Transfer: evidence for scaling improving transfer learning





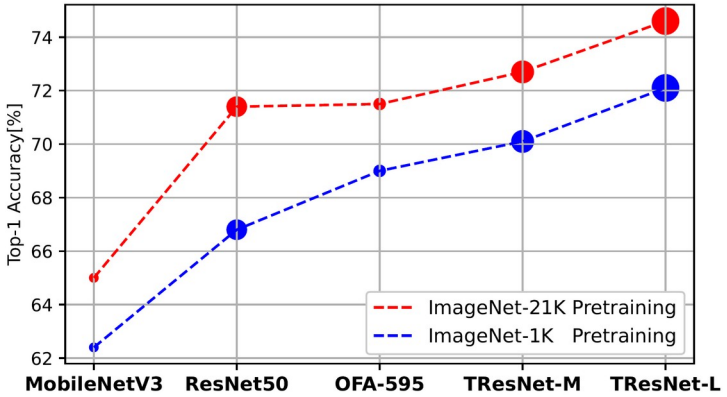


BiT - Big Transfer, Kolesnikov et al, ECCV, 2020

Foundation models: generic transferable learning

- ImageNet-21k for the masses: evidence for scaling improving transfer learning

Downstream Task	Dataset	MobileNetV3		OFA-595		ResNet50		TResNet-M		TResNet-L	
		1K	21K	1K	21K	1K	21K	1K	21K	1K	21K
Single-label Classification	iNaturalist[1]	62.4	**65.0**	69.0	**71.5**	66.8	**71.4**	70.1	**72.7**	72.4	**74.8**
	CIFAR100[1]	86.7	**88.5**	88.3	**90.3**	86.8	**90.3**	89.5	**91.7**	90.2	**92.5**
	Food 251[1]	70.1	**70.3**	72.9	**73.5**	72.2	**74.0**	75.1	**76.1**	76.3	**77.0**
Multi-label Classification	MS-COCO[2]	73.0	**74.9**	74.9	**77.7**	76.7	**80.5**	79.5	**82.2**	81.1	**83.7**
	Pascal-VOC[2]	72.1	**72.4**	72.4	**81.5**	86.9	**87.9**	85.8	**89.8**	88.2	**92.5**
Video Action Recognition	Kinetics 200[3]	72.2	**74.3**	73.2	**78.1**	78.2	**81.3**	80.5	**84.3**	82.1	**84.6**



Ridnik et al, arXiv:2104.10972, NeurIPS Datasets and Benchmarks Track, 2021

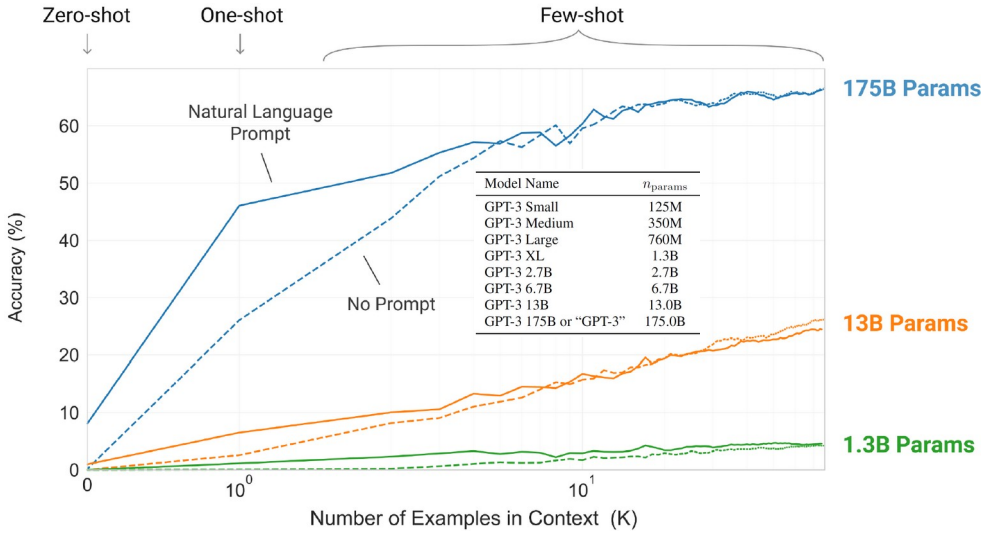
Foundation models: generic transferable learning

- Since ca. 2019: „**foundation models**“ -
- Following generic pre-training, **transfer** to various conditions and tasks possible
- **Scaling laws:** transfer, especially zero and few-shot, **improves when scaling up**
- Natural language as interface for flexible task formulation
- Advanced functionality like in-context learning gets stronger at larger scales

Few-shot

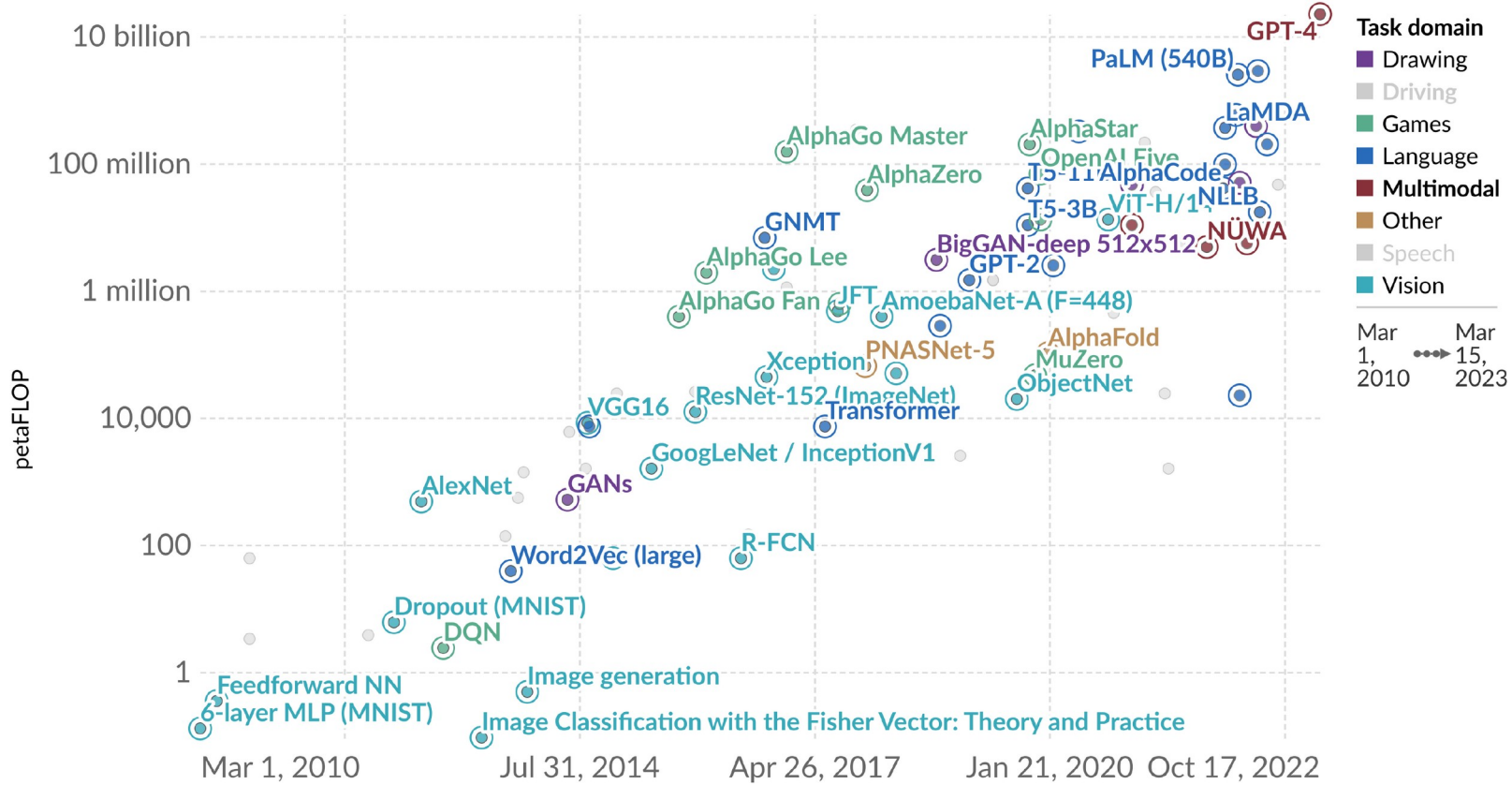
In addition to the task description, the model sees a few examples of the task. No gradient updates are performed.

```
1  Translate English to French:        ← task description
2  sea otter => loutre de mer          ← examples
3  peppermint => menthe poivrée        ← examples
4  plush girafe => girafe peluche      ← examples
5  cheese =>                           ← prompt
```



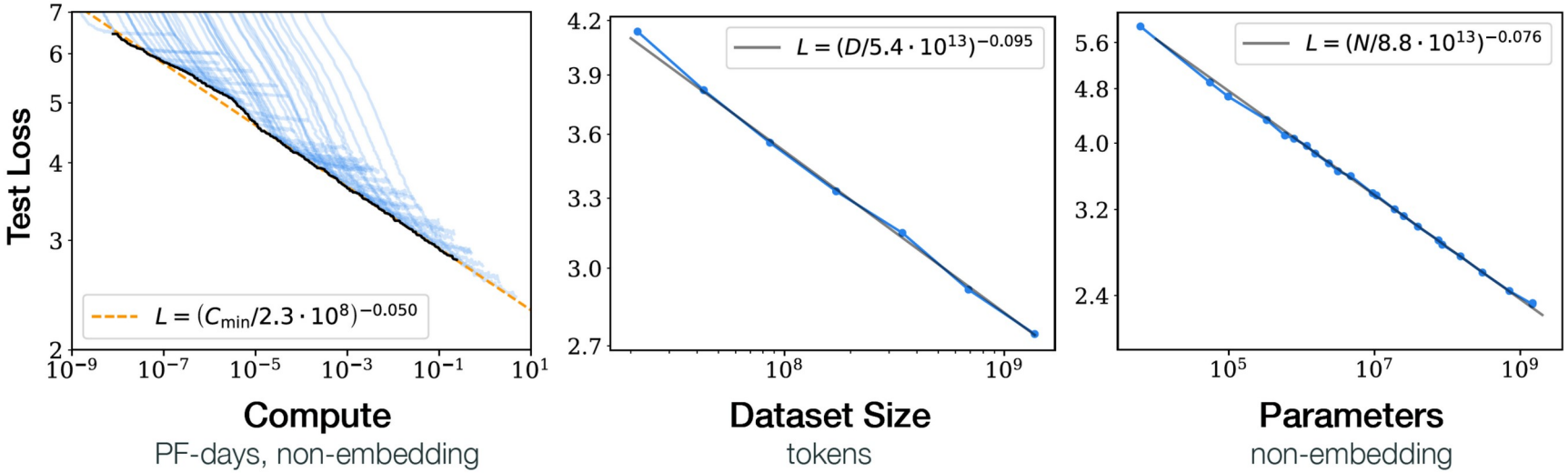
Brown et al, NeurIPS, 2020

Foundation models: larger scale, stronger function



Foundation models: scaling laws

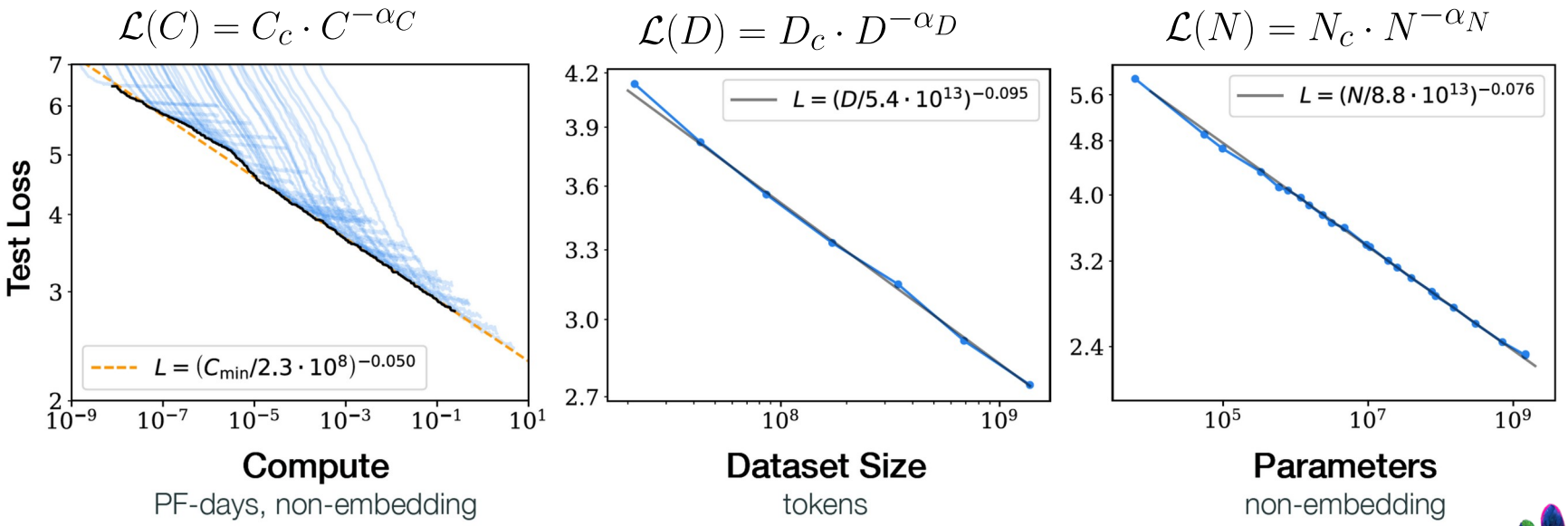
- **Scaling Laws**: larger model, data and compute scale during pre-training – **stronger generalization & transferability**
- **No change** in core algorithmic procedure required! Scaling up alone improves important core functions



Kaplan et al, 2020

Foundation models: scaling laws

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

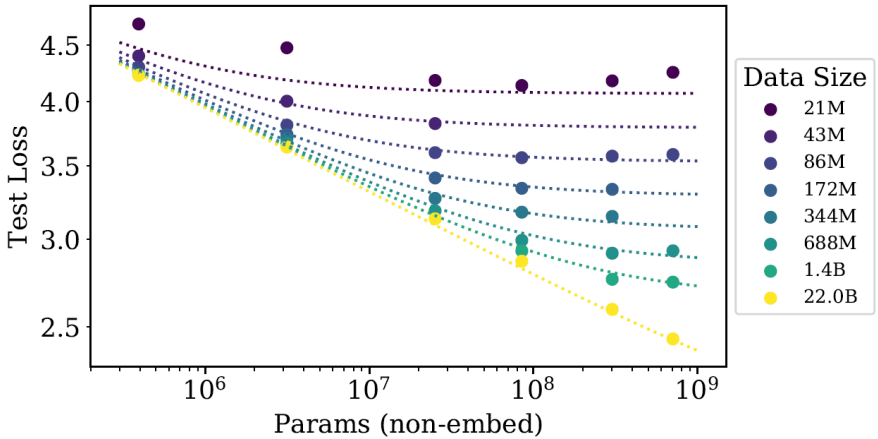


Kaplan et al, 2020

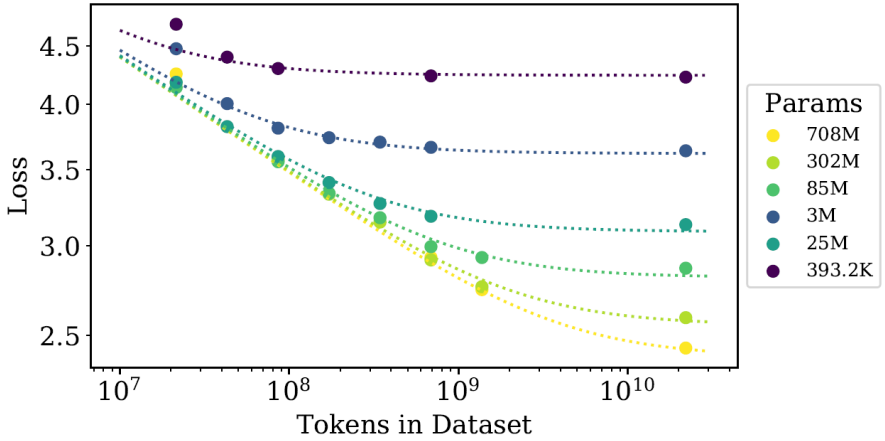
Foundation models: scaling laws

- **Scaling Laws**: predicting model properties and function across scales
- Bottlenecks: imposing limit on a scale prevents improvement when increasing others

$$\mathcal{L}(N) = N_c \cdot N^{-\alpha_N}$$



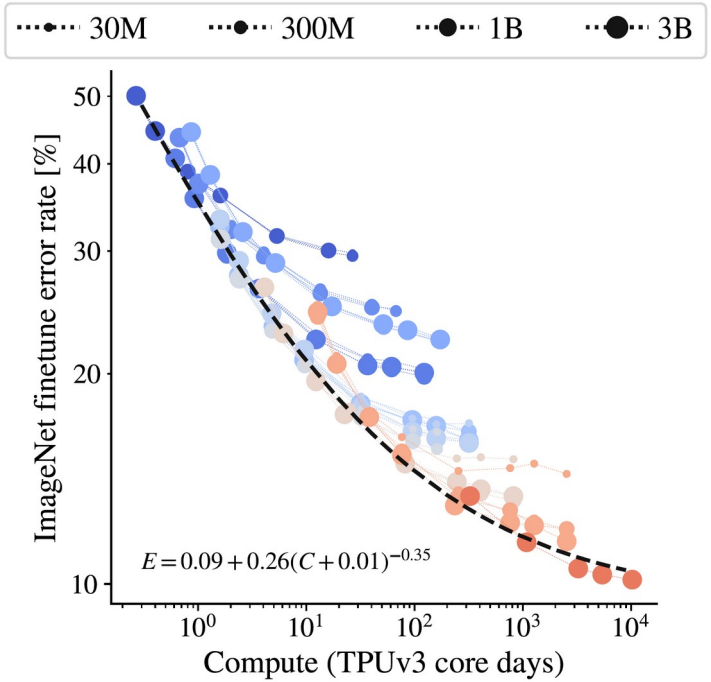
$$\mathcal{L}(D) = D_c \cdot D^{-\alpha_D}$$

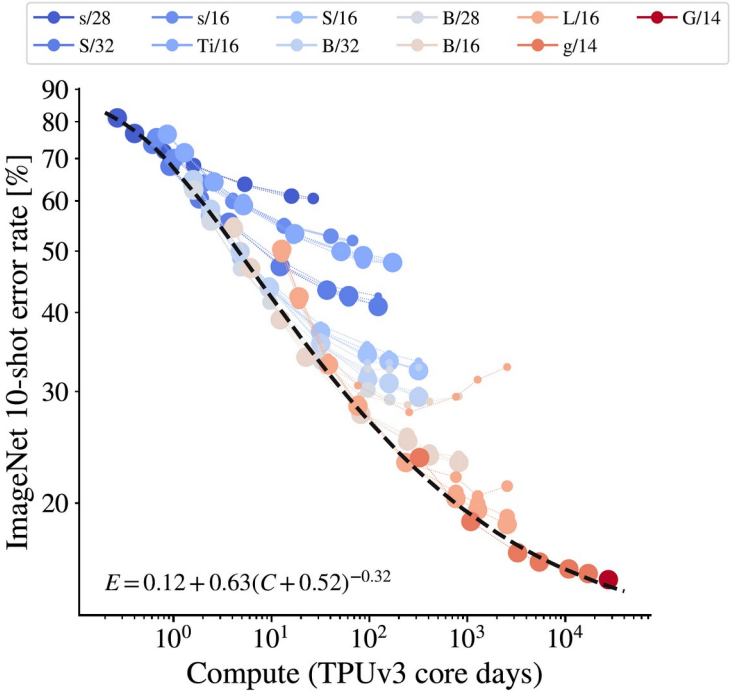


Kaplan et al, 2020

Foundation models: scaling laws

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

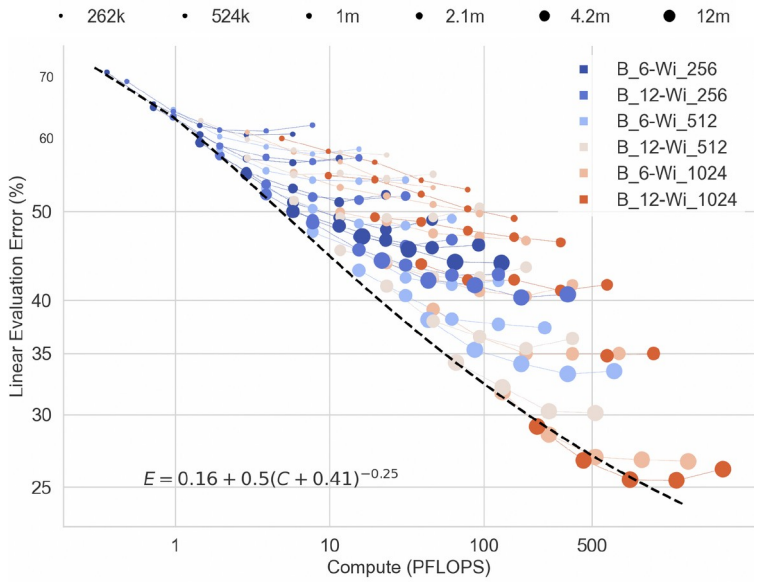


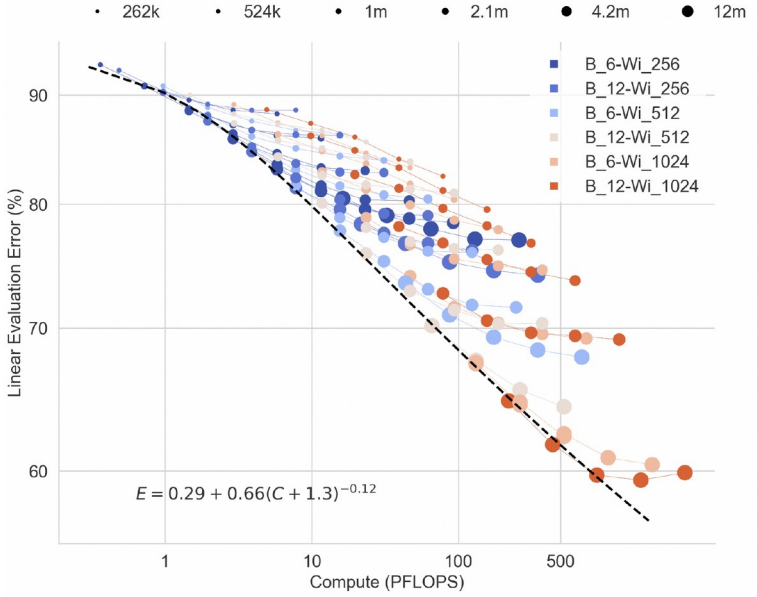


Zhai et al, arXiv:2106.04560, CVPR, 2022

Foundation models: scaling laws

- Scaling Laws: exist for various generalist learning procedures
- Example: Superised classification, plain MLPs (CIFAR100, ImageNet-1k)





Bachmann et al, arXiv:2306.13575, NeurIPS, 2023

Foundation models: scaling laws

- Example: CLIP – language-vision foundation model, contrastive multi-modal loss
- **self-supervised** language-vision learning (scalable data – public web scale)

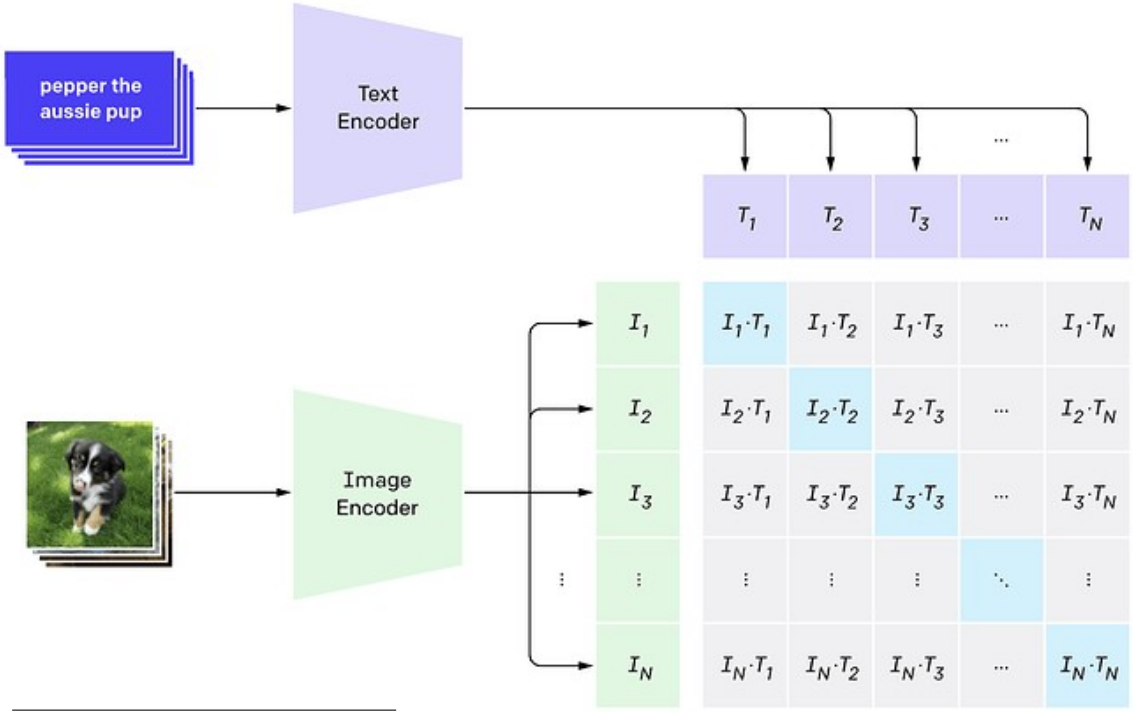
C: Green Apple Chair

C: sun snow dog

C: pink, japan, aesthetic image

C: french cat

Schuhmann et al, NeurIPS, 2022

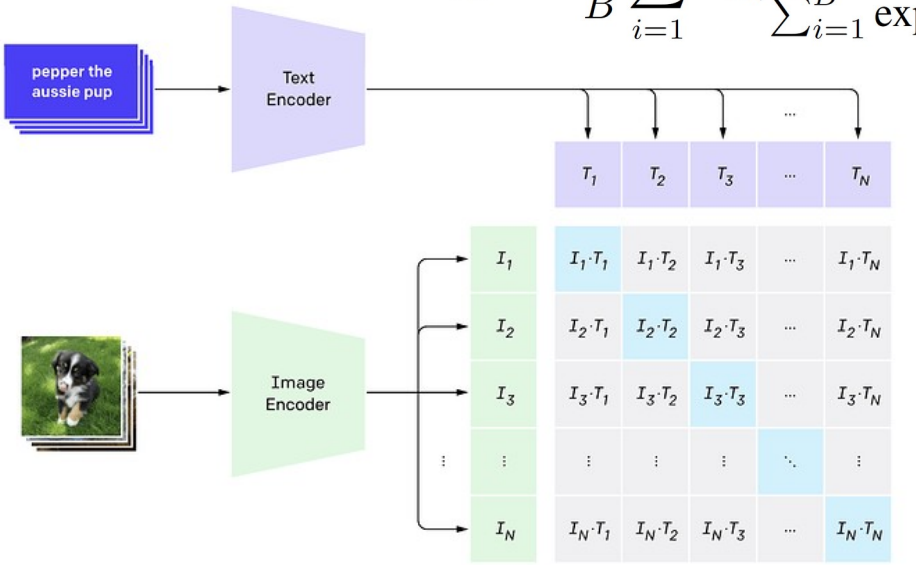


Radford et al, ICML, 2021

Foundation models: scaling laws

- Self-supervised language-vision learning (web-scale data)
- Contrastive image-text pair based loss (positive & negative pairs)

$$L_{\text{I2T}} = -\frac{1}{B}\sum_{i=1}^{B}\log\left(\frac{\exp(v_i l_i/\tau)}{\sum_{i=1}^{B}\exp(v_i l_j/\tau)}\right). \quad L_{\text{T2I}} = -\frac{1}{B}\sum_{i=1}^{B}\log\left(\frac{\exp(v_i l_i/\tau)}{\sum_{j=1}^{B}\exp(v_i l_j/\tau)}\right)$$

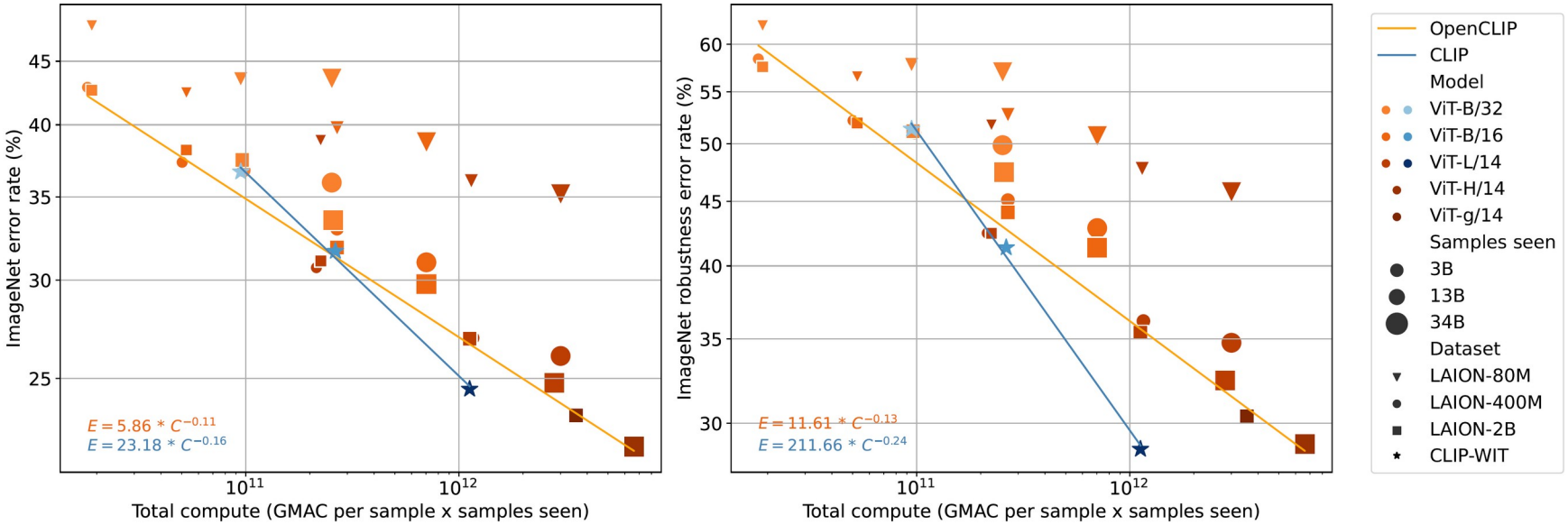


$$L_{contrastive} = L_{\text{I2T}} + L_{\text{T2I}}$$

Radford et al, ICML, 2021

Reproducible scaling laws for foundation models

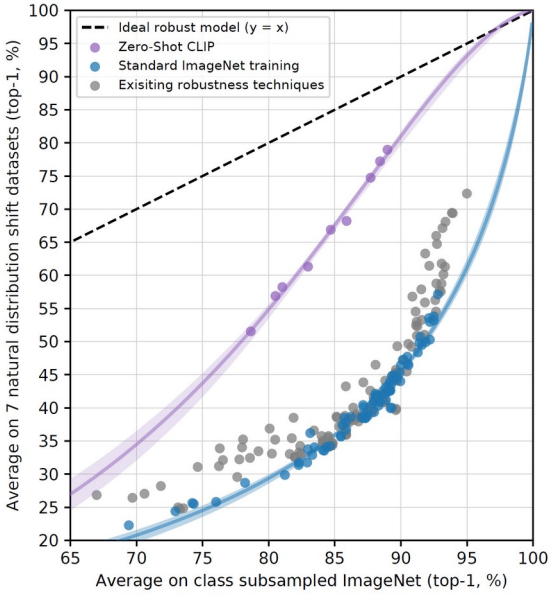
- Scaling laws with LAION and openCLIP: open-source data, models and code - reproducible science of foundation models

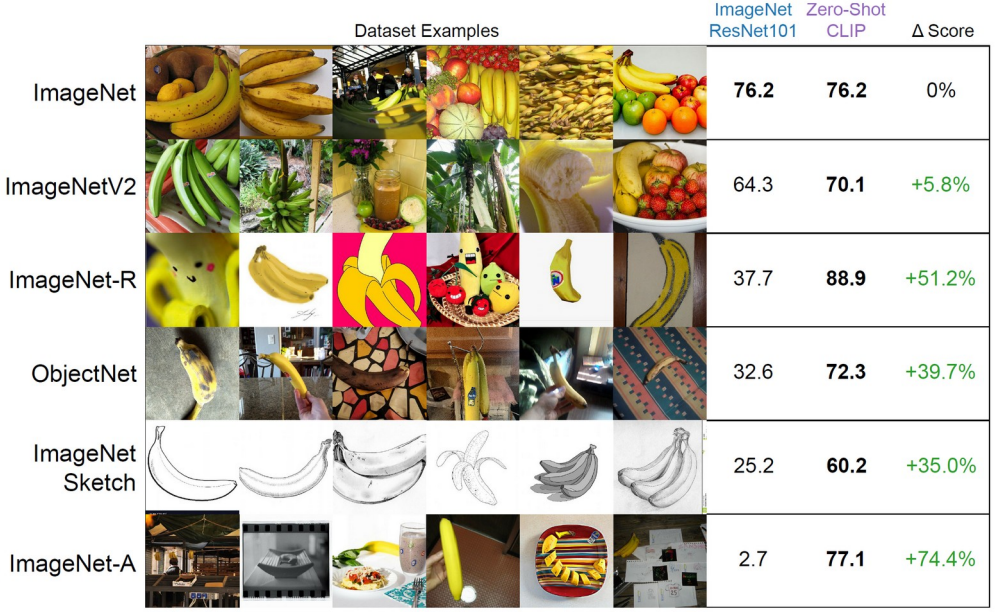


Schuhmann et al, NeurIPS, 2022; Cherti et al, CVPR 2023

Scaling laws: predicting transfer improvement

- CLIP: Out-of-distribution robustness for strong transfer from contrastive language-vision learning

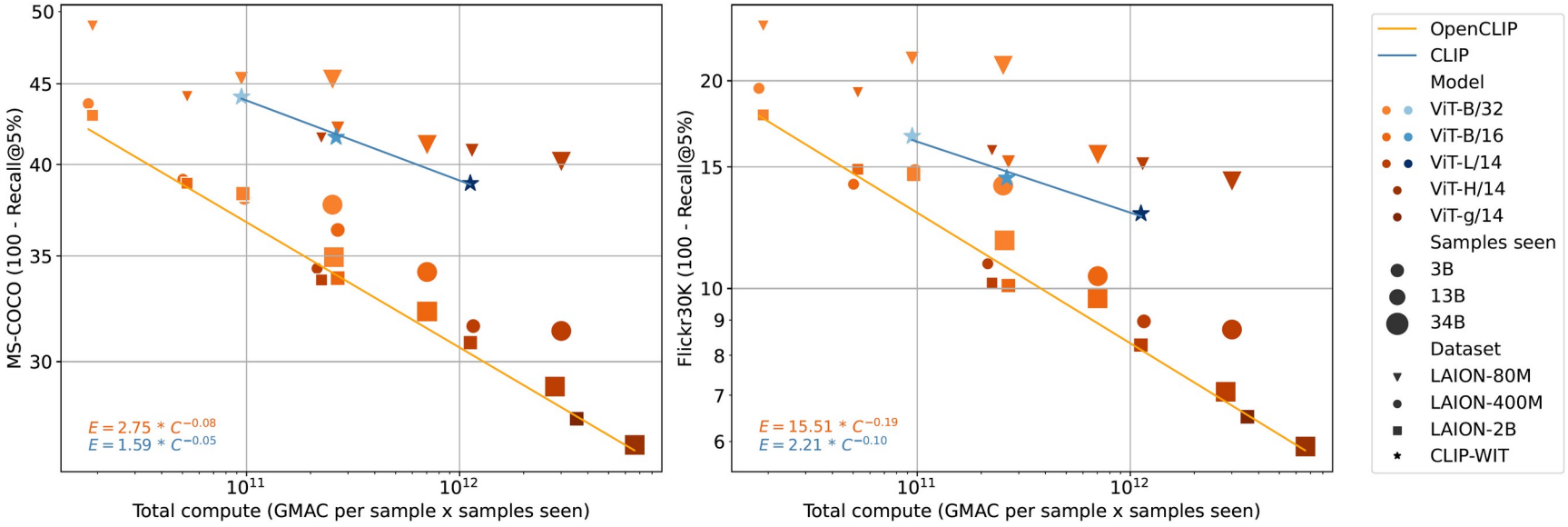




Radford et al, ICML, 2021

Reproducible scaling laws

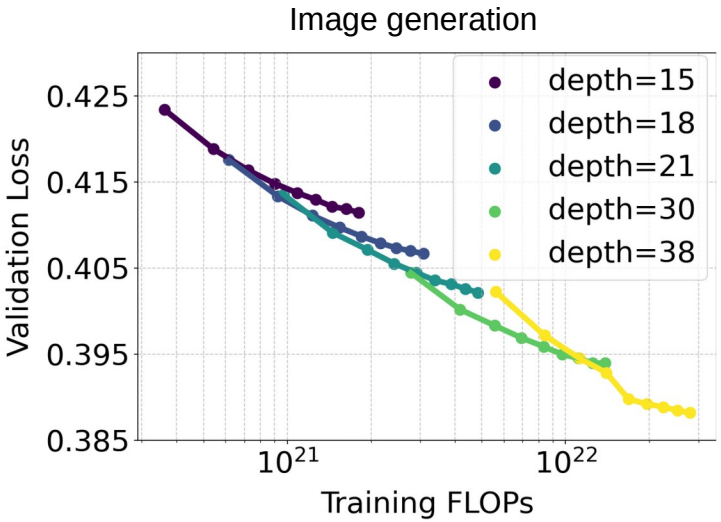
- Scaling laws: zero-shot image retrieval, MS-COCO & Flickr30K

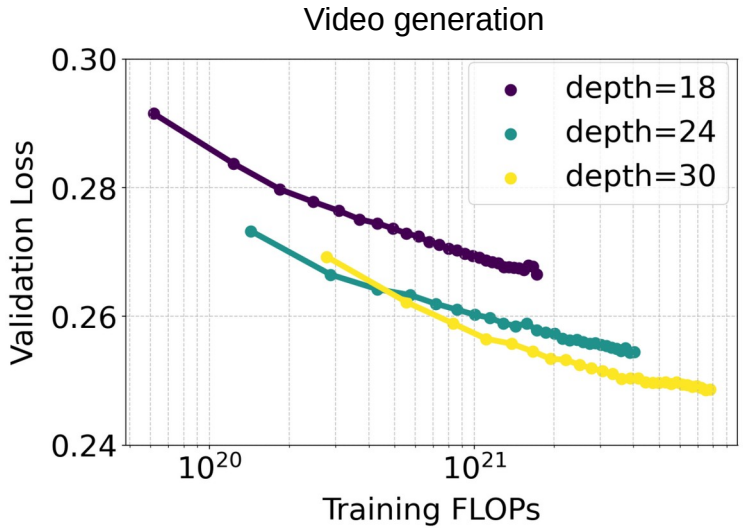


Cherti et al, CVPR 2023

Foundation models: scaling laws

- Example: rectified flow transformers, diffusion models, image and video generation (as used in FLUX models, Black Forest Labs)
- Diffusion/flow based losses, DiT variants architecture

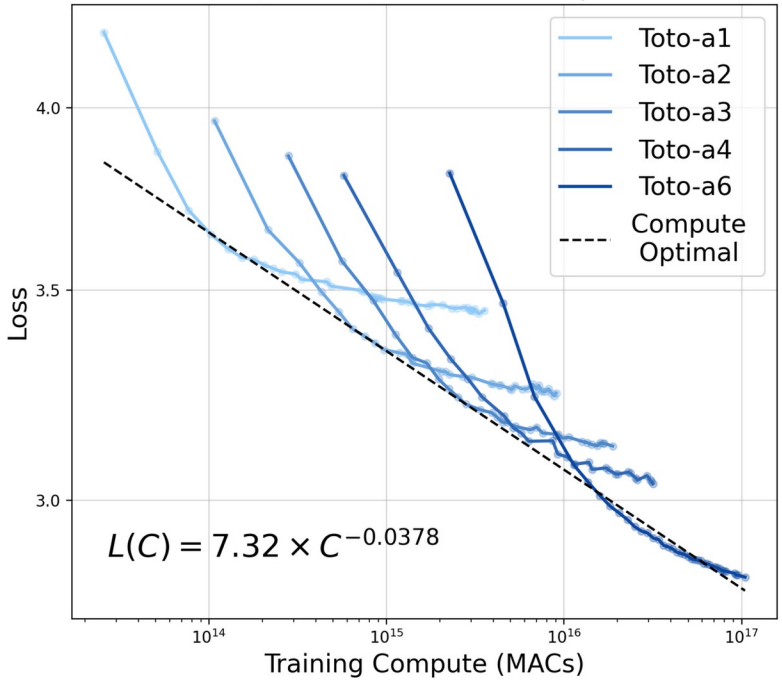




Esser et al, arXiv:2403.03206, ICML, 2024

Foundation models: scaling laws

- Scaling law: derivation requires measurements on **well tuned** (hyperparams) configurations across scales span



Power law: loss-total compute dependency

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

Power law: loss-(model scale, data samples) dependency

$$\mathcal{L}(N, D) = N_c \cdot N^{-\alpha_N} + D_c \cdot D^{-\alpha_D} + L_\epsilon$$

Example: openCLIP ViT, image-text samples

Model N / Data (D)	B/32 (100M)	B/16 (150M)	L/14 (300M)	H/14 (600M)	g/14 (1B)	G/14 (1.8B)
12.8M	L(N,D)	L(N,D)	L(N,D)	L(N,D)	L(N,D)	L(N,D)
128M	L(N,D)	L(N,D)	L(N,D)	L(N,D)	L(N,D)	L(N,D)
1.28B	L(N,D)	L(N,D)	L(N,D)	L(N,D)	L(N,D)	L(N,D)
12.8B	L(N,D)	L(N,D)	L(N,D)	L(N,D)	L(N,D)	L(N,D)

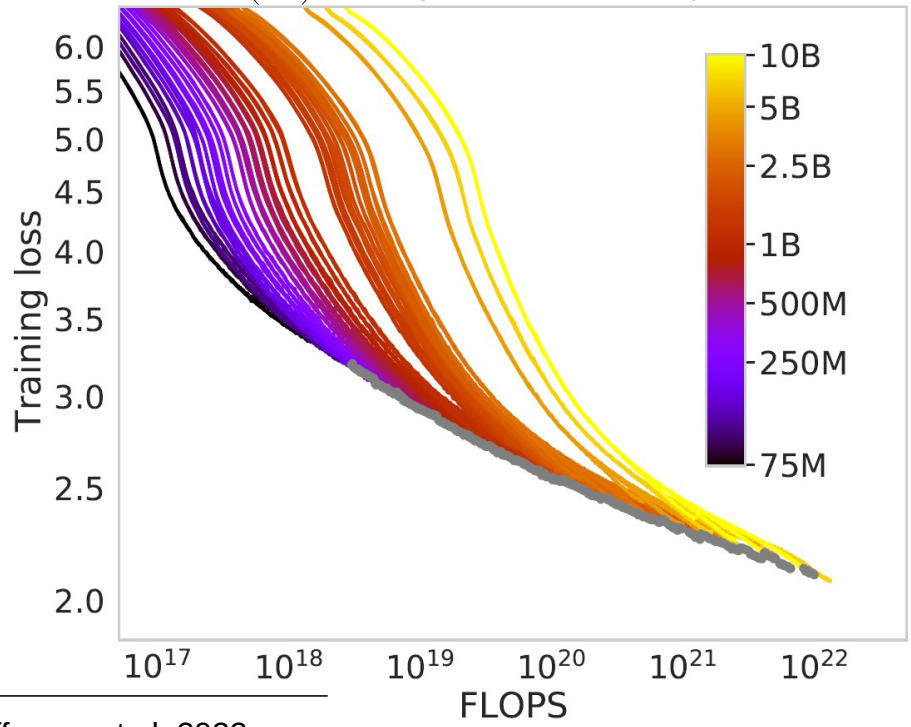
Rajasegaran et al, arXiv:2501.05453, 2025

Foundation models: scaling laws

- Scaling law: prediction

Min Loss-total compute scaling law

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



Hoffmann et al, 2022

Approx. for compute, for LM C=6ND (Kaplan et al, 2020)

$$C = \xi N D$$

- Measure loss for various N, D combinations
  - Eg, fix N, go through increasing D
- For each C(N,D), get min L(C(N,D))
- Fit L(C) through those points

LM, text data tokens

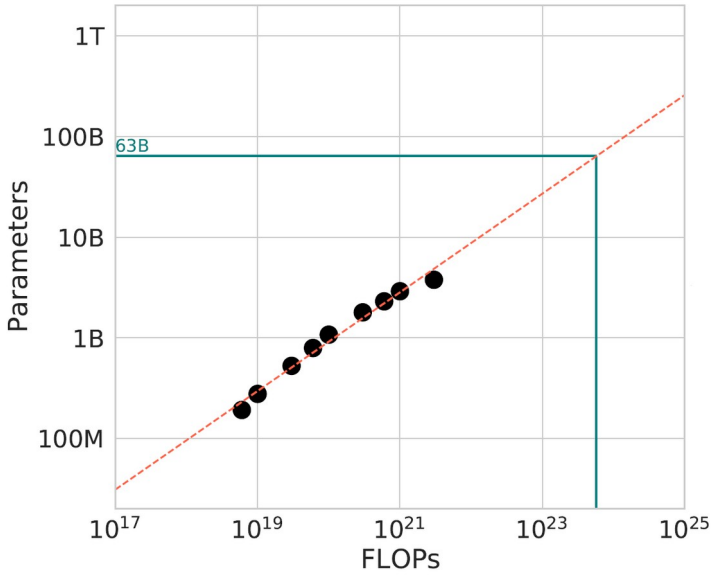
Model N / Data (D)	75M	250M	500M	1B	2.5B	5B	10B
10B	L(N,D)	L(N,D)	L(N,D)	L(N,D)	L(N,D)	L(N,D)	L(N,D)
20B	L(N,D)	L(N,D)	L(N,D)	L(N,D)	L(N,D)	L(N,D)	L(N,D)
50B	L(N,D)	L(N,D)	L(N,D)	L(N,D)	L(N,D)	L(N,D)	L(N,D)
100B	L(N,D)	L(N,D)	L(N,D)	L(N,D)	L(N,D)	L(N,D)	L(N,D)
300B	L(N,D)	L(N,D)	L(N,D)	L(N,D)	L(N,D)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

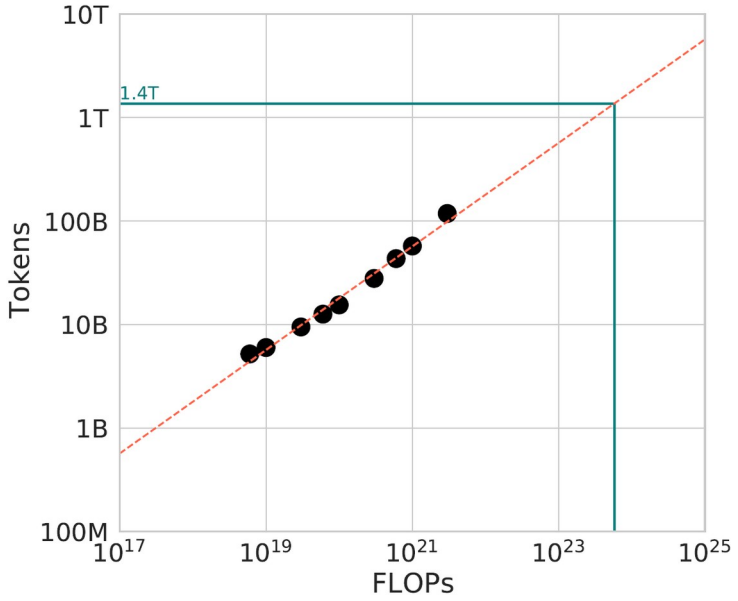
Foundation models: scaling laws

- Power law predicting N*, D* for a given compute C to gain min loss L
- Scaling coeff a, b : how much of total compute to put into params or samples seen

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

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





Foundation models: scaling laws

- Power laws predicting N*, D* for a given compute C to gain min loss L
- Scaling coeff a, b : tell how much of compute to put into params or samples seen
- a=b : params and data should grow with same rate (Hoffmann et al)
- a > b : model params should grow faster; more compute into model scale (Kaplan)
- Discrepancy Kaplan vs Hoffmann: same procedure, why different estimates?

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

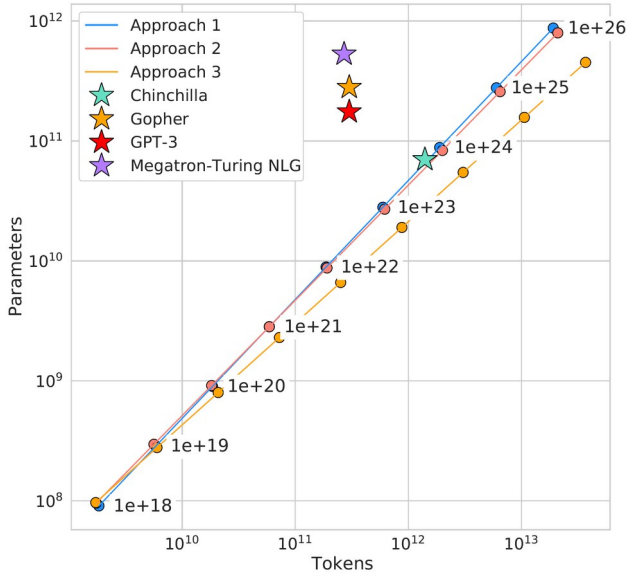
Approach	Coeff. a where $N_{opt} \propto C^a$	Coeff. b where $D_{opt} \propto C^b$
1. Minimum over training curves	0.50 (0.488, 0.502)	0.50 (0.501, 0.512)
2. IsoFLOP profiles	0.49 (0.462, 0.534)	0.51 (0.483, 0.529)
3. Parametric modelling of the loss	0.46 (0.454, 0.455)	0.54 (0.542, 0.543)
Kaplan et al. (2020)	0.73	0.27

Kaplan et al, 2020; Hoffmann et al, 2022

Foundation models: scaling laws

- Discrepancy Kaplan vs Hoffmann: different estimates give different predictions - has to be resolved

Prediction Kaplan et al : 3T! ...



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

Kaplan et al : 300B! …
BLOOM followed that ...

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

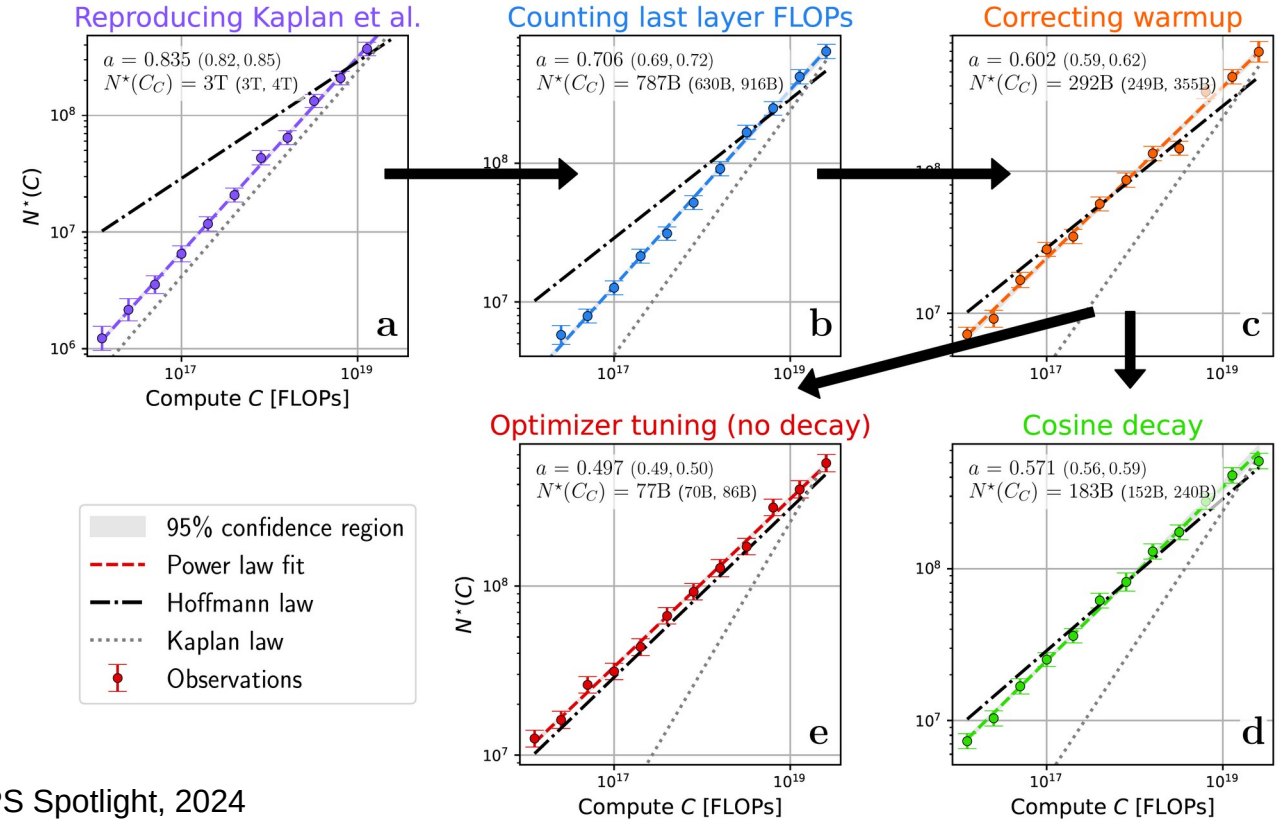
Approach	Coeff. a where $N_{opt} \propto C^a$	Coeff. b where $D_{opt} \propto C^b$
1. Minimum over training curves	0.50 (0.488, 0.502)	0.50 (0.501, 0.512)
2. IsoFLOP profiles	0.49 (0.462, 0.534)	0.51 (0.483, 0.529)
3. Parametric modelling of the loss	0.46 (0.454, 0.455)	0.54 (0.542, 0.543)
Kaplan et al. (2020)	0.73	0.27

Kaplan et al, 2020; Hoffmann et al, 2022

Foundation models: scaling laws

- Kaplan vs Chinchilla: measured same phenomena, with different measurement procedure
- EXTREMELY IMPORTANT:

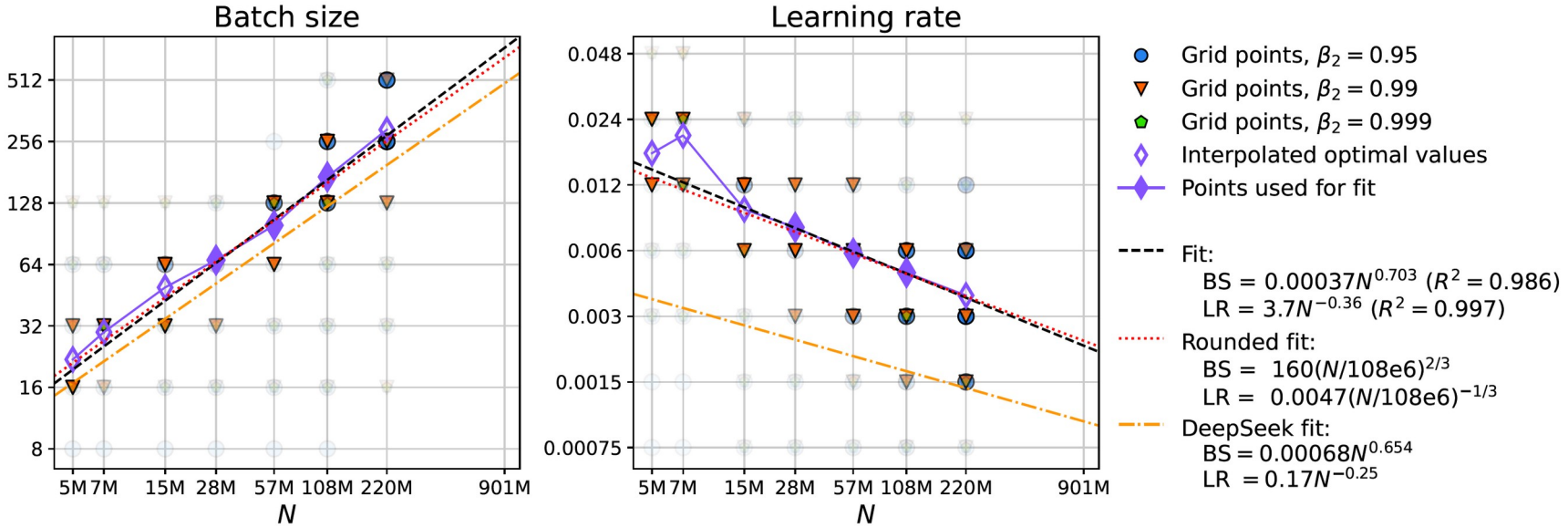
  TUNE hyperparams for each measurement !



Porian et al, arXiv:2406.19146, NeurIPS Spotlight, 2024

Foundation models: scaling laws

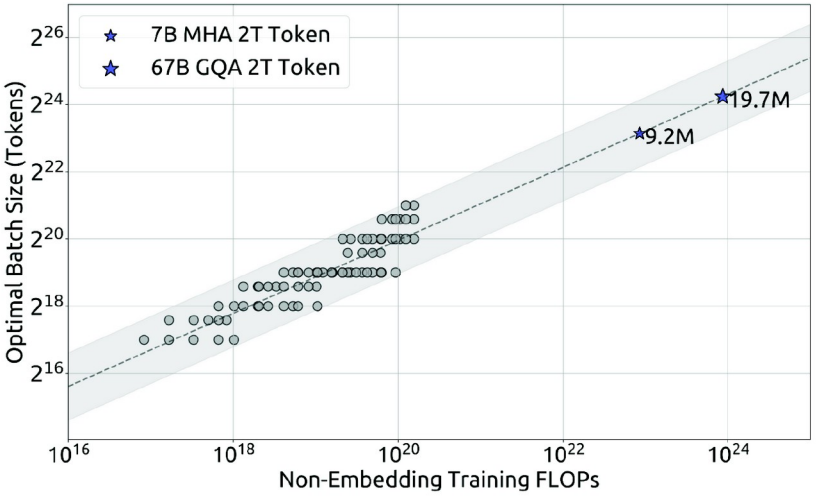
- EXTREMELY IMPORTANT: TUNE hyperparams for each measurement
- Scaling laws for hyperparameter prediction
- Other approaches for prediction, eg muP



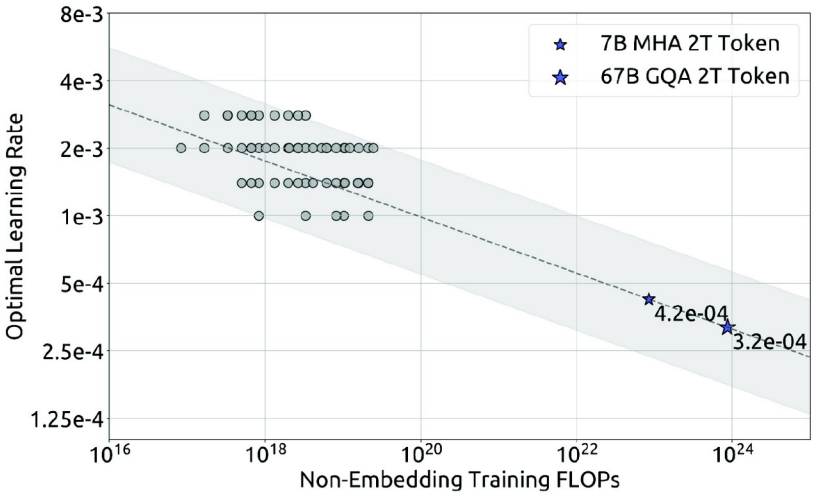
Porian et al, arXiv:2406.19146, NeurIPS Spotlight, 2024

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

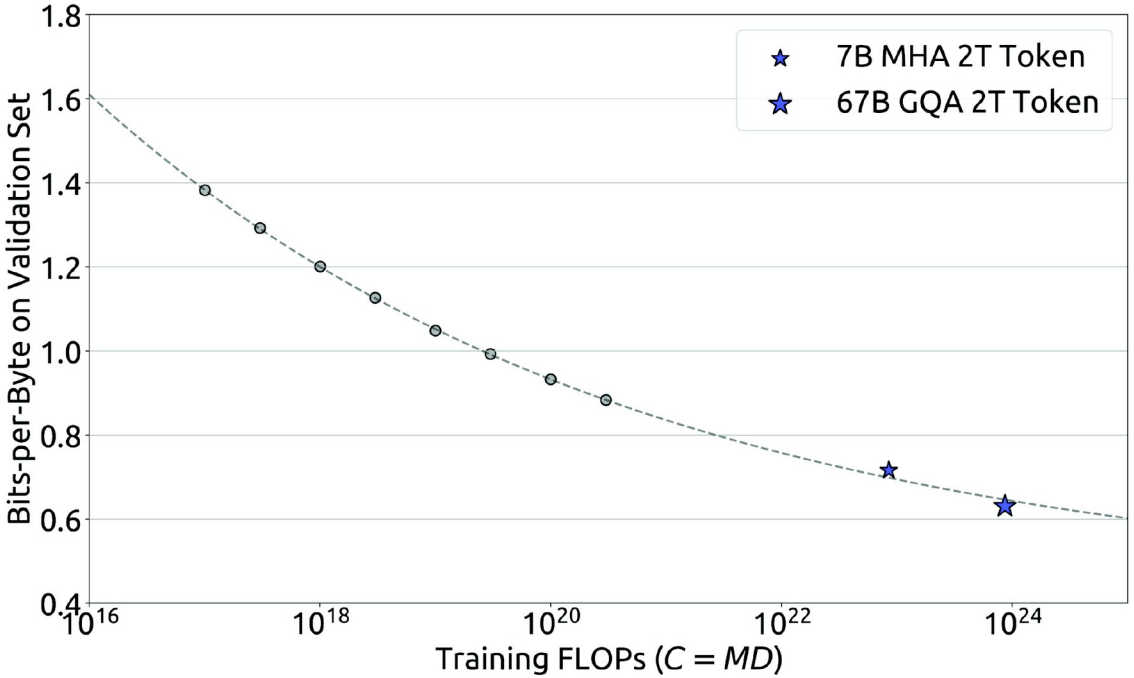


(b) Learning rate scaling curve

Bi et al, ,2024

Foundation models: scaling laws

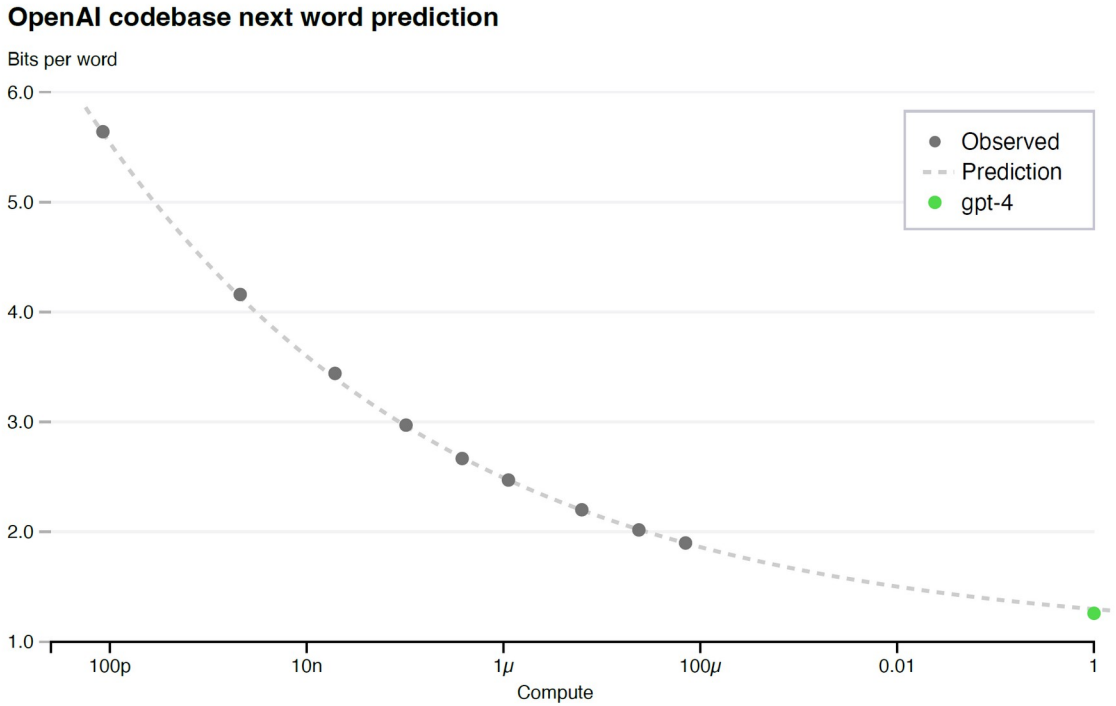
- Scaling law: predicting training/model properties and function
- Predictions are accurate if scaling law derivation is done properly



Bi et al, ,2024

Foundation models: scaling laws

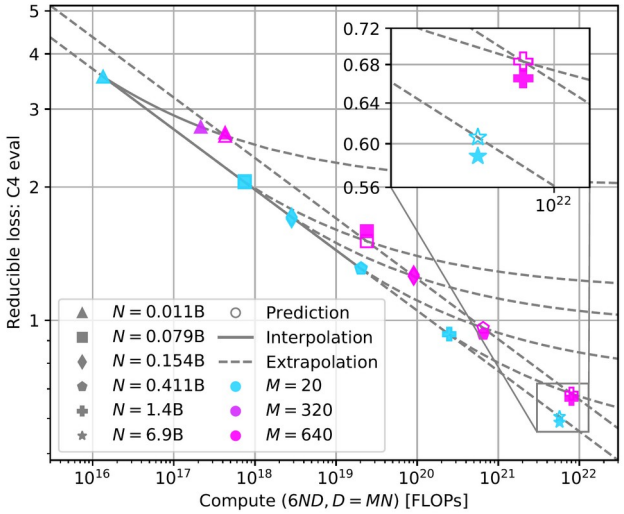
- Scaling law: predicting training/model properties and function
- Predictions are accurate if scaling law derivation is done properly

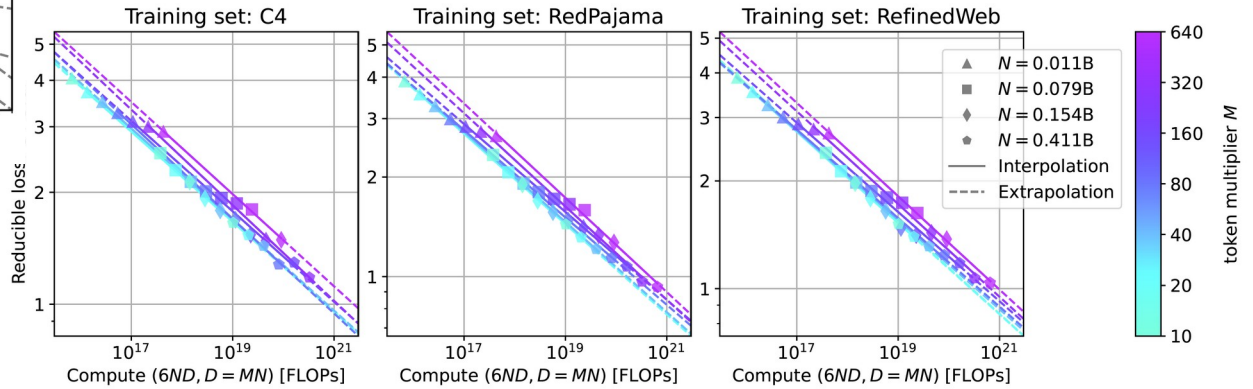


et al, openAI, GPT-4 Technical Report ,2024

Foundation models: scaling laws

- Scaling law: predicting training/model properties and function
- „Overtraining“ : scaling laws for suboptimal compute budgets

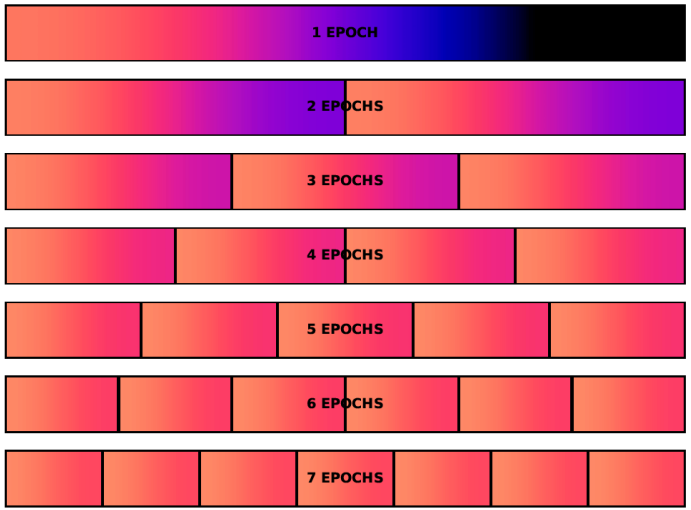


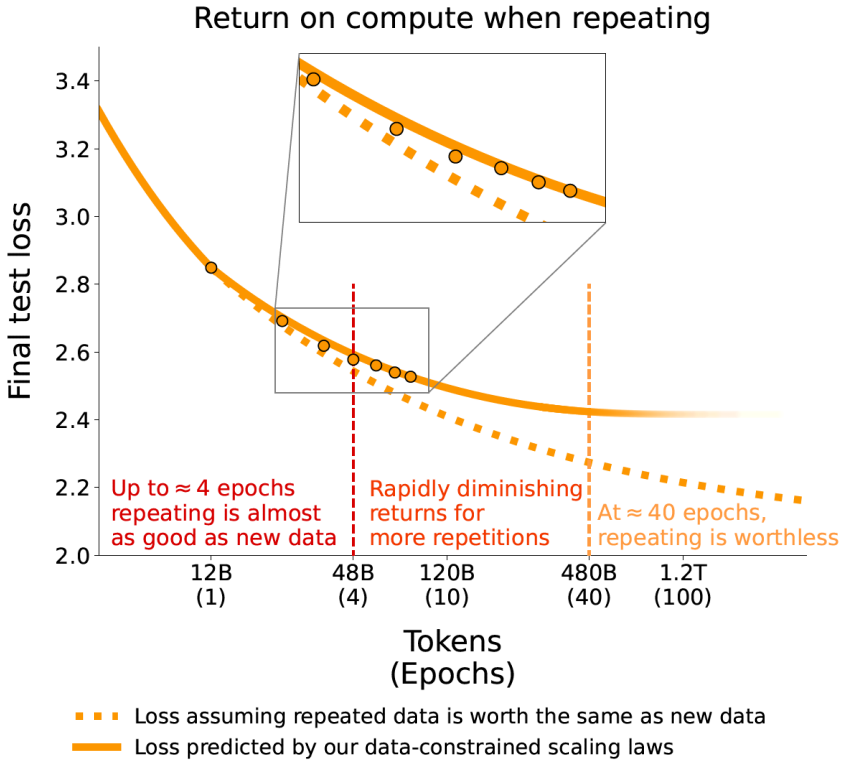


Gadre et al, arXiv:2403.08540, 2024; ICLR, 2025

Foundation models: scaling laws

- Scaling law: predicting training/model properties and function
- Repeating data is an important factor

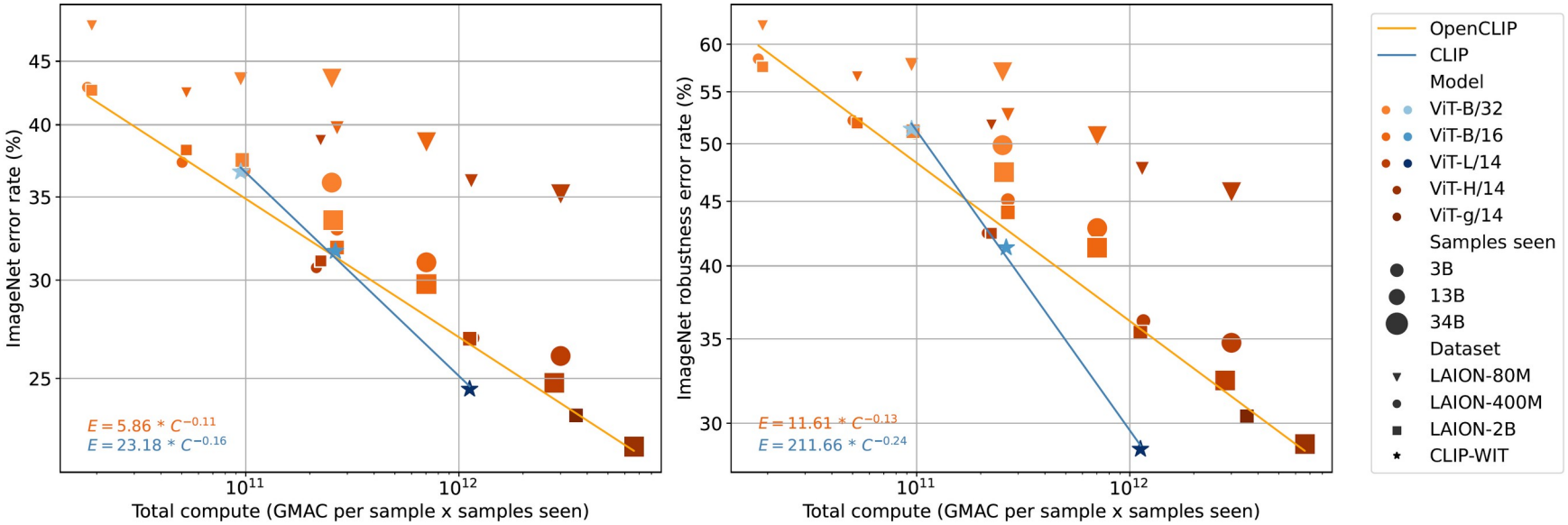




Muennighof et al, arXiv:2305.16264, NeurIPS, 2023

Scaling laws: pre-training procedure comparison

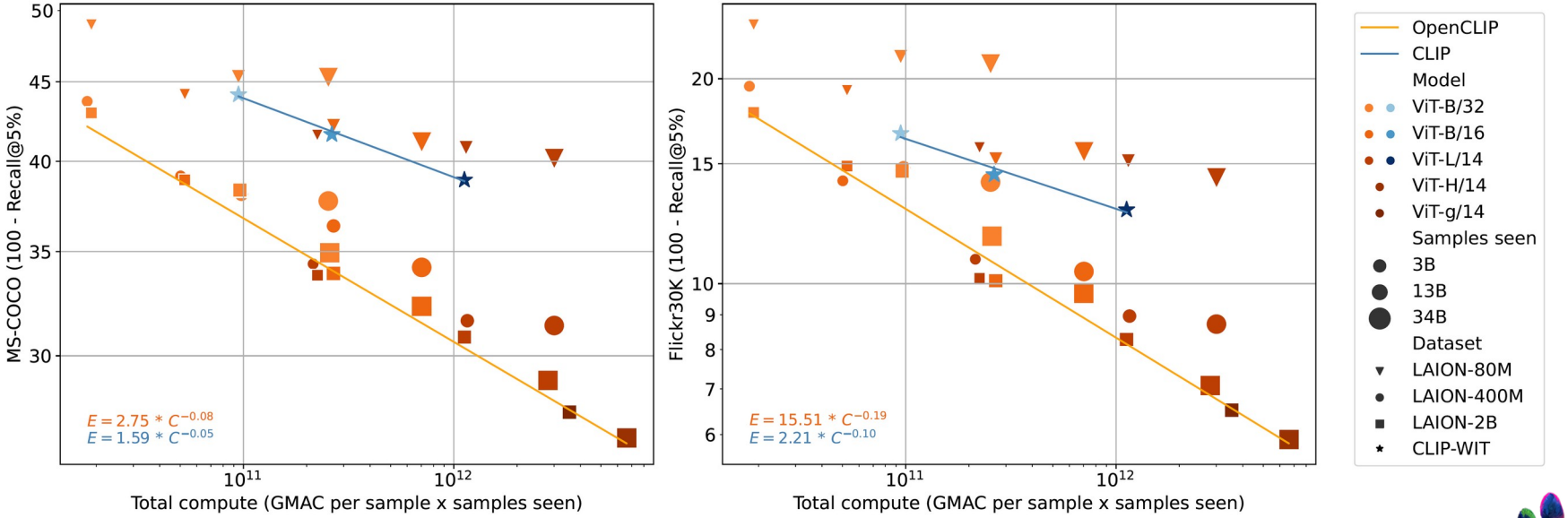
- Example: comparing CLIP pre-training with two different datasets, LAION-400M/2B (LAION) and WIT (openAI)
- Zero-shot ImageNet-1k classification; ImageNet robustness



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

Scaling laws: pre-training procedure comparison

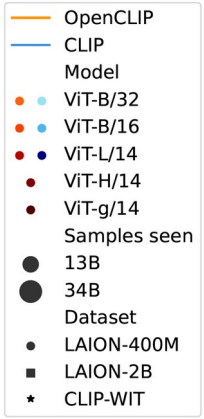
- Example: comparing CLIP pre-training with two different datasets, LAION-400M/2B (LAION) and WIT (openAI)
- Zero-shot image retrieval, MS-COCO & Flickr30K
- Task dependency (classification vs retrieval) revealed

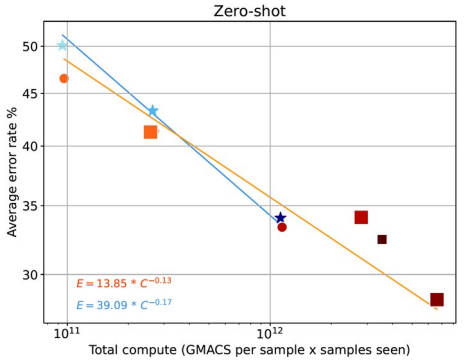


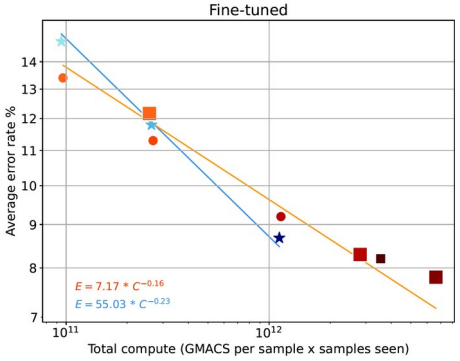
Cherti et al, CVPR 2023

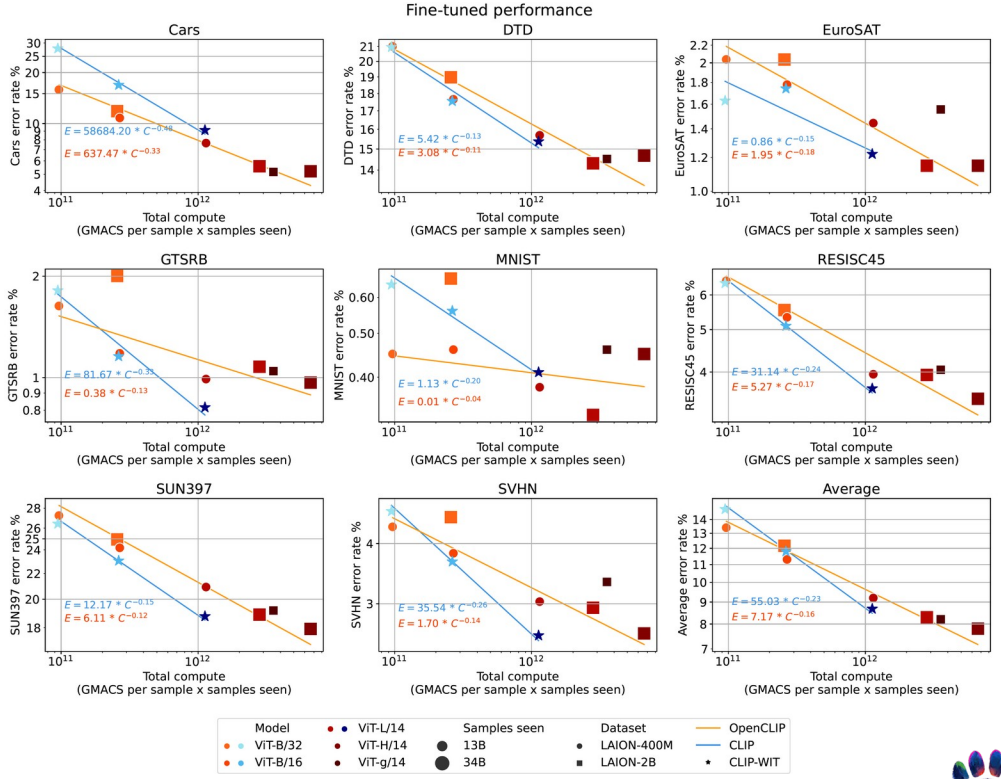
Scaling laws: pre-training procedure comparison

- Scaling laws for various downstream datasets, tasks & transfer procedures





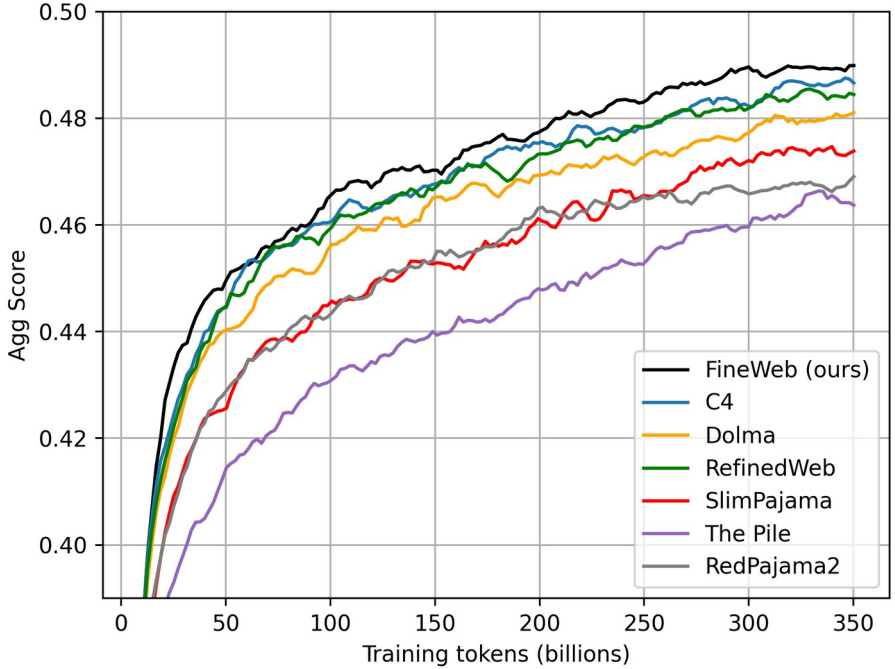




Cherti et al, CVPR 2023

Scaling laws: pre-training procedure comparison

- Principled dataset comparison: what is a better dataset for training?
- Fixed reference scale alone is not enough (here: 1.7B model scale)



Penedo et al, arXiv:2406.17557, NeurIPS Datasets and Benchmarks Spotlight, 2024

Scaling laws: pre-training procedure comparison

- Comparison by deriving full scaling law per dataset. Example : LLMs pre-training with different datasets, DCLM-baselines (DataComp-LM), FineWeb-Edu, OLMO, Llama, RPJ, C4, etc

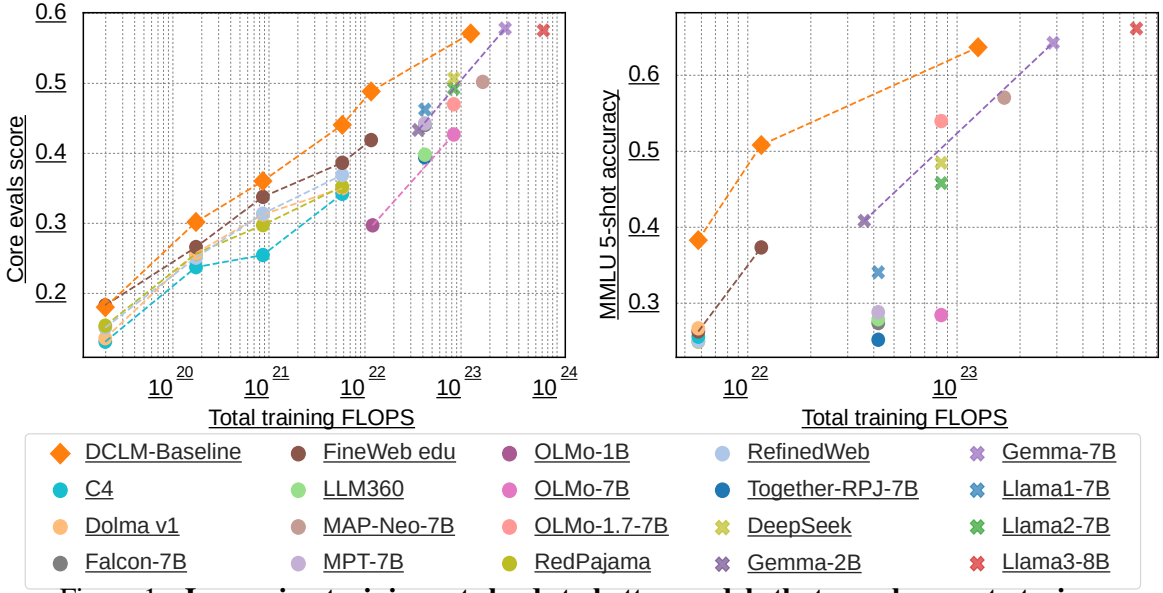


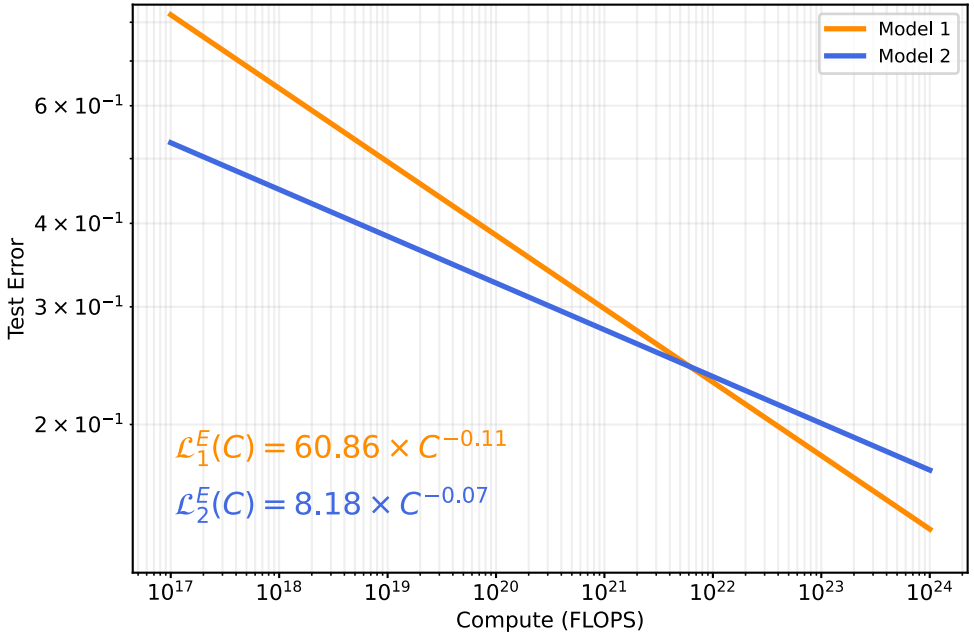
Figure 1: **Improving training sets leads to better models that are cheaper to train.**

Li et al, arXiv:2406.11794, NeurIPS Datasets and Benchmarks, 2024

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)

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

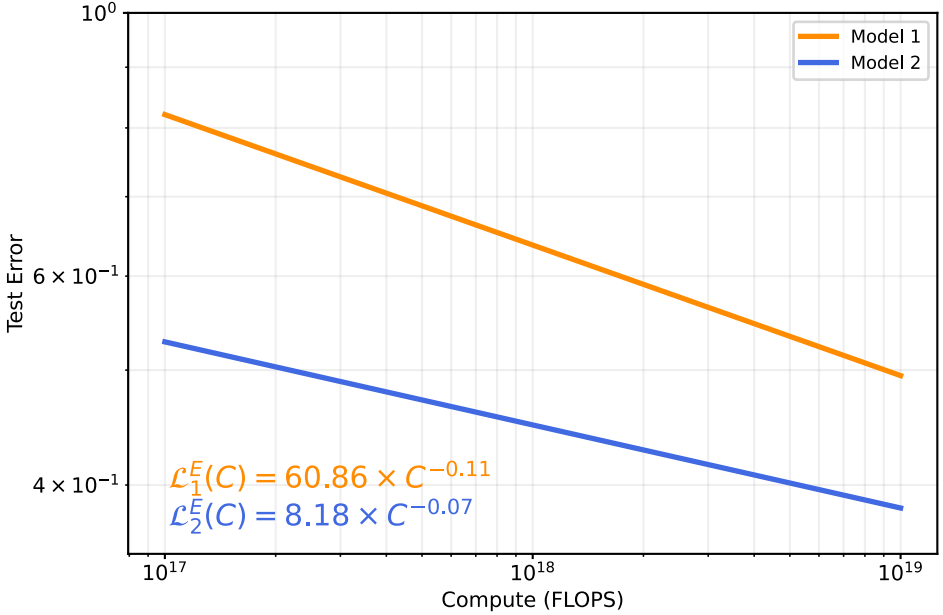


- Model 1 takes over Model 2 at larger compute scales – in higher performance region
- Model 2 outperforms Model 1 at smaller compute scales – in low performance region ...

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)

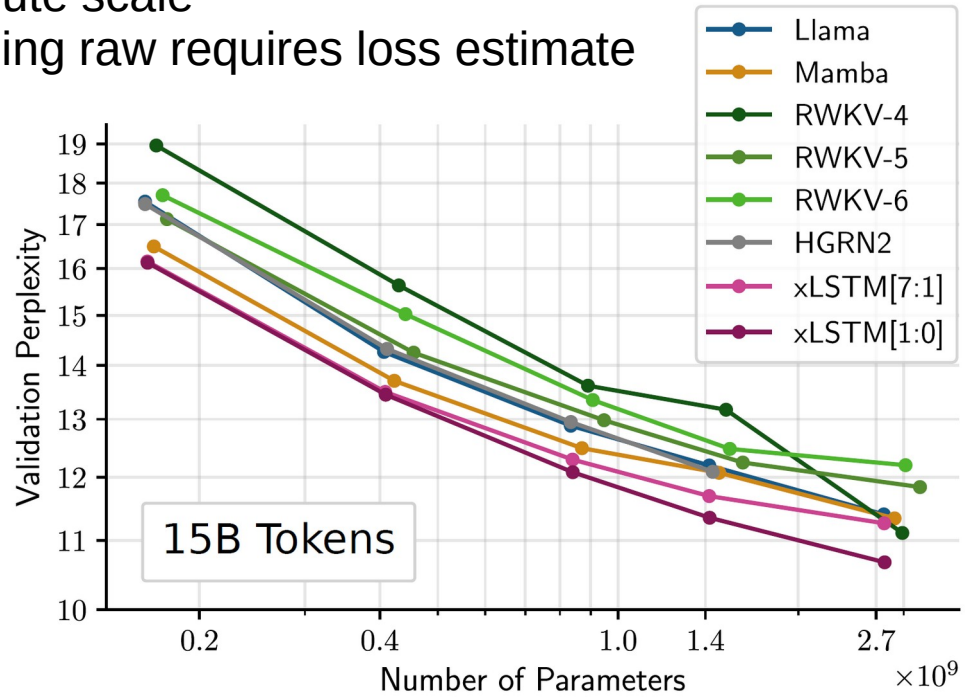
$$\mathcal{L}(C) = C_c \cdot C^{-\alpha_C} + L_\epsilon$$



- 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

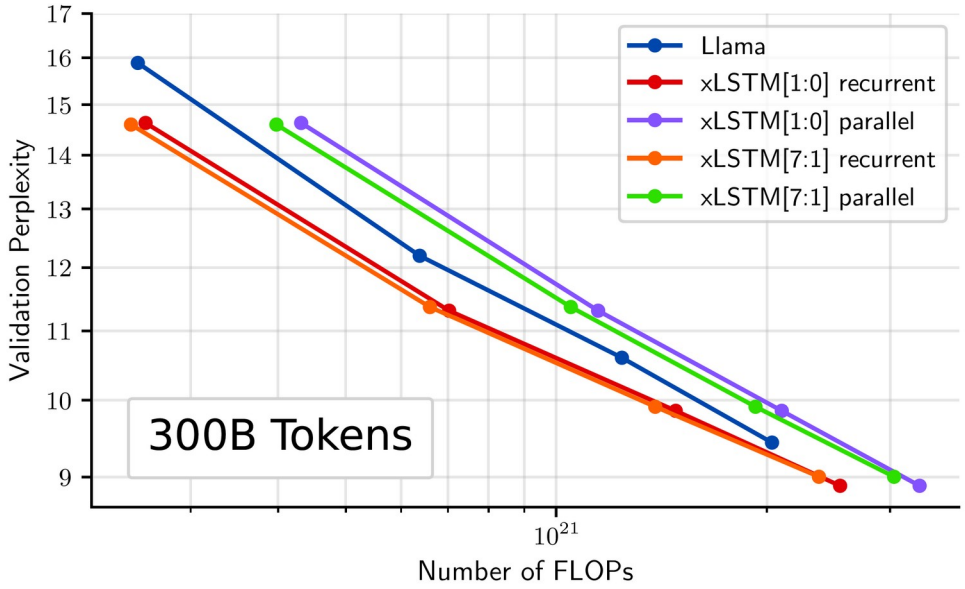
- Improper comparison: example xLSTM
- Comparison in poor performance region: inconclusive
- Small token budget: token/compute scale too small, bottlenecks expected
- No alignment on common compute scale
- Fixed token budget: Proper scaling raw requires loss estimate across combinations of N,D



Beck et al, NeurIPS, 2024

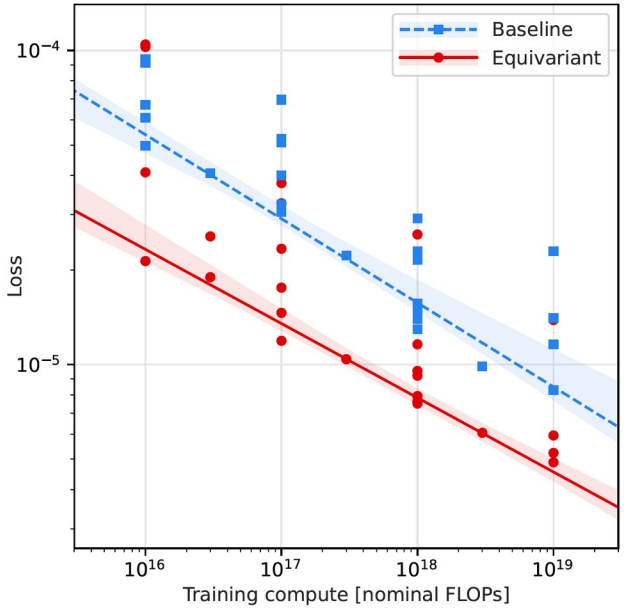
Scaling laws: pre-training procedure comparison

- Improper comparison: example xLSTM
- Fixed token budget 300B: proper scaling raw requires loss estimate across combinations of N,D
- Scale span too narrow (less than 2 orders of magnitude)
- Line crossing probable (Llama taking over xLSTM)



Scaling laws: pre-training procedure comparison

- Insufficient comparison: example strong (equivariancy) vs weak inductive biases
- Upper scale compute insufficient for extrapolation
- Line crossing probably (baseline taking over equivariant model)



DOES EQUIVARIANCE MATTER AT SCALE?

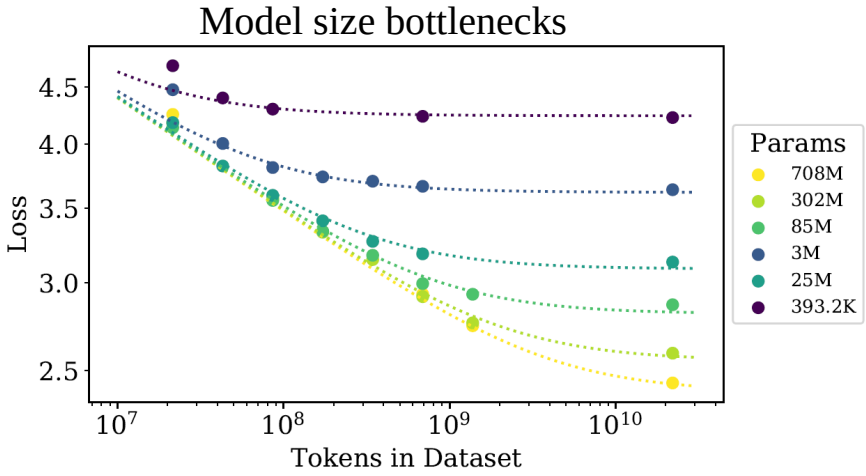
Johann Brehmer* **Sönke Behrends** **Pim de Haan*** **Taco Cohen***

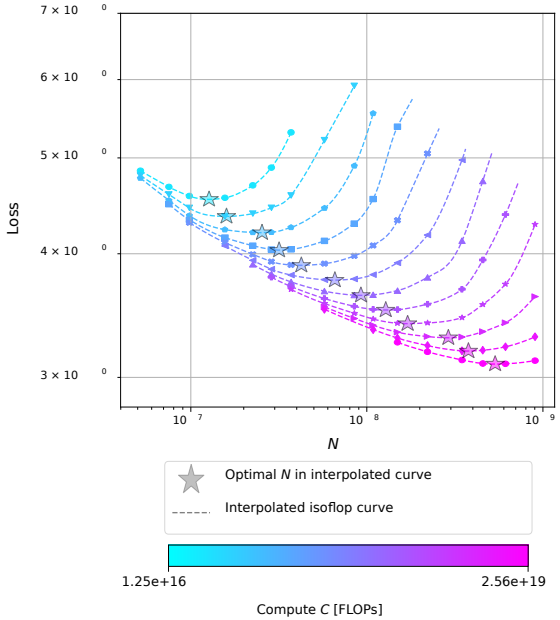
Scaling law	Param.	Baseline			Equivariant		
		Central	Lower	Upper	Central	Lower	Upper
Eq. (2): $\hat{L}(N,D) = A/N^{\alpha} + B/D^{\beta}$	A	**1.27**	0.484	5.07	**0.000282**	0.000162	0.000607
	B	**0.202**	0.0108	0.361	**469**	159	592
	α	**0.909**	0.832	1.03	**0.348**	0.293	0.417
	β	**0.379**	0.256	0.404	**0.734**	0.689	0.747
Eq. (4): $N^*(C) \propto C^a$	a	**0.294**	0.215	0.307	**0.678**	0.619	0.711
	b	**0.706**	0.693	0.785	**0.322**	0.289	0.381
Eq. (5): $L^*(C) = F/C^{\gamma}$	F	**1.03**	0.124	1.89	**0.14**	0.0524	0.517
	γ	**0.268**	0.213	0.284	**0.236**	0.212	0.267

Brehmer et al, arXiv:2410.23179, 2024

Foundation models: scaling laws

- Studying scaling laws: **imperative** for foundation models research
- Without scaling laws: no way to **properly compare models** across scales, no way to predict what happens on various scale combinations

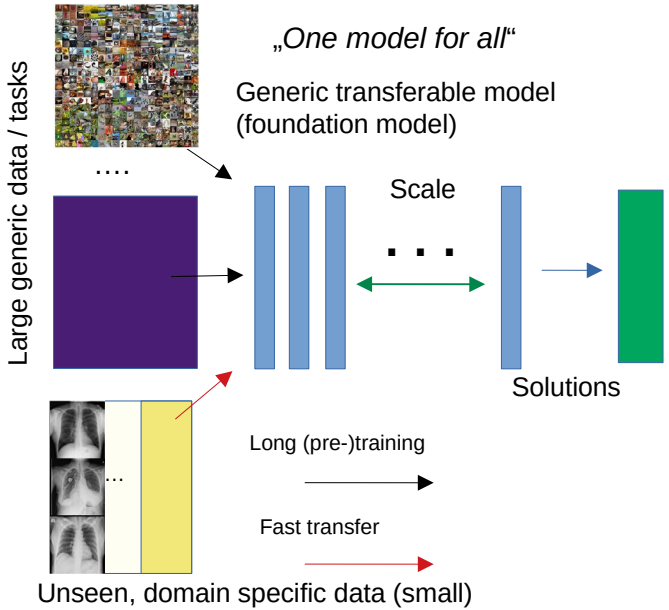




Kaplan et al, 2020; Porian et al, NeurIPS, 2024

Foundation models: reproducibility & progress

- **Problem**: research on foundation models & scaling laws executable and reproducible only by few large industry labs (Google; openAI; Microsoft; Facebook; 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 as well**
- **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 (mostly driven by HuggingFace)
- **LAION** (EU, Germany, 2021; **important hub** at **JSC**): multi-modal language-vision, language-audio – LAION-400M/5B, openCLIP, CLAP, openFlamingo, Open Assistant, open-LM, DataComp, Leo-LM
- **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

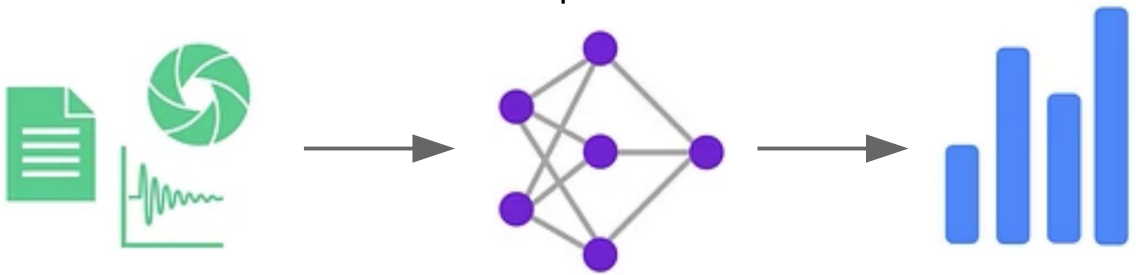
Open-source foundation models & datasets

- 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/

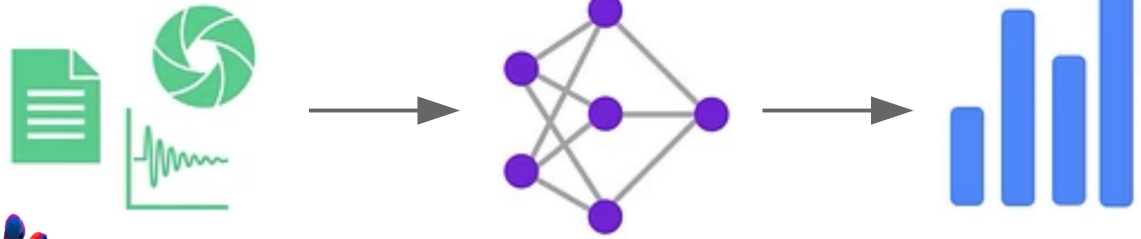
Open-source foundation models & datasets

- 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

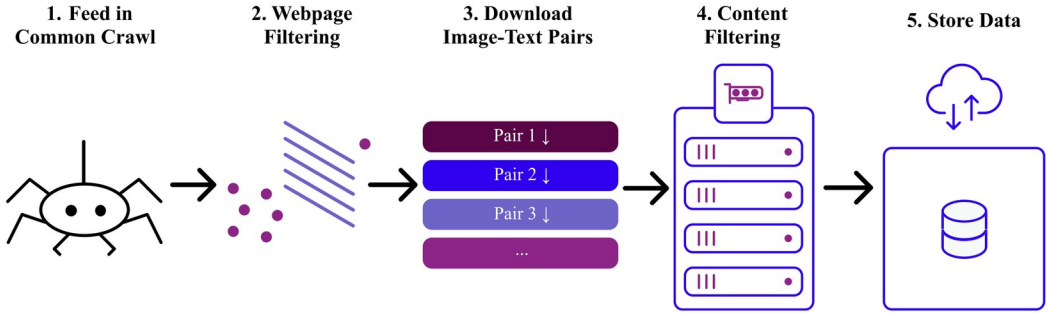
Lm-eval-harness, bigcode-evaluation-harness, LAION-AIW

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

EleutherAI
AI2
BigScience

Open large-scale 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 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

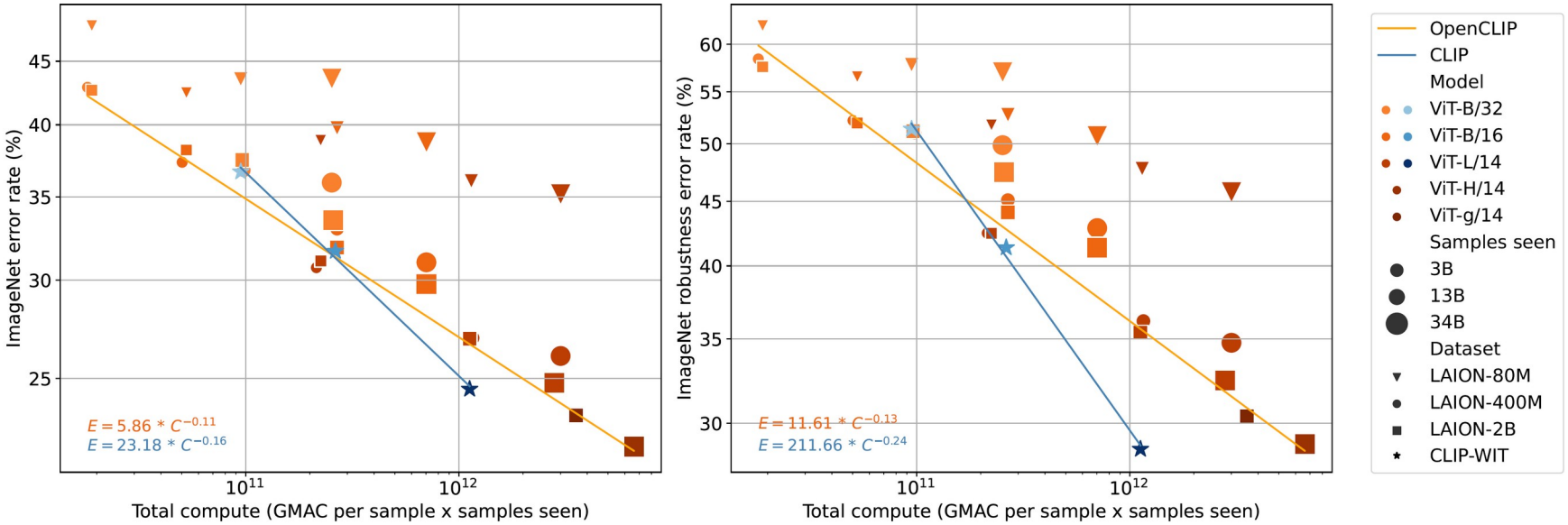
Dataset	# English Img-Txt Pairs
Public Datasets	
MS-COCO	330K
CC3M	3M
Visual Genome	5.4M
WIT	5.5M
CC12M	12M
RedCaps	12M
YFCC100M	100M[2]
LAION-5B (Ours)	**2.3B**
Private Datasets	
CLIP WIT (OpenAI)	400M
ALIGN	1.8B
BASIC	6.6B

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

Schuhmann et al, NeurIPS, 2022

Reproducible scaling laws for foundation models

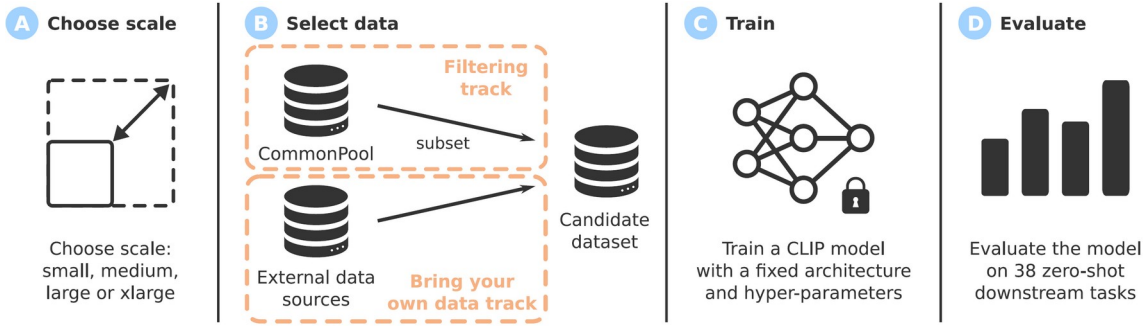
- Scaling laws with LAION-400M/2B and openCLIP: open-source data, models and code - **reproducible** science of foundation models

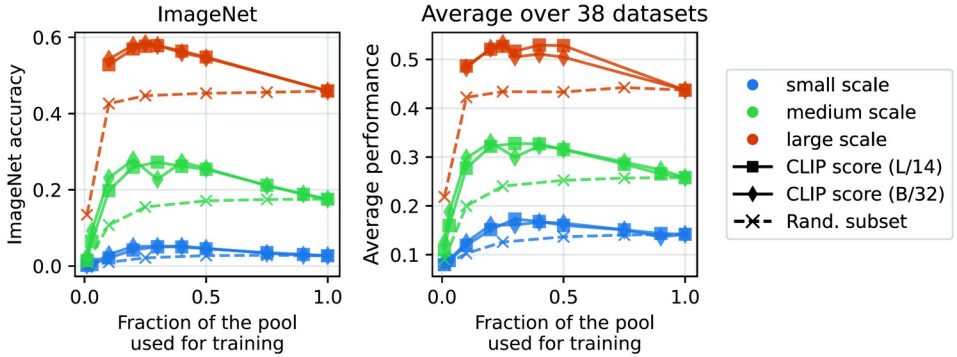


Schuhmann et al, NeurIPS, 2022; Cherti et al, CVPR 2023

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: 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 (open-LM-1B/7B)

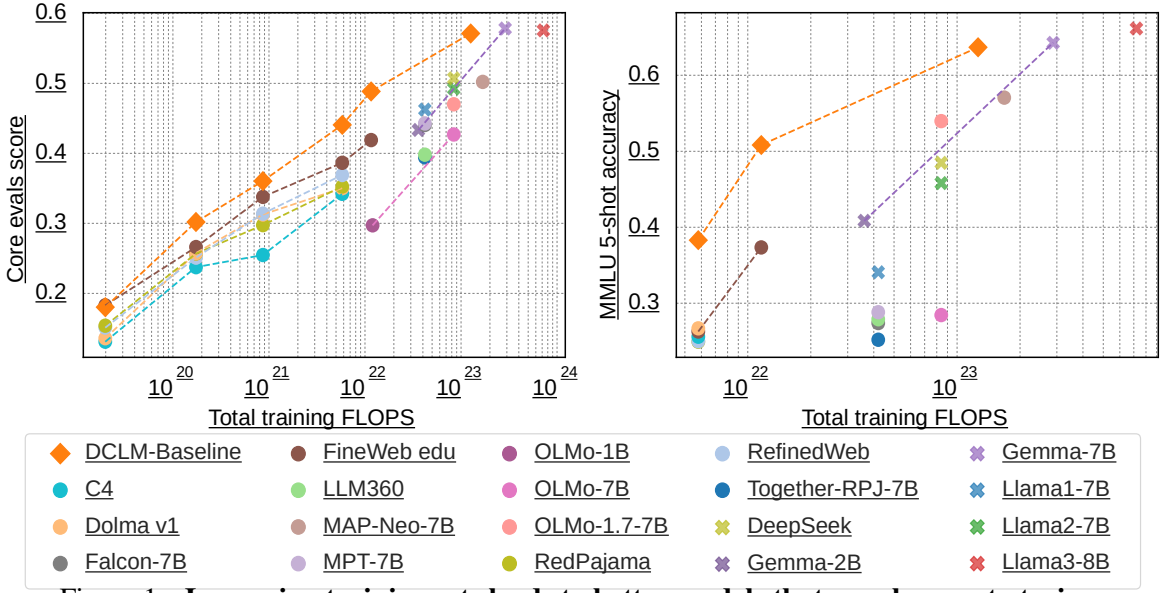


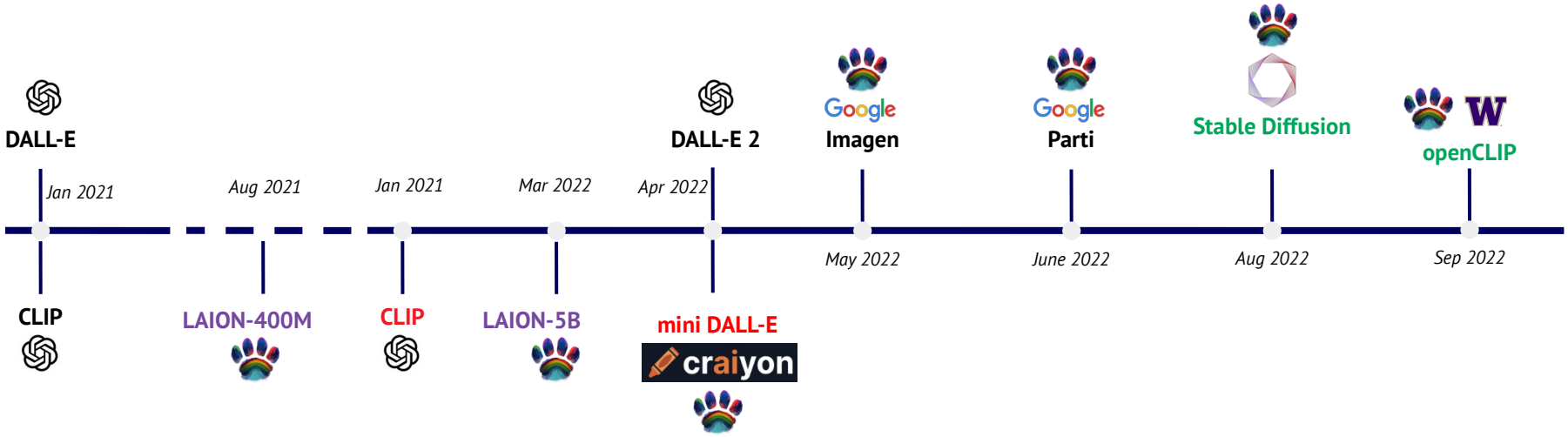
Figure 1: **Improving training sets leads to better models that are cheaper to train.**

Li et al, ArXiv:2406.11794, NeurIPS, 2024

https://github.com/mlfoundations/dclm

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

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

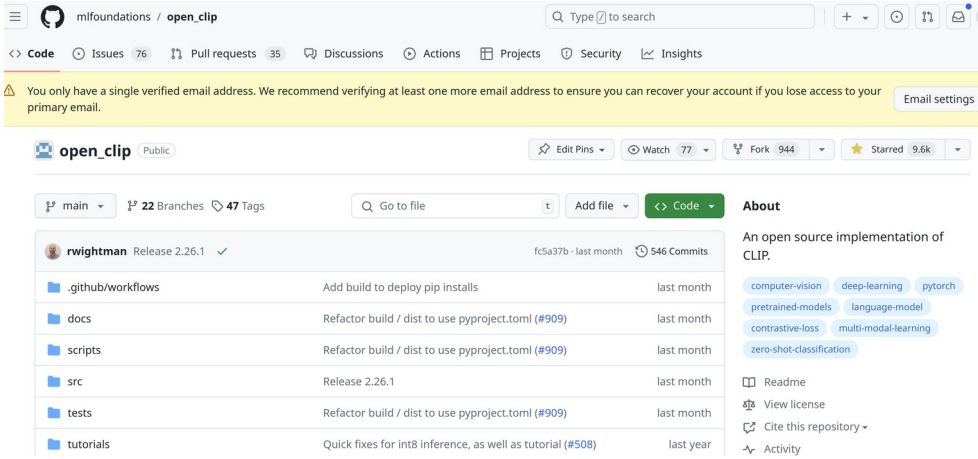
laion/CLIP-ViT-g-14-laion2B-s12B-b42K
Updated Feb 23 • ↓ 38.2k • ♡ 39

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

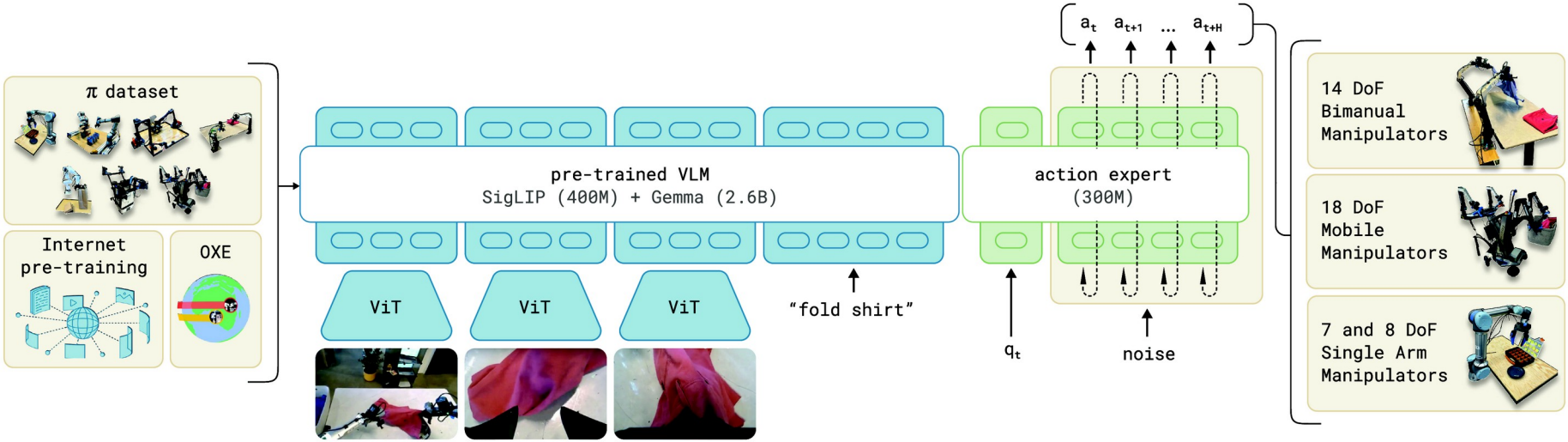
laion/CLIP-ViT-B-16-laion2B-s34B-b88K
Zero-Shot Image Classification • Updated Apr 19, 2023 • ↓ 5.81M • ♡ 27

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



Foundation models as re-usable components

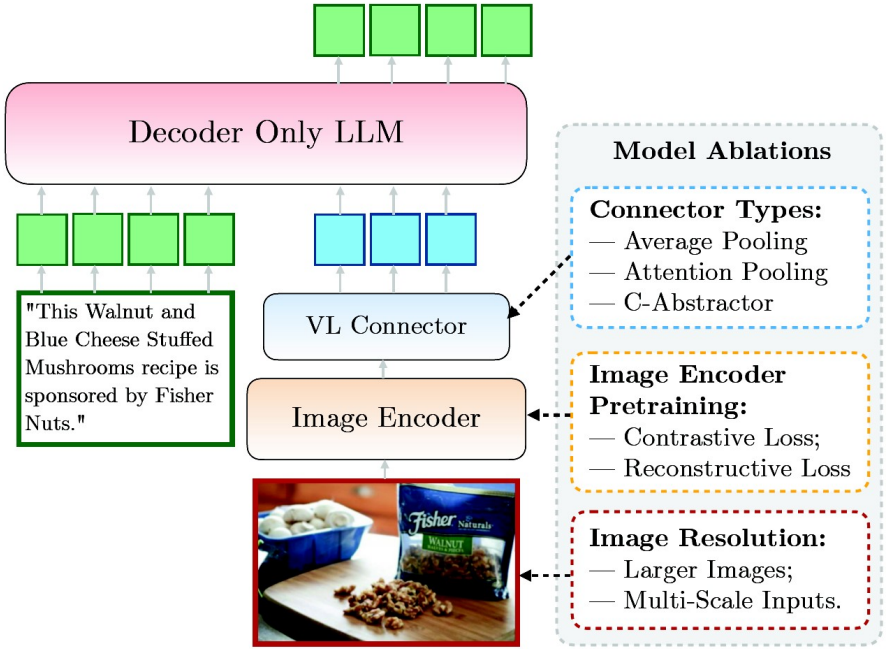
- Combining pre-trained foundation models for multi-modal generalist function (no or little adaptation required): Flamingo, BLIP-2, ImageBind, LENS, LlaVA, EMU, MM-1, PaliGemma, ...

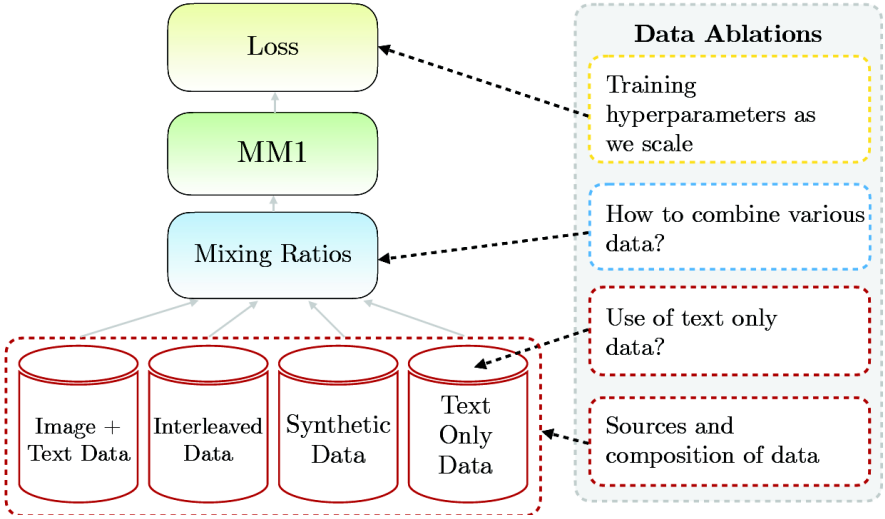


Black et al, 2024

Foundation models with strong generalization

- Systematic search for multimodal generalist foundation models with **improved scalability and generalization**: Open-Sci (Open-ψ), Project Nucleus

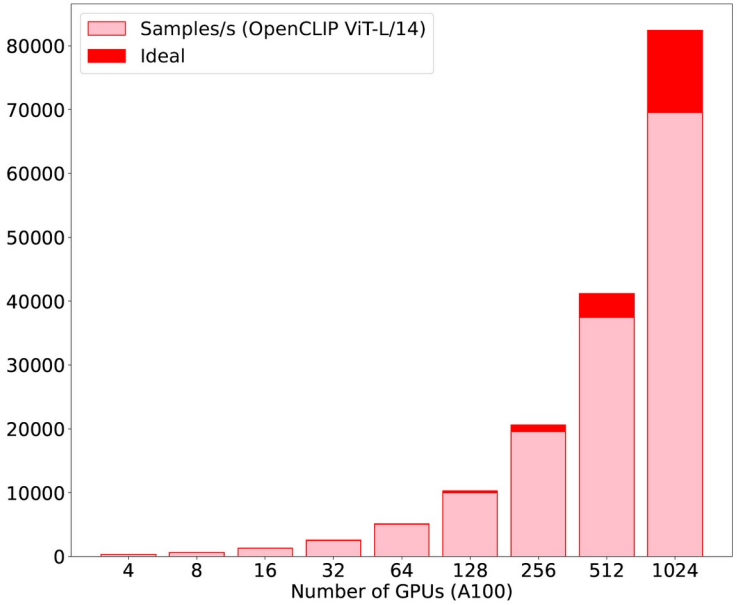


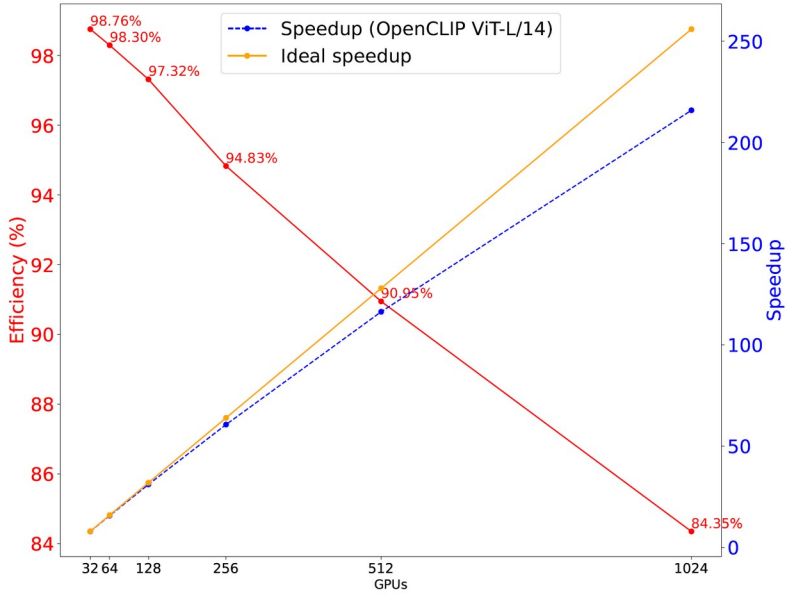


McKinzie et al, 2024

Supercomputers for foundation model research

- Supercomputers: necessary for scaling law derivation (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“

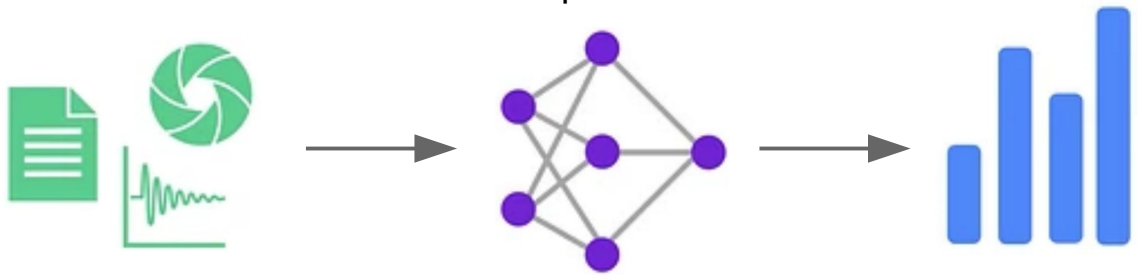
Open foundation models: improving scaling

- Interventions along the **whole reproducible pipeline** – dataset composition, model training, benchmarks & evaluation – to boost scaling

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 to measure generalization, transfer capability

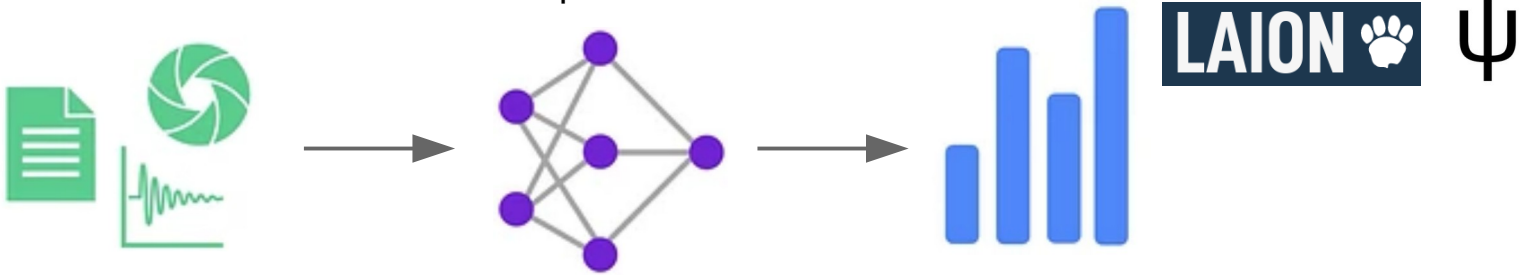
Open foundation models: improving scaling

- 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

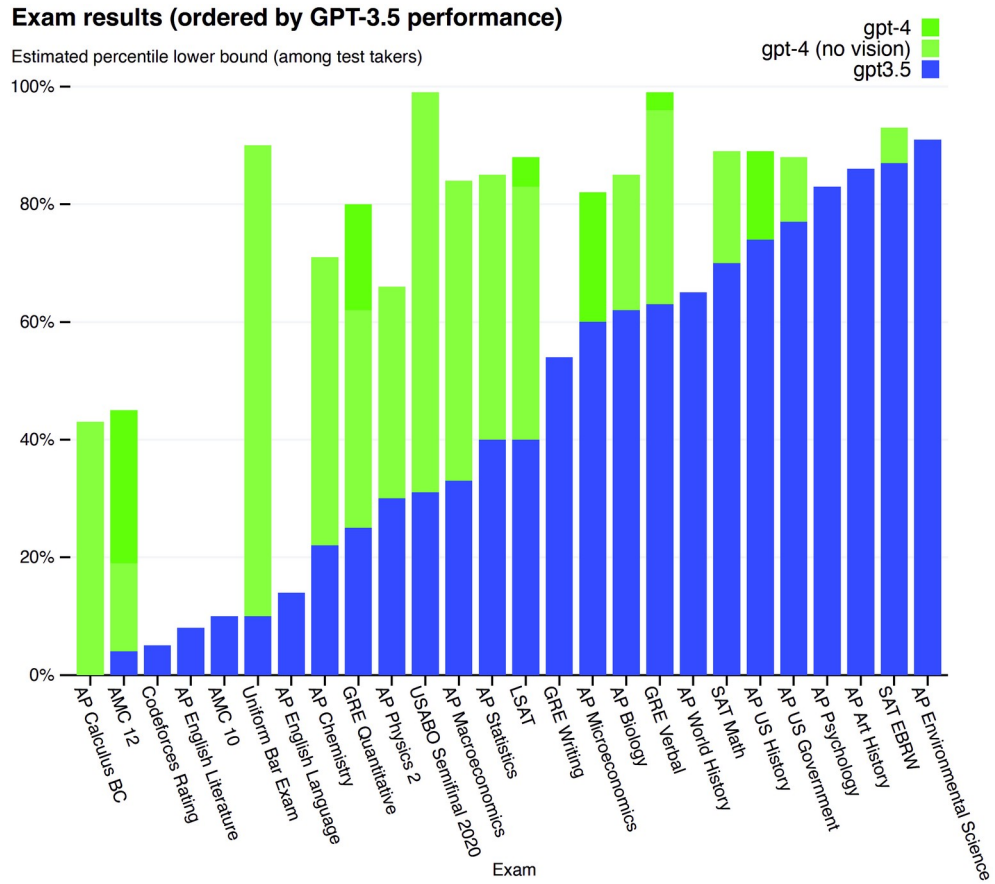
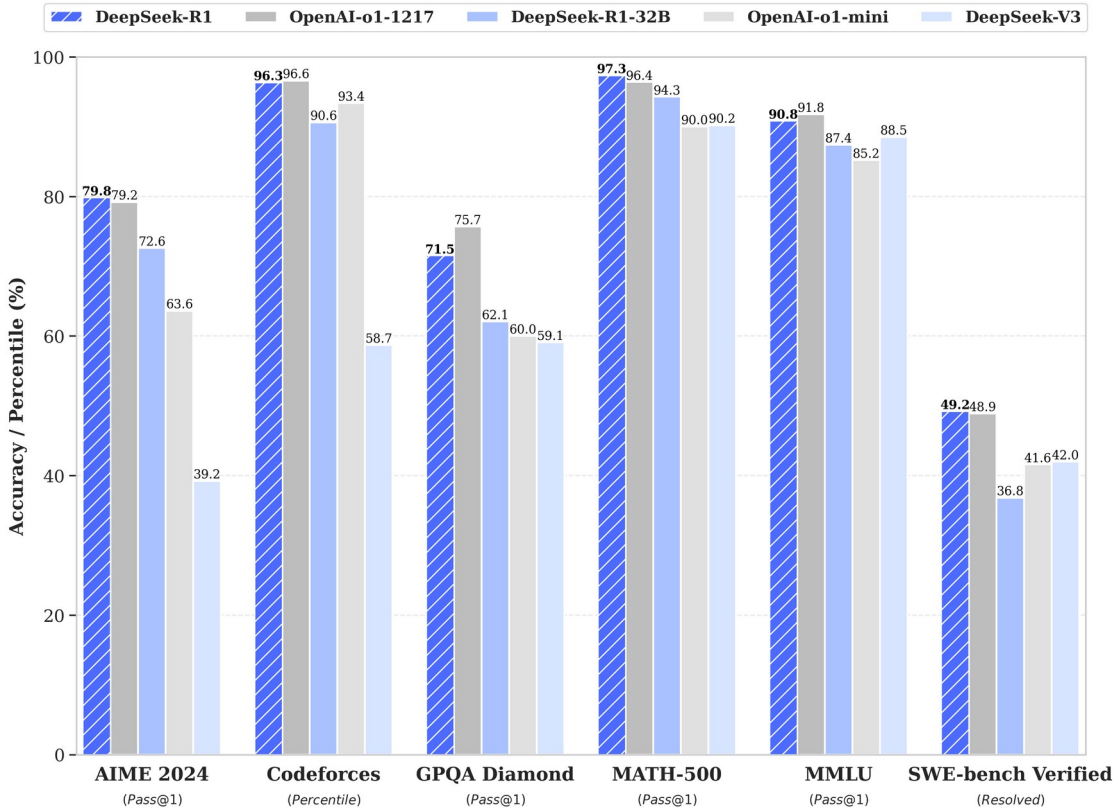


Figure 4. GPT performance on academic and professional exams. In each case, we simulate the conditions and scoring of the real exam. Exams are ordered from low to high based on GPT-3.5 performance. GPT-4 outperforms GPT-3.5 on most exams tested. To be conservative we report the lower end of the range of percentiles, but this creates some artifacts on the AP exams which have very wide scoring bins. For example although GPT-4 attains the highest possible score on AP Biology (5/5), this is only shown in the plot as 85th percentile because 15 percent of test-takers achieve that score.

GPT-4 Technical Report, openAI, 2023

Scaling laws: predicting generalization

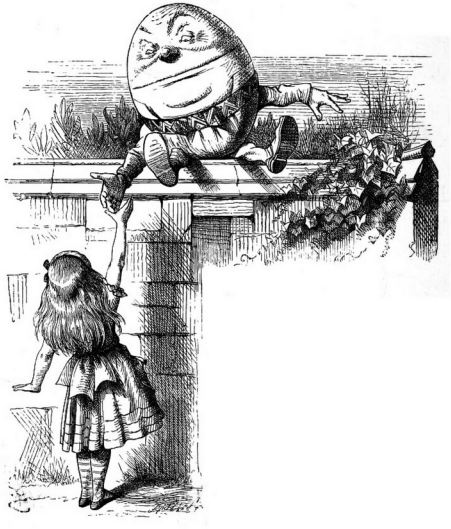
- Do benchmark downstream tasks reflect generalization properly?



DeepSeek R1 Technical Report, DeepSeek, 2025

Scaling laws: predicting generalization

- Do benchmark downstream tasks reflect generalization properly?
- Big issue: insensivity to basic deficits, eg lack of robustness to problem variations
- Big issue: test set leakage, 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?

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

LAION

open-sci collective

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

Foundation models: scaling laws & generalization

- Questions Break

Scaling laws: predicting generalization

- Using variations of simple problem templates to measure model robustness

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**]

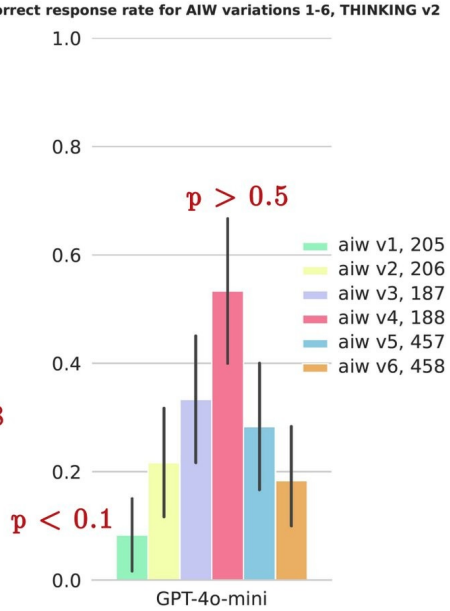
How many sisters does Alice's brother have?

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

- 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

AIW Ext Alice and Bob, Alice's Brothers, Variations 1-6

Alice and Bob are sister and brother.

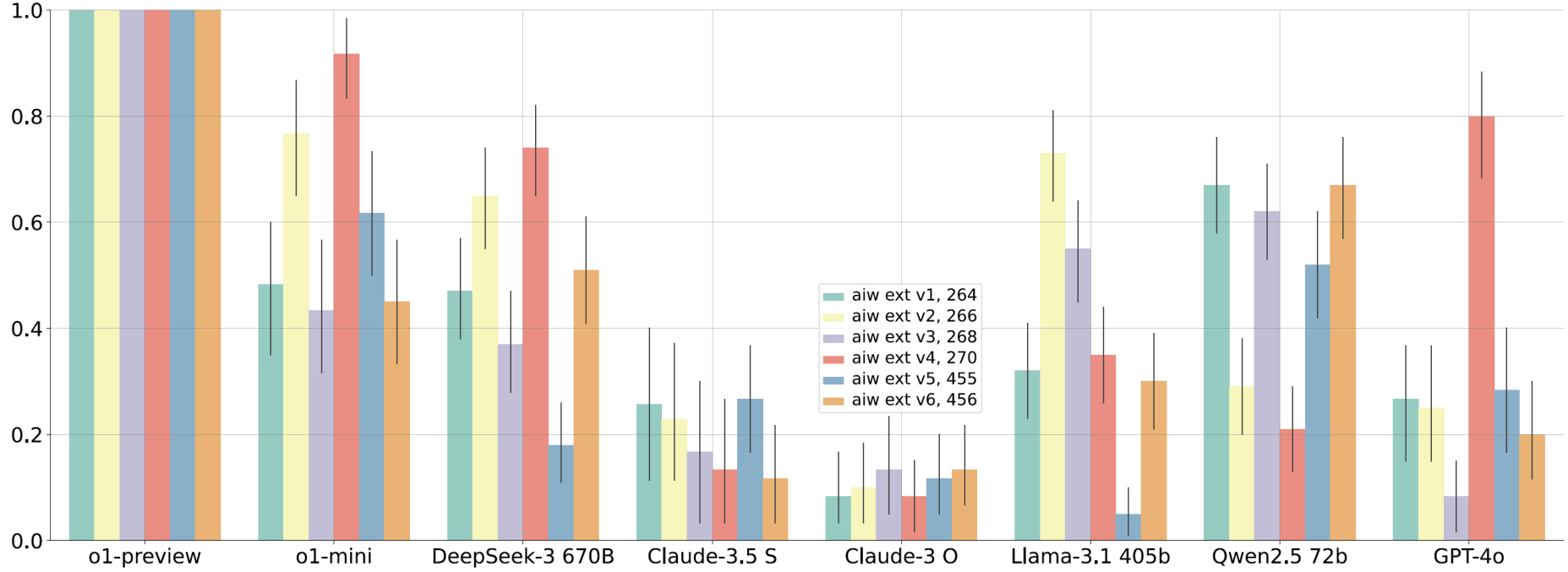
Variation 1: Alice has **3 sisters** and Bob has **6 brothers**. [Correct answer: **7**]
Variation 2: Alice has **2 sisters** and Bob has **2 brothers**. [Correct answer: **3**]
Variation 3: Alice has **1 sister** and Bob has **4 brothers**. [Correct answer: **5**]
Variation 4: Alice has **3 sisters**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

How many brothers does Alice 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 Friends, Variations 1-6, Prompt IDs: 577 580 581 582 583 584

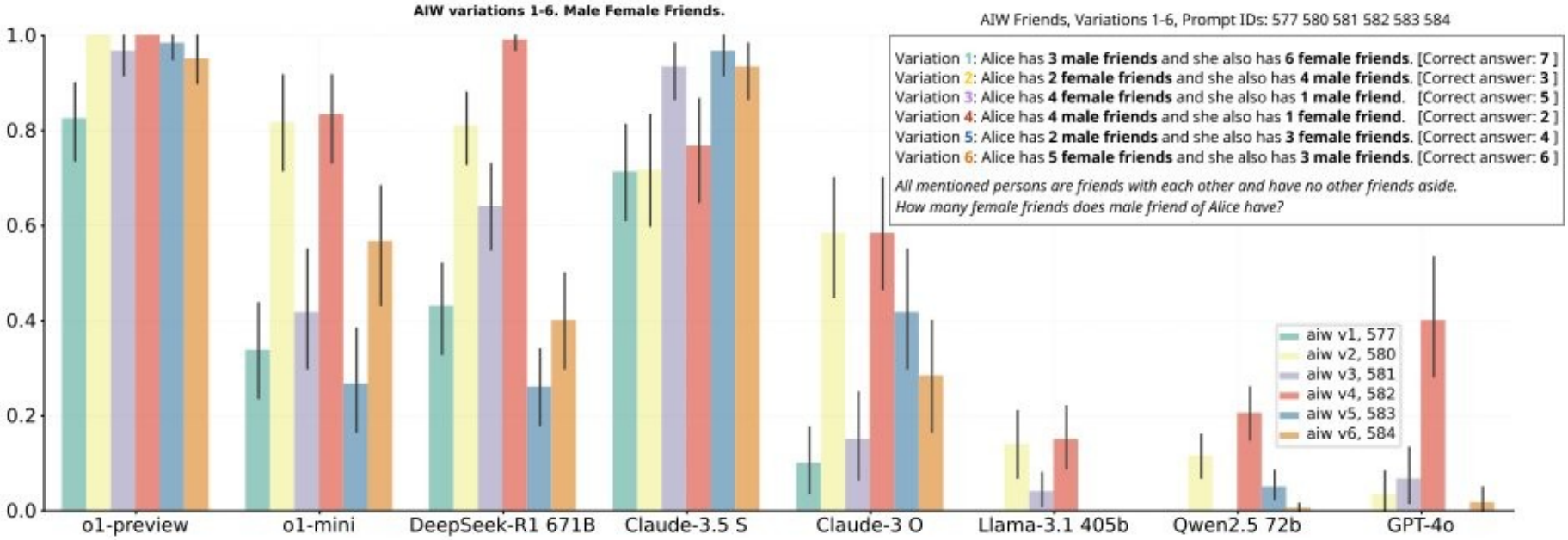
Variation 1: Alice has **3 male friends** and she also has **6 female friends**. [Correct answer: **7**]
Variation 2: Alice has **2 female friends** and she also has **4 male friends**. [Correct answer: **3**]
Variation 3: Alice has **4 female friends** and she also has **1 male friend**. [Correct answer: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

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

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

Scaling laws: predicting generalization

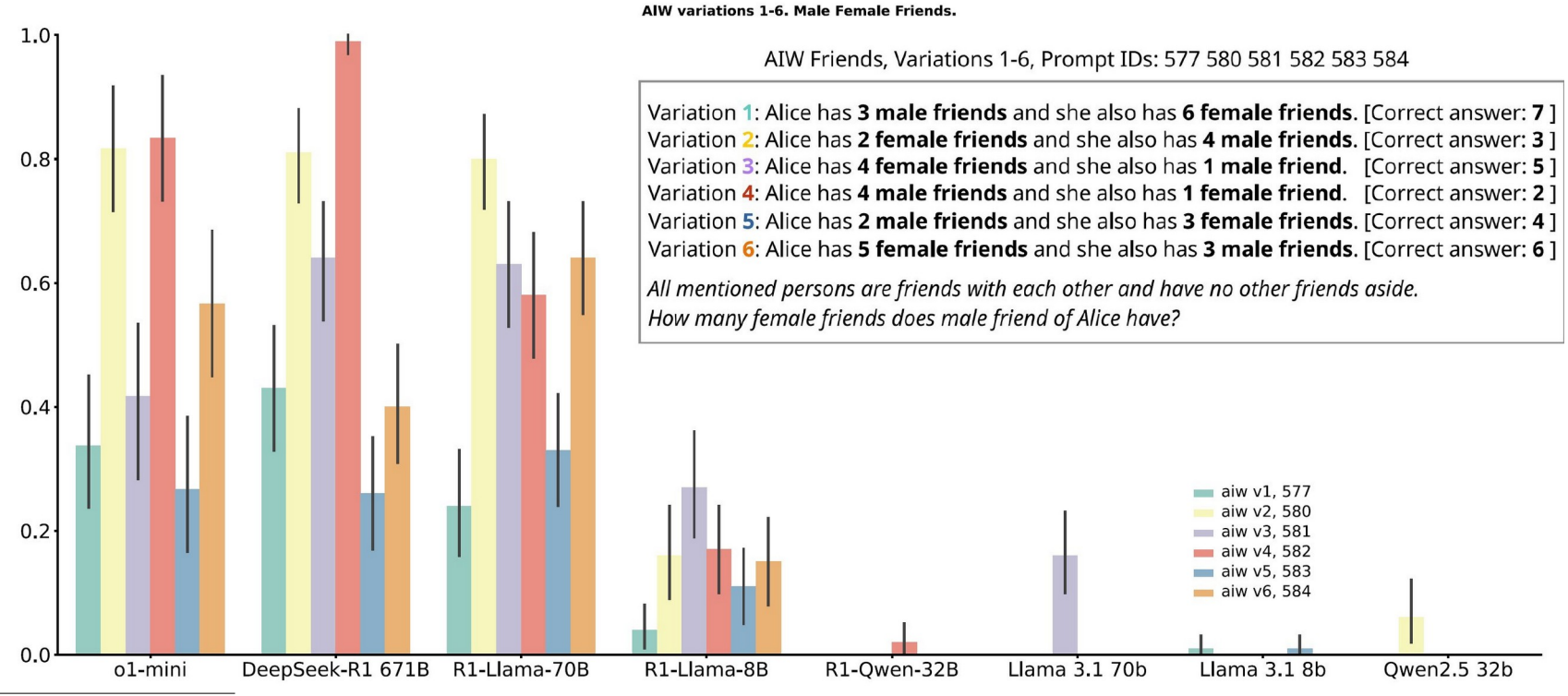
- SOTA LLMs show strong fluctuations across variations that DO NOT CHANGE problem structure at all



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



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 1-6. THINKING v2, Prompt IDs: 559 560 561 562 563 564

Variation 1: Alice has ***1*** sister and ***1*** brother in total. Her mother has 2 brothers. She also has 1 sister who does not have children and who has ***6*** nephews and nieces in total. Alice's father has 2 sisters. He also has a brother who has ***5*** nephews and nieces in total, and who also has ***2 sons***. [Correct answer: ***7***]

Variation 2: Alice has ***2*** sisters and ***1*** brother in total. Her mother has 2 brothers. She also has 1 sister who does not have children and who has ***6*** nephews and nieces in total. Alice's father has 2 sisters. He also has a brother who has ***4*** nephews and nieces in total, and who also has ***1 son***. [Correct answer: ***3***]

Variation 3: Alice has ***2*** sisters and ***1*** brother in total. Her mother has 2 brothers. She also has 1 sister who does not have children and who has ***7*** nephews and nieces in total. Alice's father has 2 sisters. He also has a brother who has ***5*** nephews and nieces in total, and who also has ***1 son***. [Correct answer: ***5***]

Variation 4: Alice has ***1*** sister and ***3*** brothers in total. Her mother has 2 brothers. She also has 1 sister who does not have children and who has ***6*** nephews and nieces in total. Alice's father has 2 sisters. He also has a brother who has ***5*** nephews and nieces in total, and who also has ***1 daughter***. [Correct answer: ***2***]

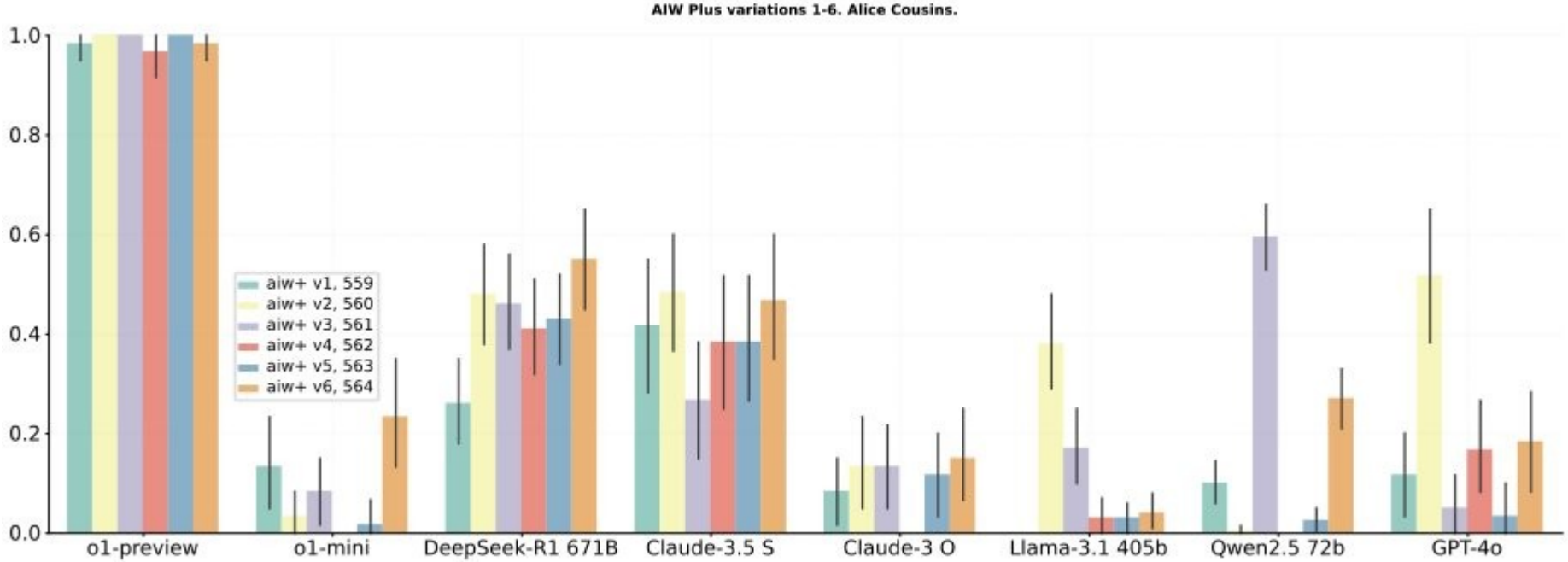
Variation 5: Alice has ***2*** sisters and ***1*** brother in total. Her mother has 2 brothers. She also has 1 sister who does not have children and who has ***6*** nephews and nieces in total. Alice's father has 2 sisters. He also has a brother who has ***5*** nephews and nieces in total, and who also has ***1 son***. [Correct answer: ***4***]

Variation 6: Alice has ***1*** sister and ***1*** brother in total. Her mother has 2 brothers. She also has 1 sister who does not have children and who has ***6*** nephews and nieces in total. Alice's father has 2 sisters. He also has a brother who has ***5*** nephews and nieces in total, and who also has ***1 daughter***. [Correct answer: ***6***]

How many cousins does Alice's sister have?

Scaling laws: predicting generalization

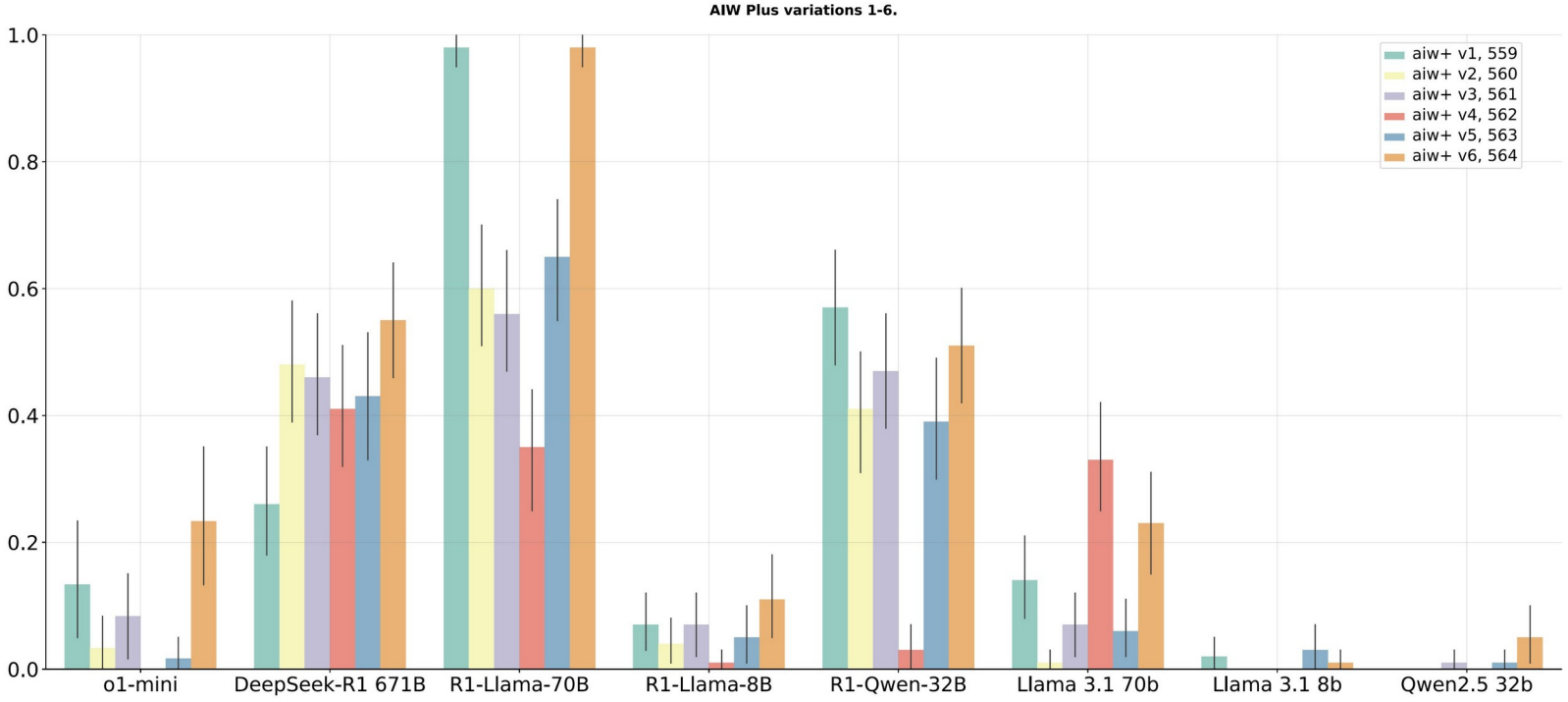
- SOTA LLMs show strong fluctuations across variations that DO NOT CHANGE problem structure at all



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



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

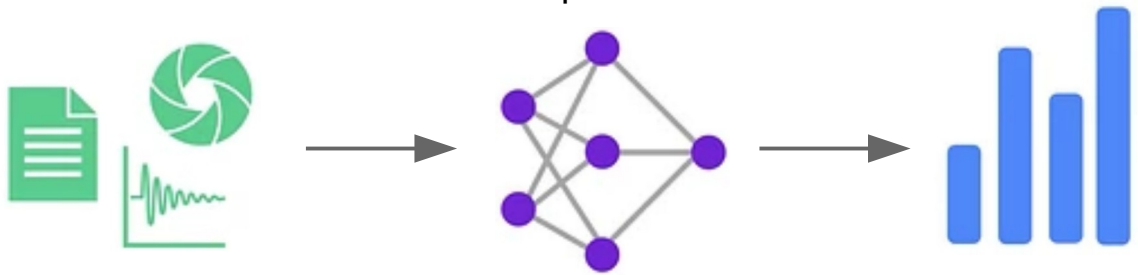
Open foundation models: improving scaling

- Interventions along the **whole reproducible pipeline** – dataset composition, model training, benchmarks & evaluation – to boost scaling

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 to measure generalization, transfer capability

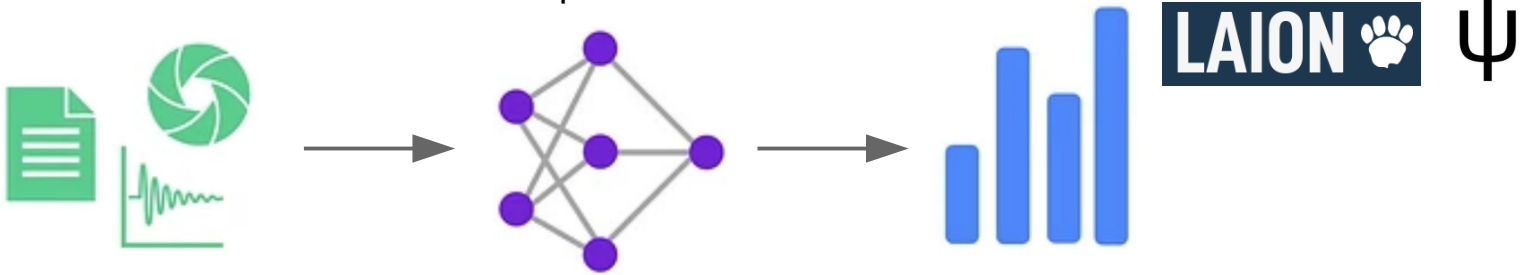
Open foundation models: improving scaling

- 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!

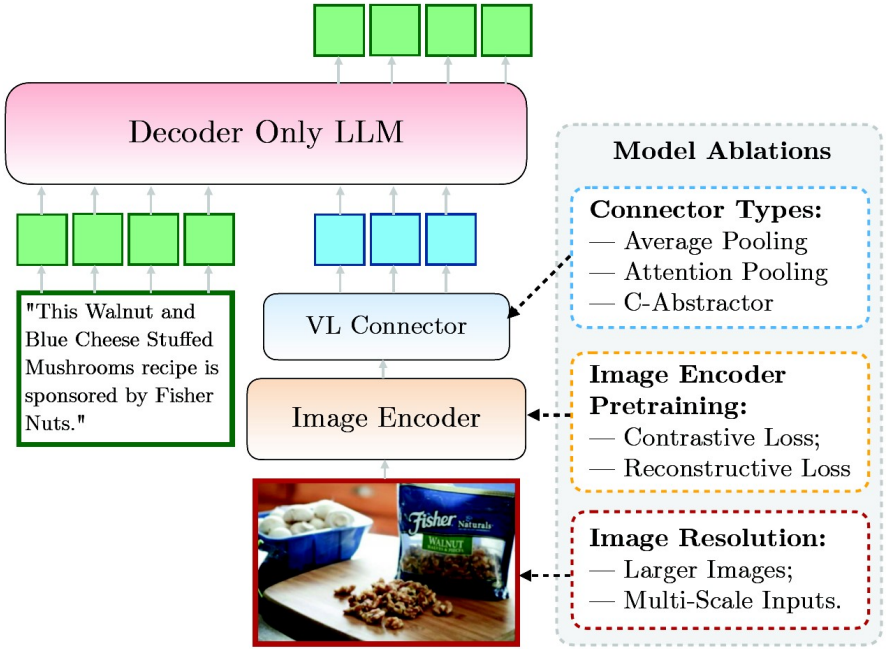
Dataset composition studies, scaling laws

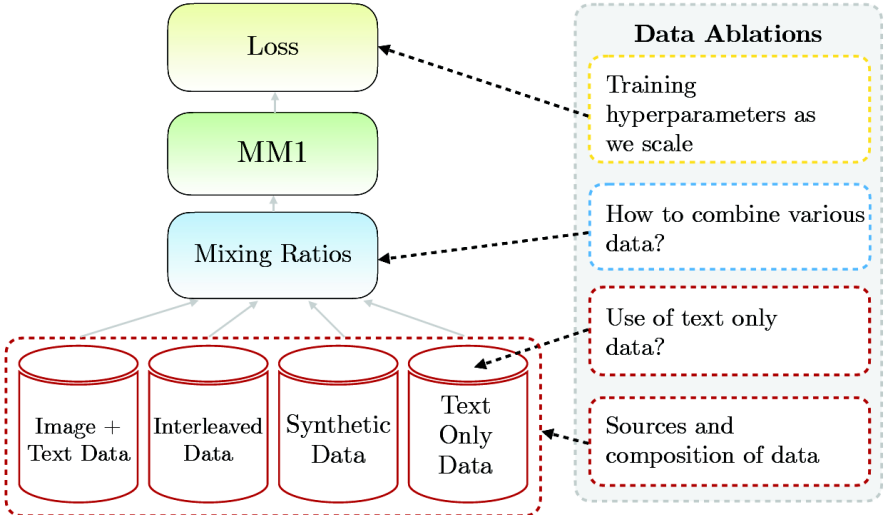
Learning procedure studies, scaling laws

Novel benchmarks for model capabilities, transfer

Foundation models with strong generalization

- Systematic search for multimodal generalist foundation models with **improved scalability and generalization**: Open-Sci (Open-ψ), Project Nucleus





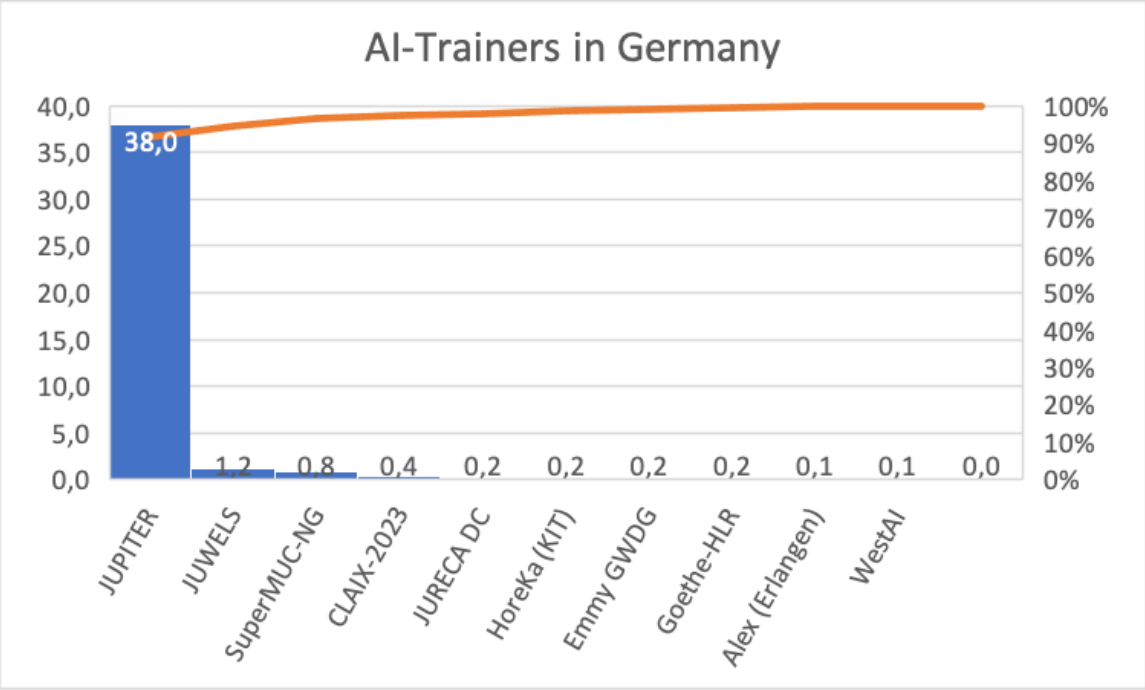
McKinzie et al, 2024

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 (Q2 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

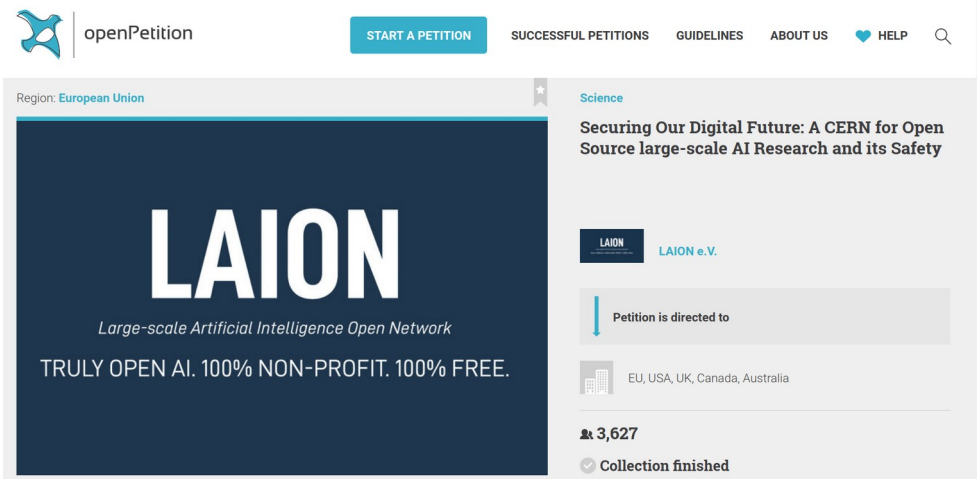
- Now: GPT-4 level models (already 1 year old), language only: ca. 10M H100 GPU hours → ca. 2.5 weeks on **whole** JUPITER for **single training run ...**

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

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

LAION: research community & alliances

- Various alliances in EU: **ELLIS**, Tuebingen AI Center, MPI for Intelligent Systems, WestAI, U Freiburg, HuggingFace, U Turku (**HPLT**), …
- Various alliances worldwide: Stanford, U Washington, Allen AI Institute, Together AI, Ontocord AI, Tokyo Tech, U Berkeley, U Tel Aviv, ...



LAION: strong grassroot research community

- Collaborative work of broadly distributed community: **Outstanding NeurIPS 2022 paper award**, strongly impacting open source releases
- **Falling Walls Award**: Scientific Breakthrough 2023
- LAION public Discord server: > **35k members**

Open foundation models: outlook

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

- **OpenEuroLLM – LAION/ELLIS/HPLT & friends** : EU consortium for building open foundation models with strongly improved generalization & reasoning
  - Will deliver the strong language component for open-sci-MM
    → Join us! Multiple open ML reseacher (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, UoW

Christoph Schumann

Let‘s build open, robust, safe AI foundations together!

BigScience

EleutherAI

LAION
Large-scale Artificial Intelligence Open Network

Open-ψ (open-sci) Collective

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