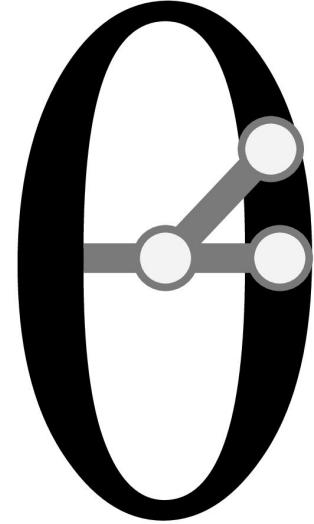
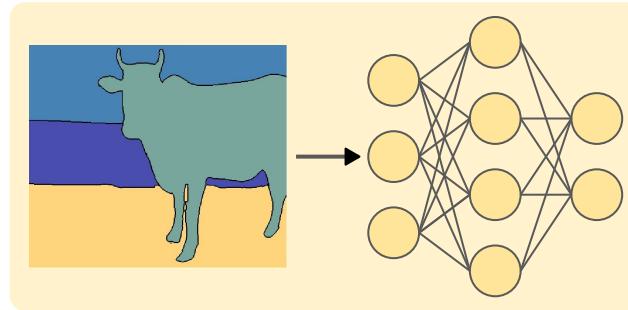
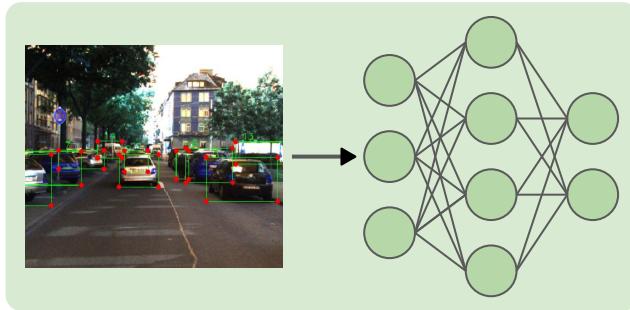
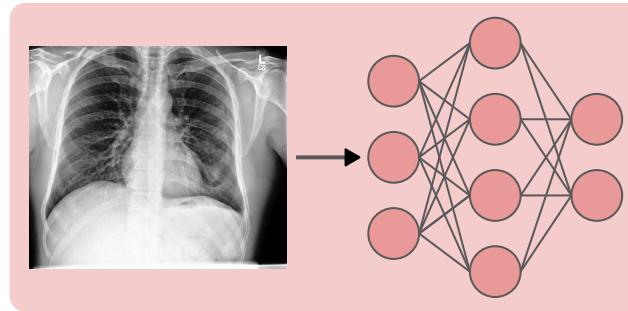
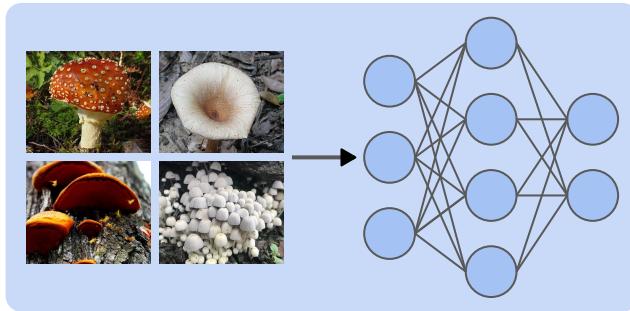


Building Machine Learning Models like Open-Source Software with git-theta

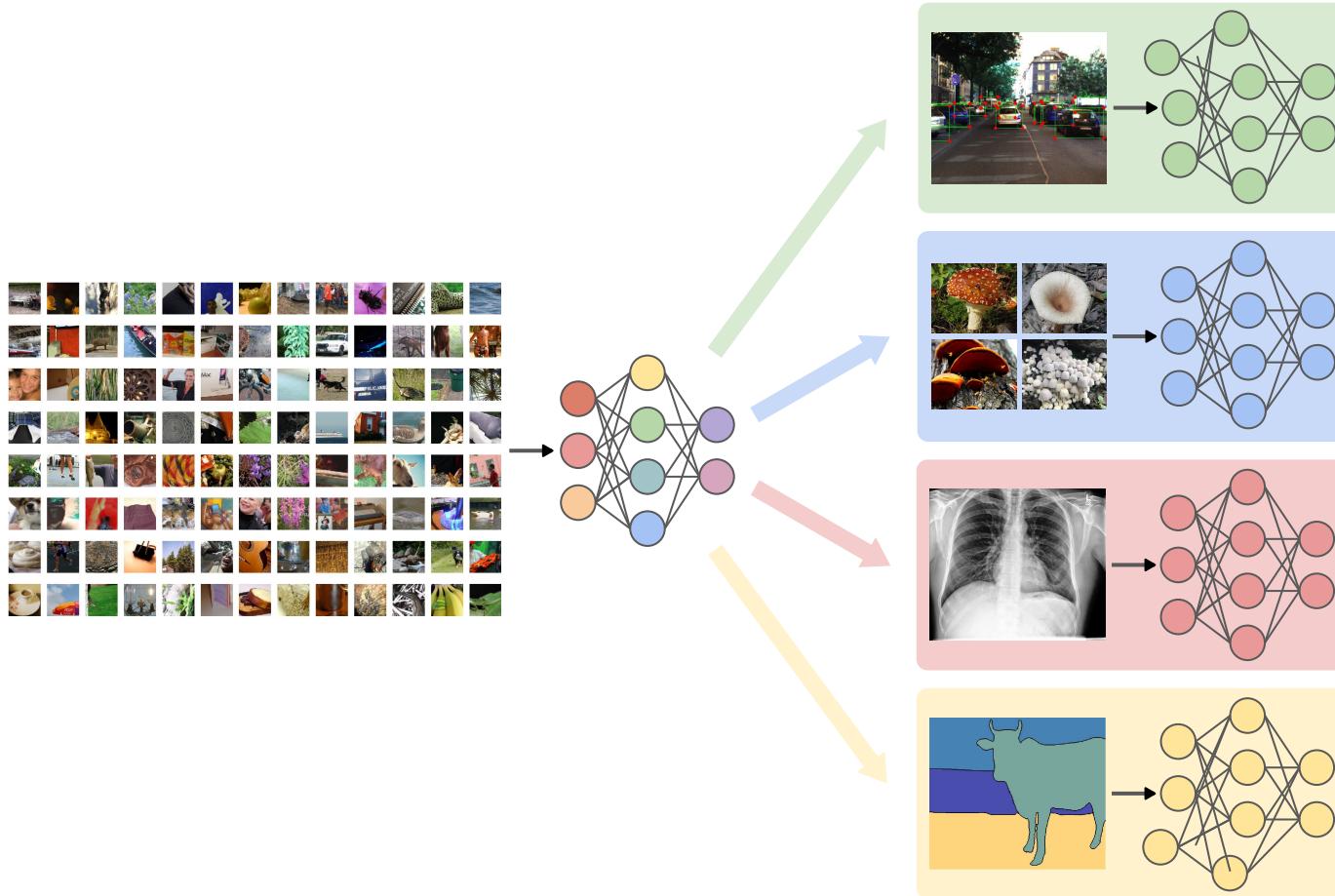


Nikhil Kandpal & Colin Raffel

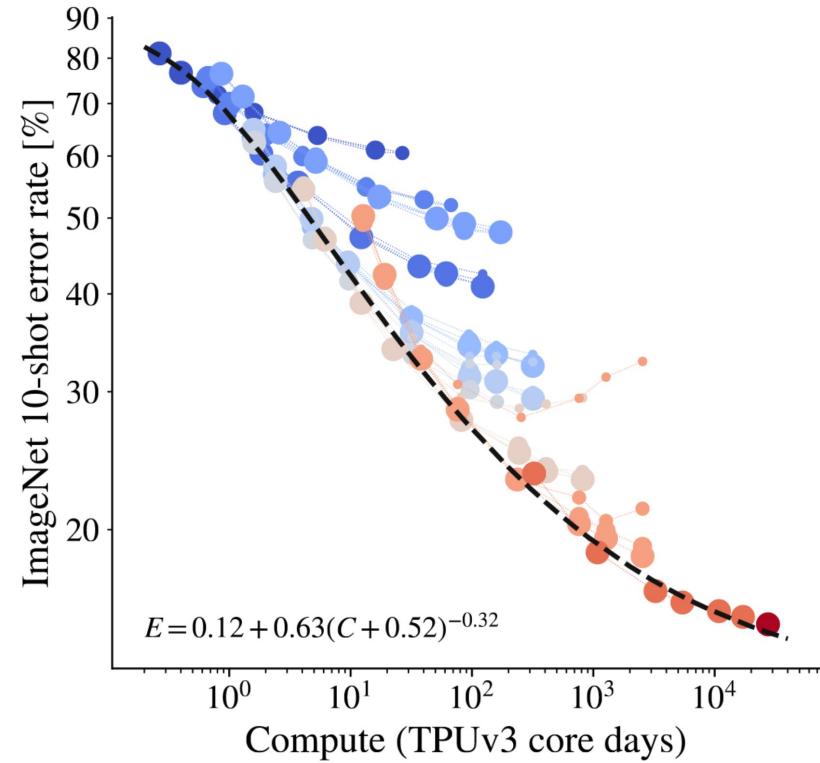
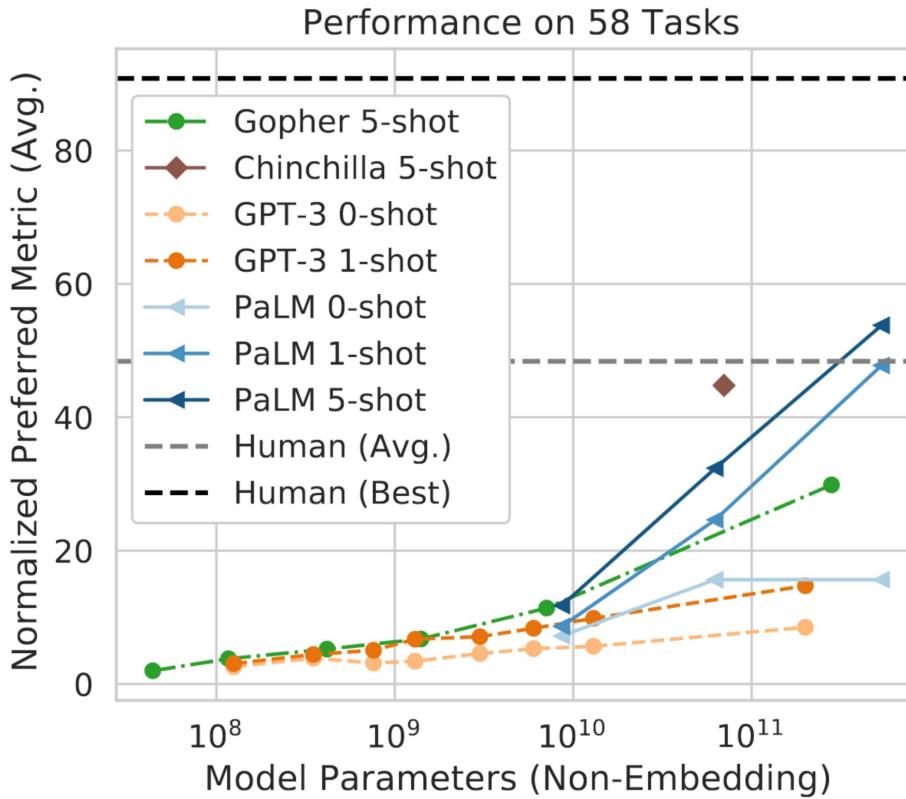
Deep learning circa 2013 – training models from scratch



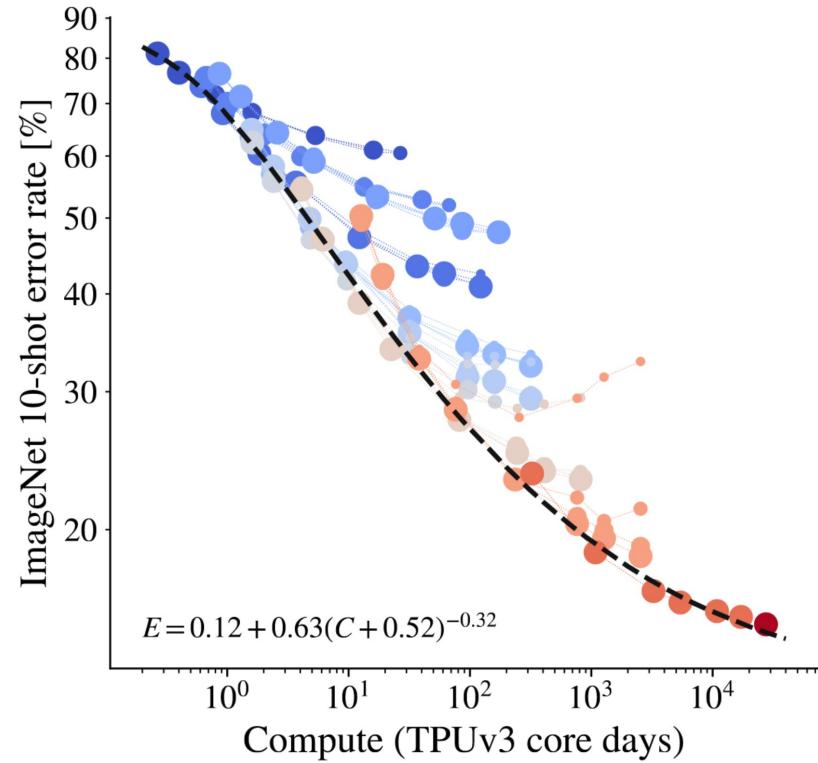
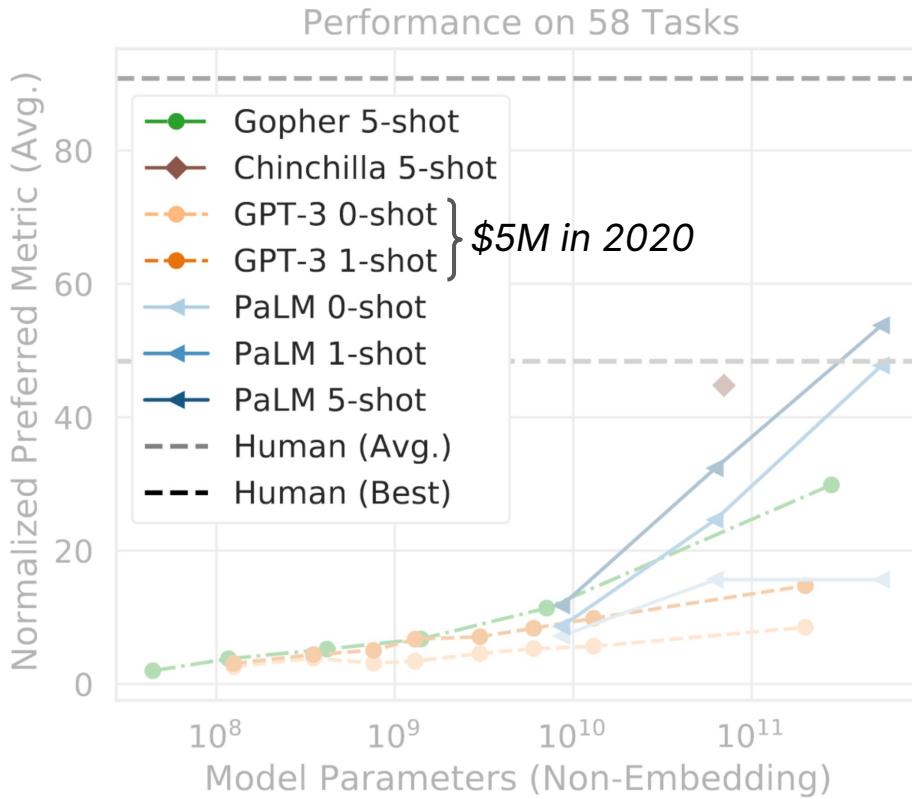
Deep learning in 2023 – pre-train then adapt



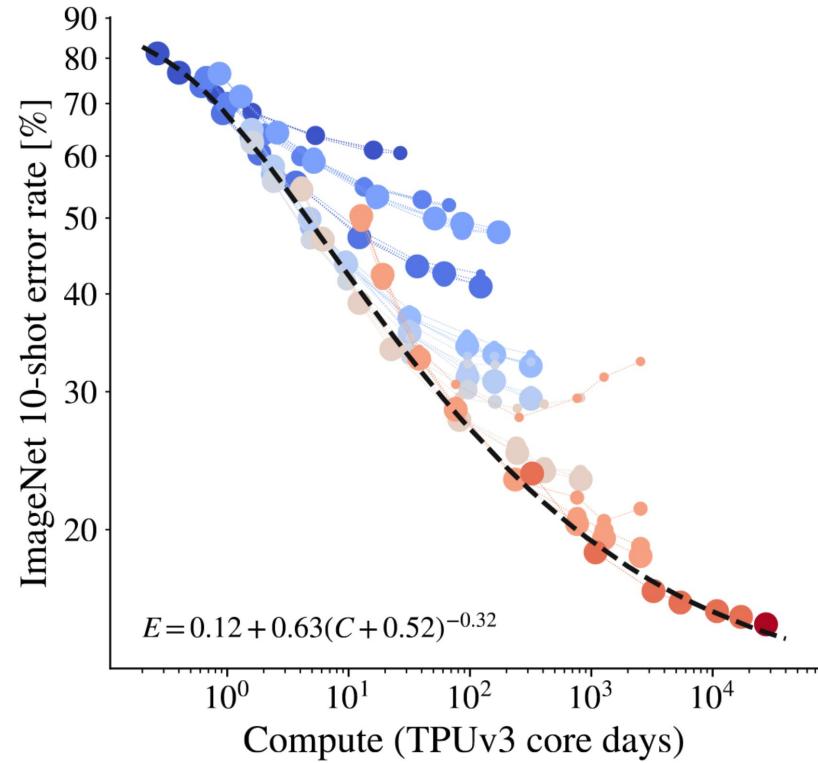
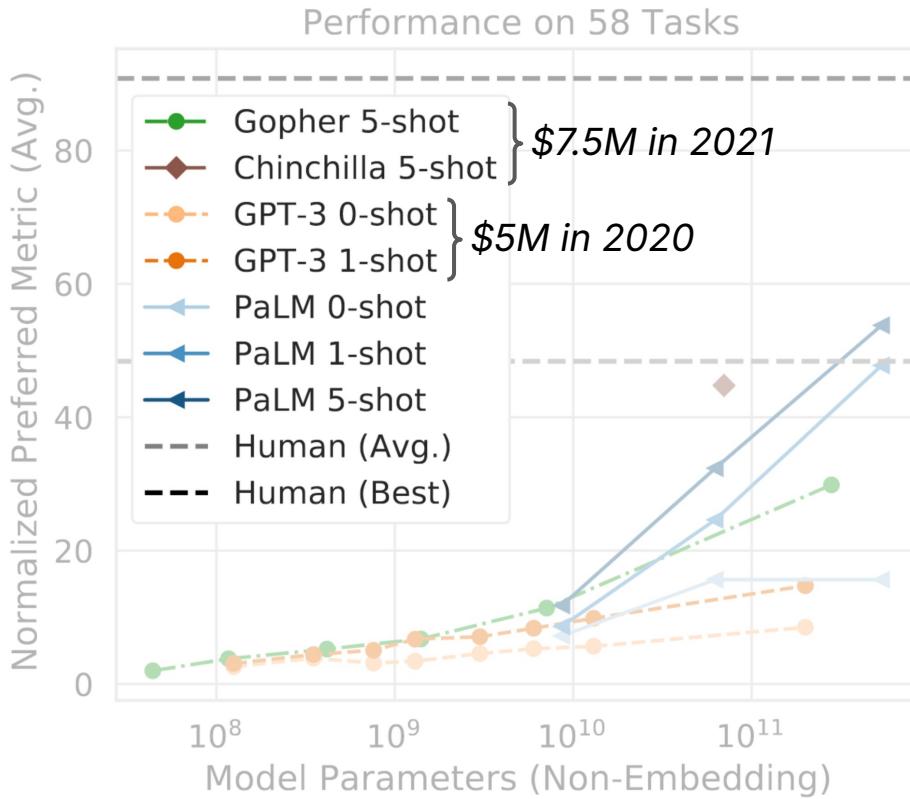
The benefits – and costs – of scale



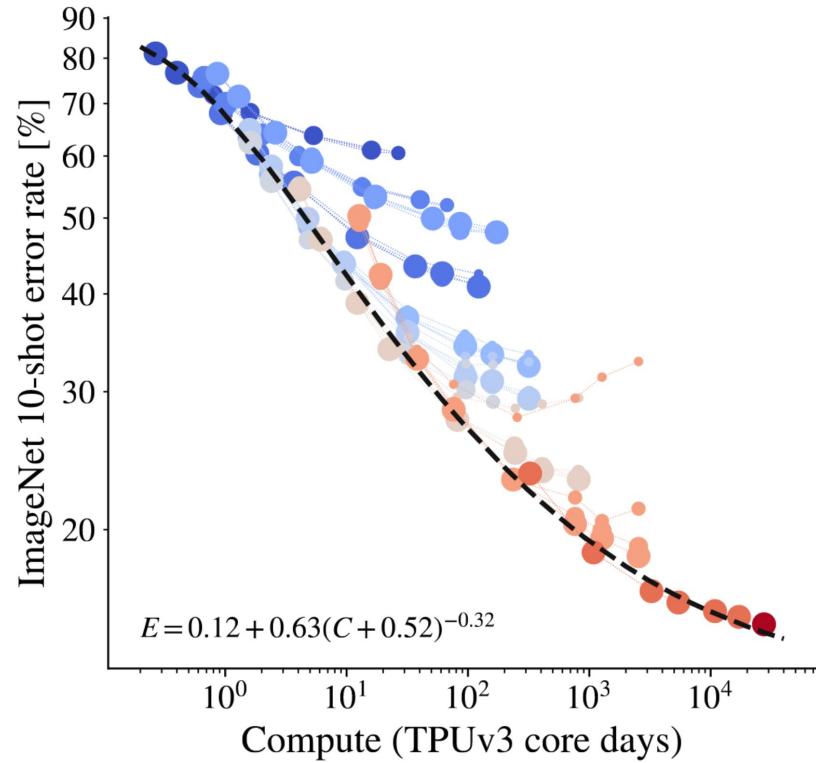
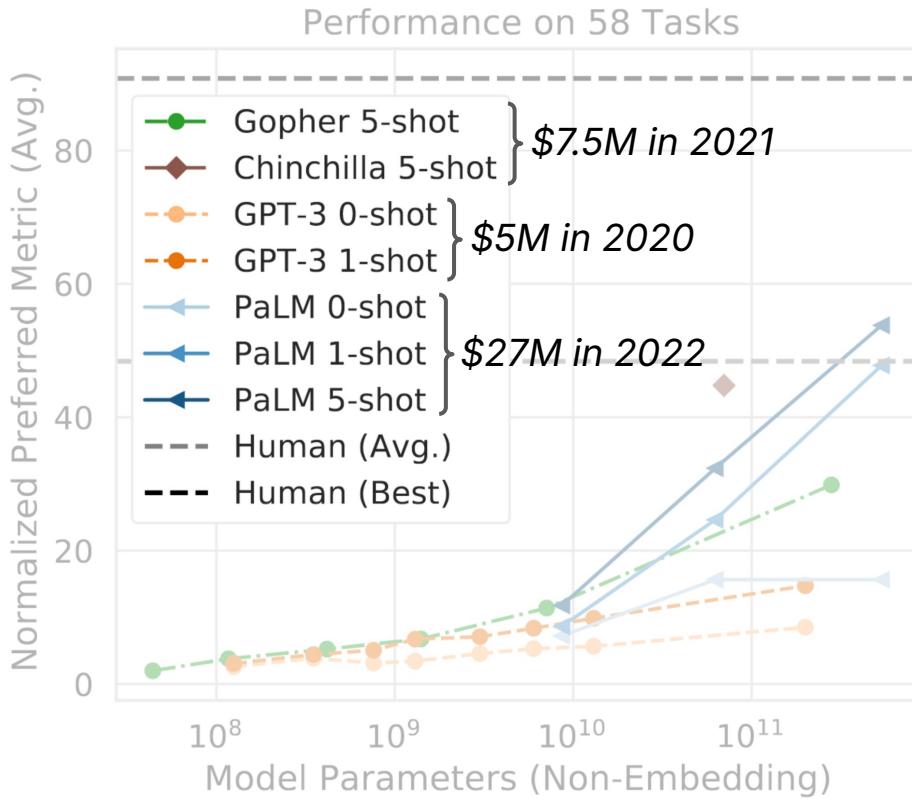
The benefits – and costs – of scale



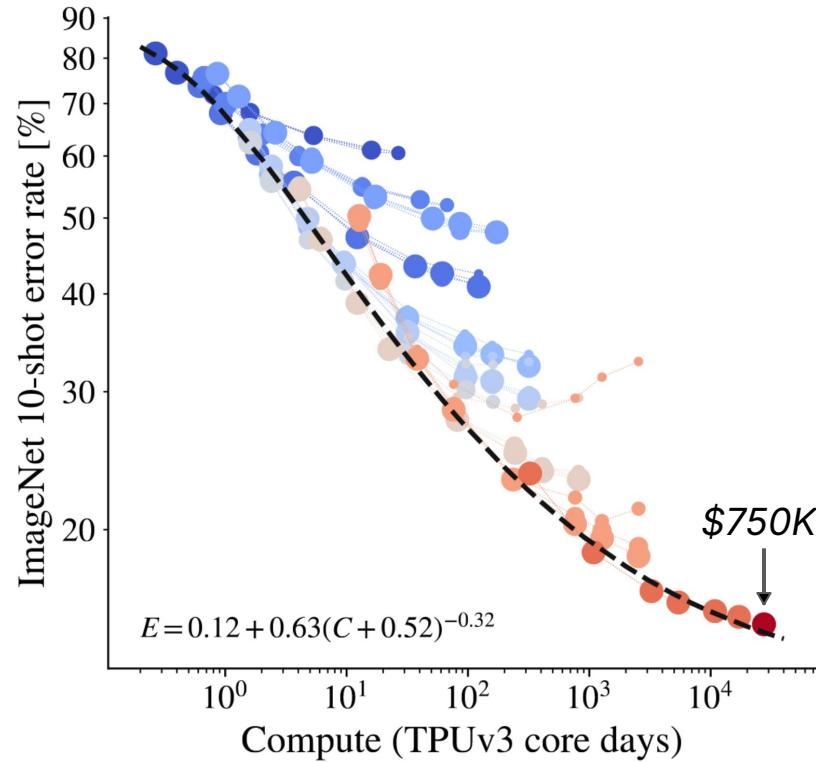
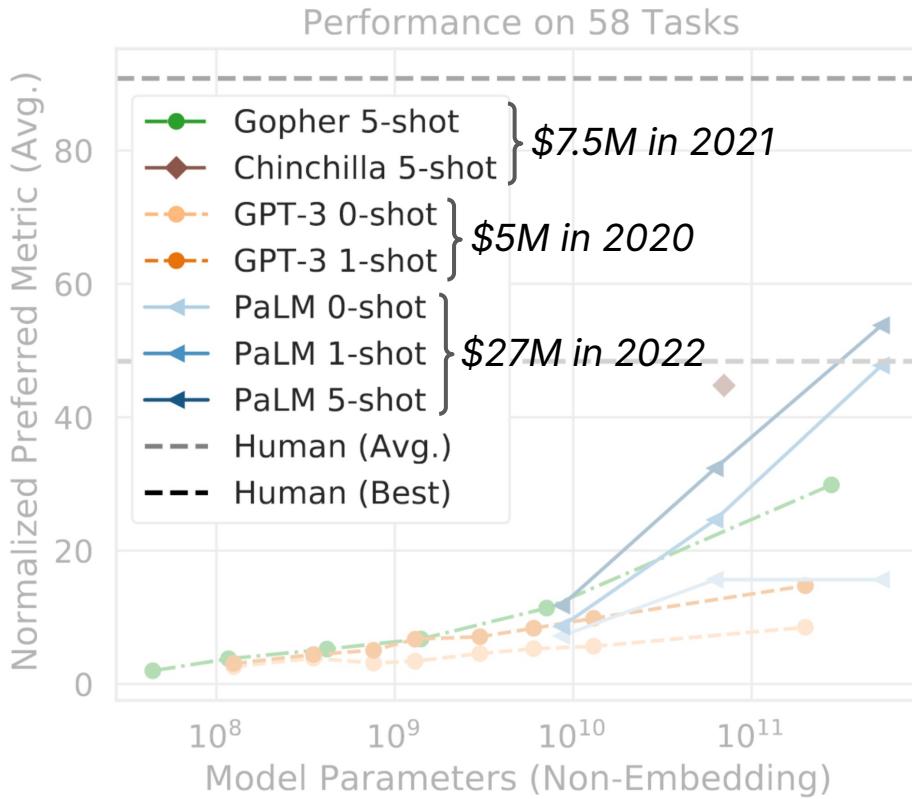
The benefits – and costs – of scale



The benefits – and costs – of scale



The benefits – and costs – of scale



Increased costs have decreased sharing

Introducing the LightOn Muse API

Create. Process. Understand. Learn.

Production-ready intelligence primitives powered by state-of-the-art language models. For the first time natively in French, Spanish, Italian, and more. Now in private beta!

AI21 studio Docs Pricing Start Building

Differentiate your product with generative text AI

AI21 Studio provides API access to Jurassic-1 large-language-models. Our models power text generation and comprehension features in thousands of live applications.

Google Research

BLOG

Pathways Language Model (PaLM): Scaling to 540 Billion Parameters for Breakthrough Performance

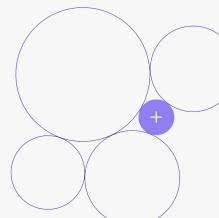
co:here Dashboard Documentation Playground Community Log In

Home Guides and Concepts API Reference Release Notes Search

Add Language AI capability to your system

Integrate state-of-the-art language models into your builds in just five minutes.

Get your API key



HyperCLOVA

Designing businesses to fill with ideas & creativity.

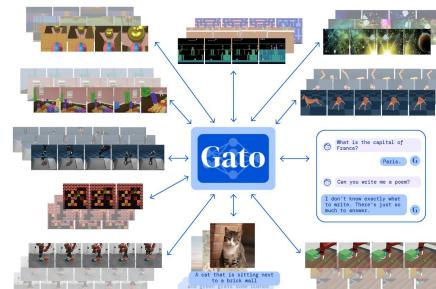
あなたならではの 想い、創造力を発揮できる余白をつくる

Imagen

unprecedented photorealism × deep level of language understanding

Microsoft Megatron-Turing NLG 530B

The World's Largest and Most Powerful Generative Language Model



OpenAI API Beta ABOUT EXAMPLES DOCS PRICING LOG IN JOIN >

OpenAI technology, just an HTTPS call away

Apply our API to any language task — semantic search, summarization, sentiment analysis, content generation, translation, and more — with only a few examples or by specifying your task in English.

Popular public models often come from resource-rich groups

The screenshot shows the Hugging Face Model Hub interface. At the top, there is a search bar with the placeholder "Search models, datasets, users..." and a logo for "Hugging Face". Below the search bar, a navigation bar includes a "Models" button (showing 33,490), a "Search Models" button, a "Add filters" button, and a "Sort: Most Downloads" button. The main content area displays a list of seven models, each in its own card:

- bert-base-uncased** (by Google): Fill-Mask, 30M, 54 downloads.
- roberta-large** (by Facebook): Fill-Mask, 13.1M, 20 downloads.
- distilbert-base-uncased** (by Hugging Face): Fill-Mask, 4.83M, 26 downloads.
- xlm-roberta-base** (by Facebook): Fill-Mask, 4.78M, 11 downloads.
- openai/clip-vit-large-patch14** (by OpenAI): 9.7M, 106 downloads.
- roberta-base** (by Facebook): Fill-Mask, 3.45M, 6 downloads.
- gpt2** (by Facebook): Text Generation, 3.34M, 24 downloads.

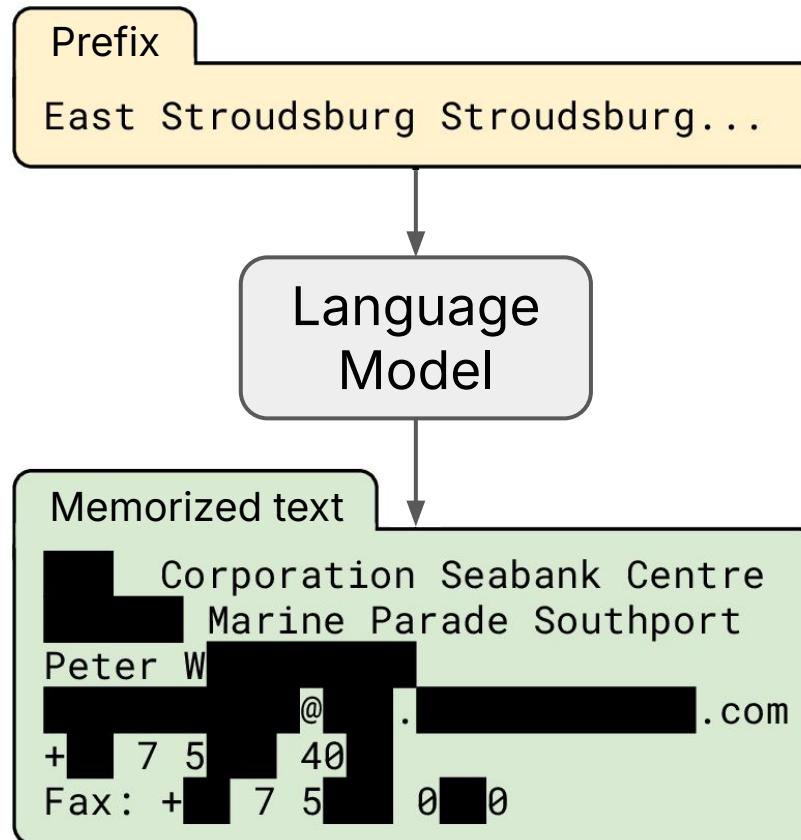
... and the models themselves are rarely updated

The screenshot shows the Hugging Face Model Hub interface. At the top, there's a search bar with the placeholder "Search models, datasets, users..." and a navigation menu icon. Below the header, a banner displays "Models 33,490". There are buttons for "Search Models", "Add filters", and "Sort: Most Downloads". The main content area lists seven pre-trained models, each in a card with a year and model name:

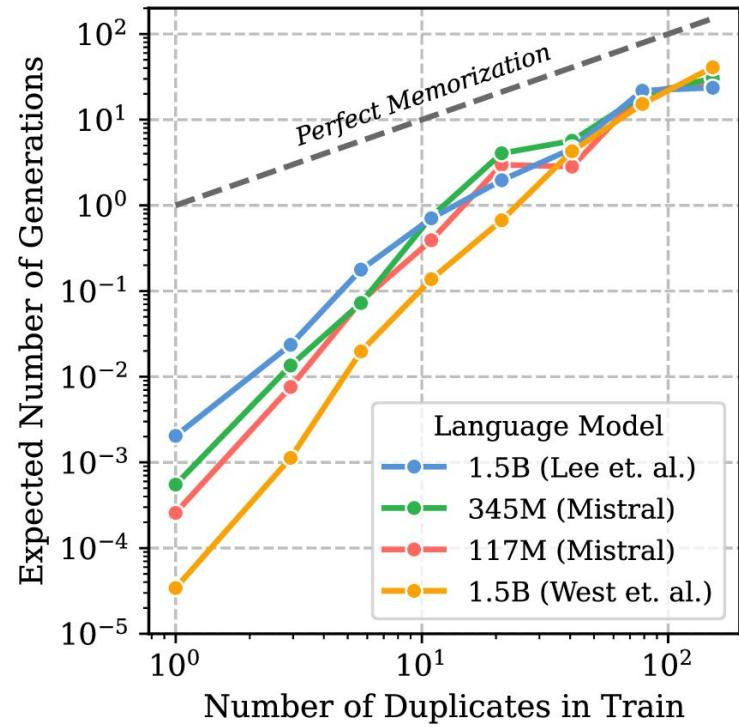
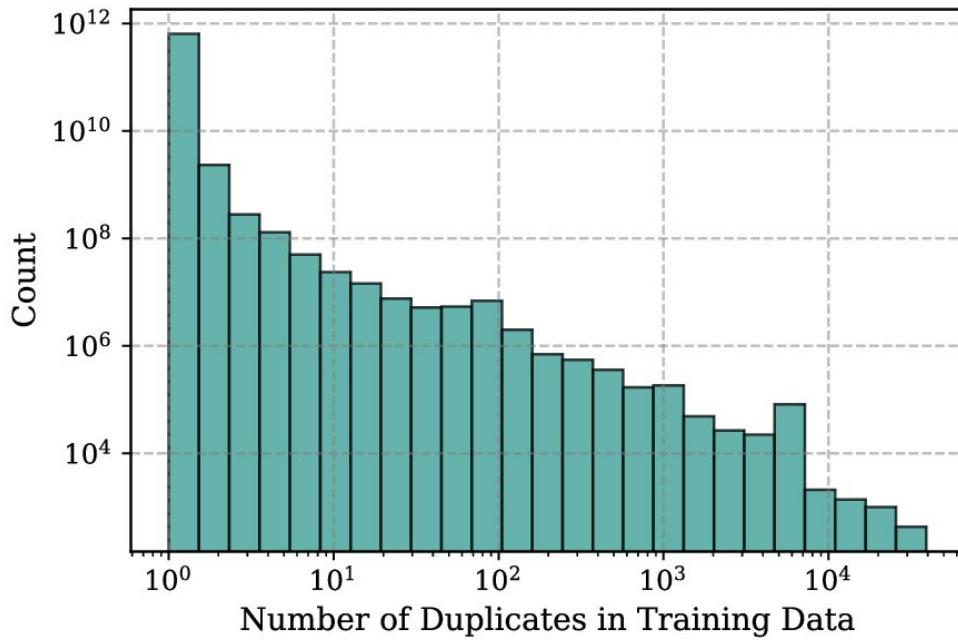
- 2018** bert-base-uncased
- 2019** roberta-large
- 2019** distilbert-base-uncased
- 2019** xlm-roberta-base
- 2021** openai/clip-vit-large-patch14
- 2019** roberta-base
- 2019** gpt2

Each card includes a small icon, the model name, its task (e.g., "Fill-Mask"), size ("30M", "13.1M", etc.), and the number of downloads ("54", "20", etc.).

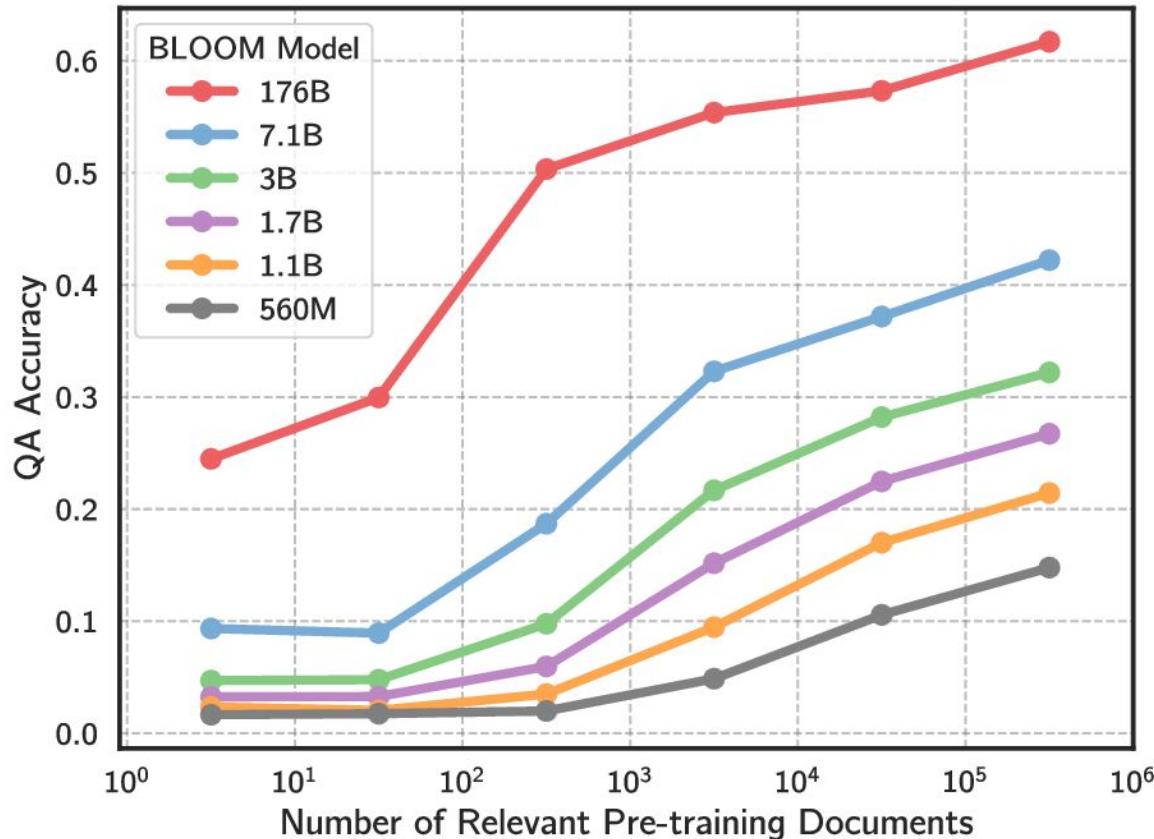
Models can exhibit issues, like memorized training data



Issues with a model can be caused by issues with a dataset



Pre-training datasets can also fail to address downstream needs



From "Large Language Models Struggle to Learn Long-Tail Knowledge" by Kandpal et al.

Pre-trained models are often used as the basis for derivative models

Models 4,184

t5

Add filters

Sort: Most Downloads

Michau/t5-base-en-generate-headline

↓ 1.64M ⚡ 27

prithivida/parrot_paraphraser_on_T5

↓ 1.24M ⚡ 46

pszemraj/long-t5-tglobal-base-1638...

↓ 1.47M ⚡ 38

mrm8488/t5-base-finetuned-question...

↓ 438k ⚡ 58

snzspeaks/t5-one-line-summary

↓ 1.28M ⚡ 23

mrm8488/t5-base-finetuned-common_g...

↓ 302k ⚡ 19



SentenceT5

Imagen

Muse

mT5

T5

T5.1.1

PaLI

mT0

ByT5

UnifiedQA

Tk-Instruct

T5+LM

Flan-T5

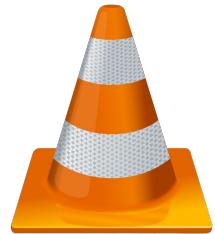
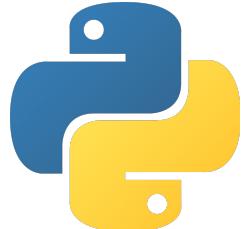
T0

→ Additional training

→ New model

→ Reuse part of the model

MACAW

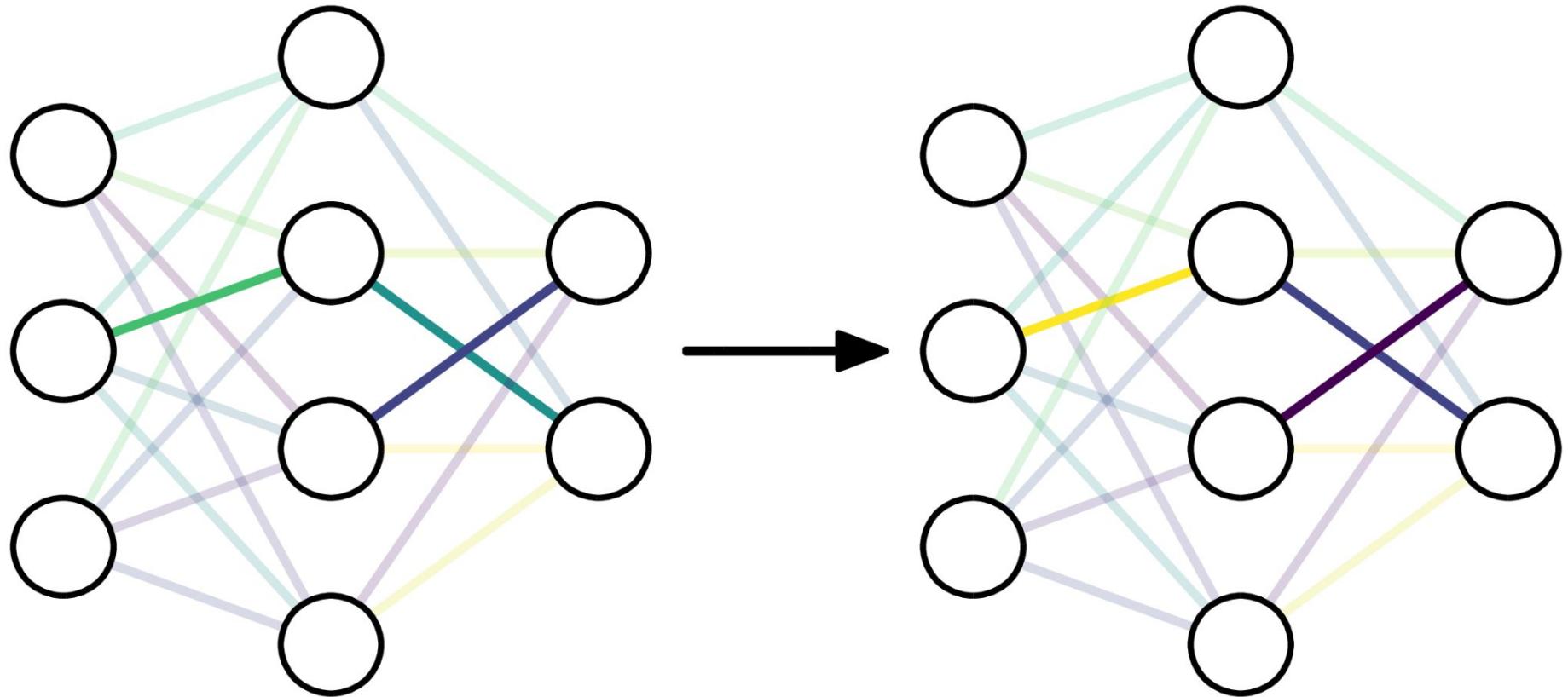


How can we enable collaborative and continual development of machine learning models?

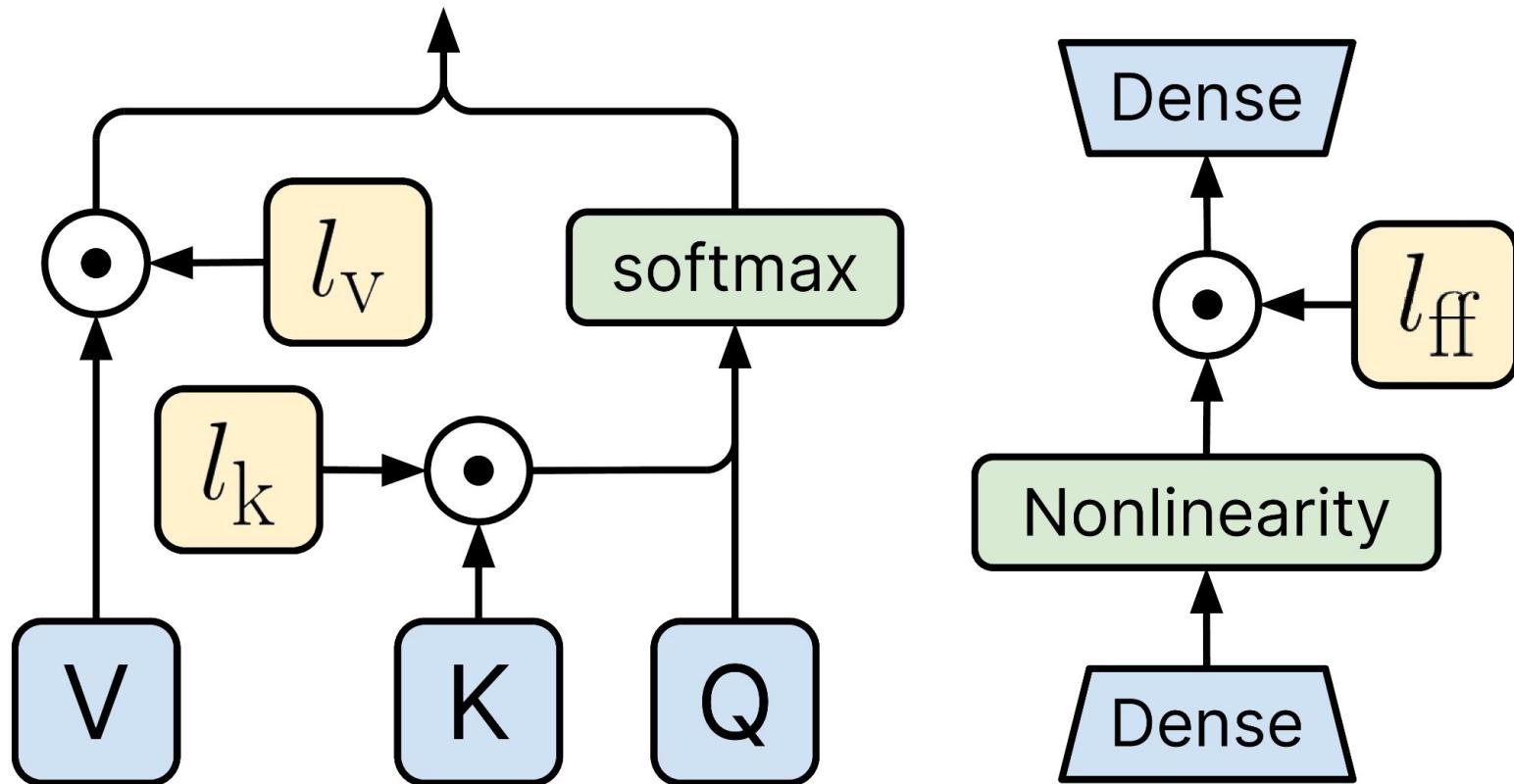
How can we enable collaborative and continual development of machine learning models?

Contributors need to be able to cheaply communicate **patches** to a model.

Updating a subset of parameters reduces communication costs



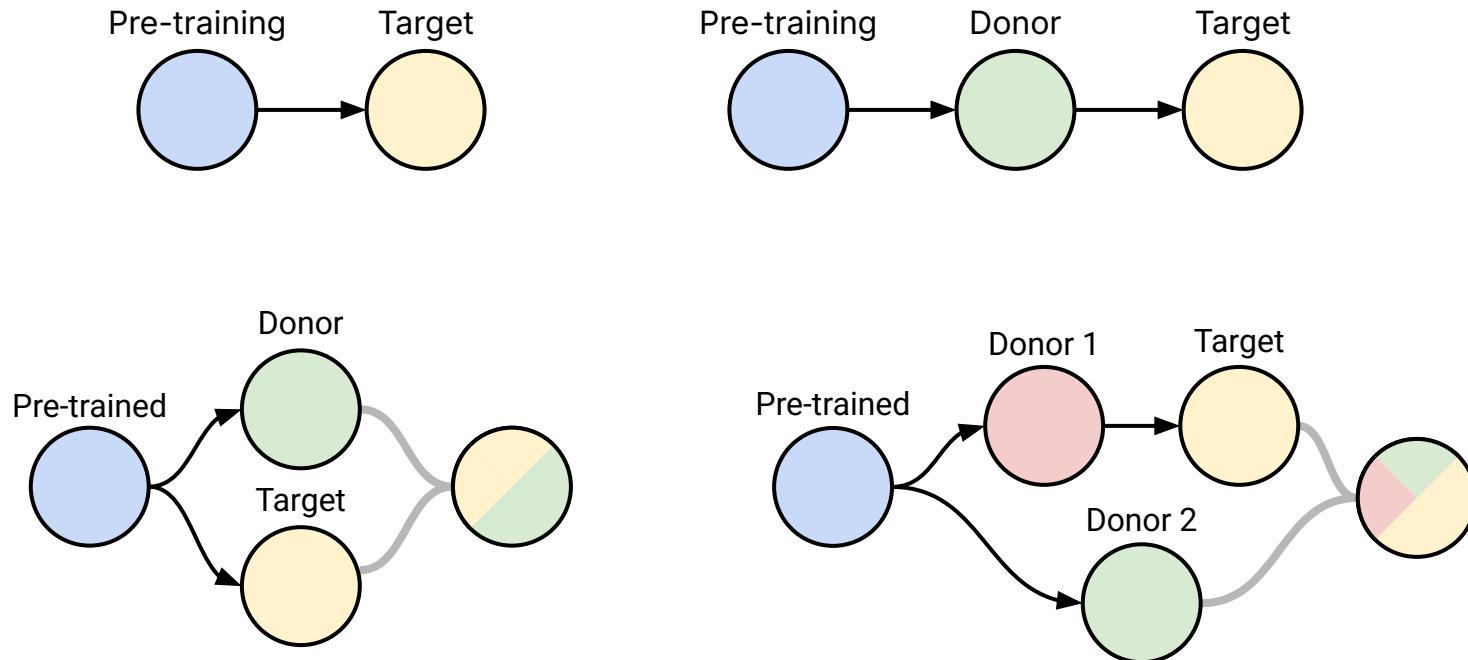
Updating models by rescaling activations



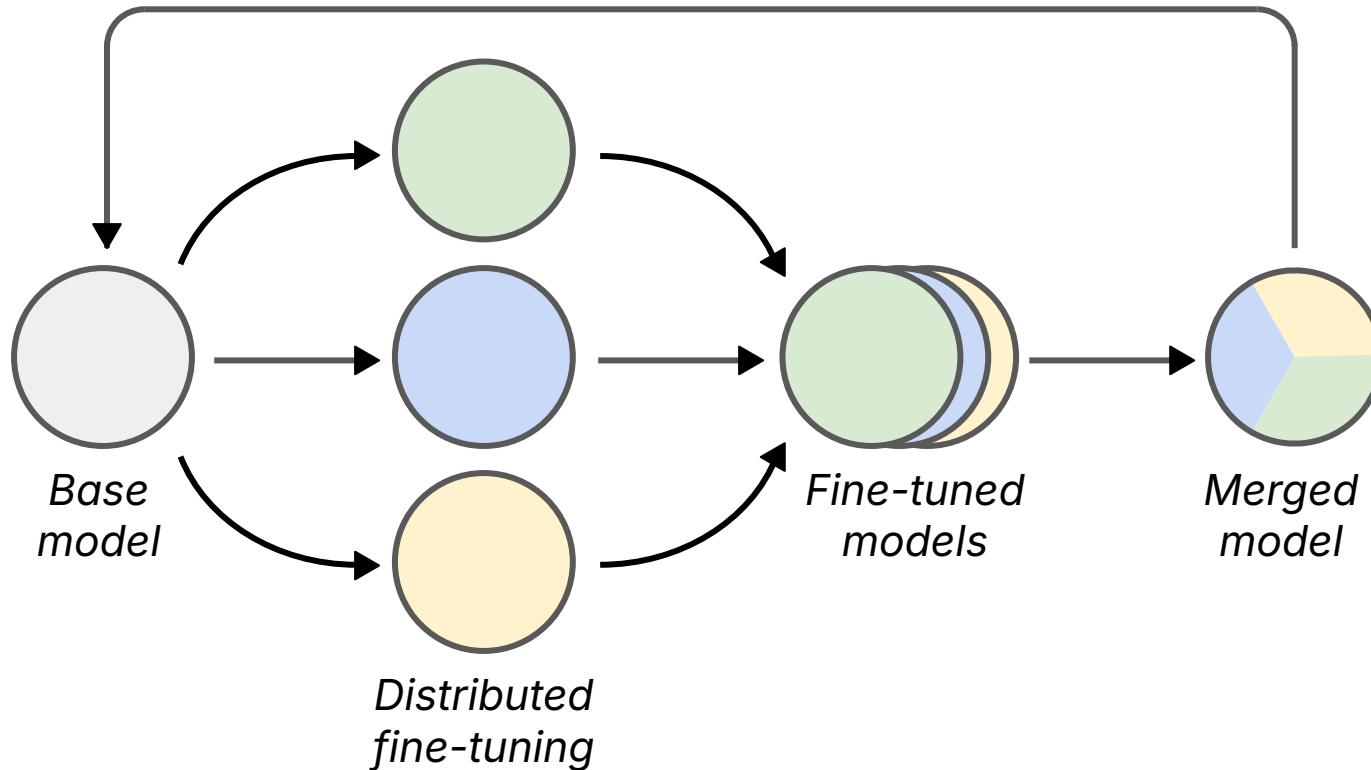
How can we enable collaborative and continual development of machine learning models?

Maintainers need to be able to merge updates from different contributors.

Model merging enables new paths for transferring capabilities



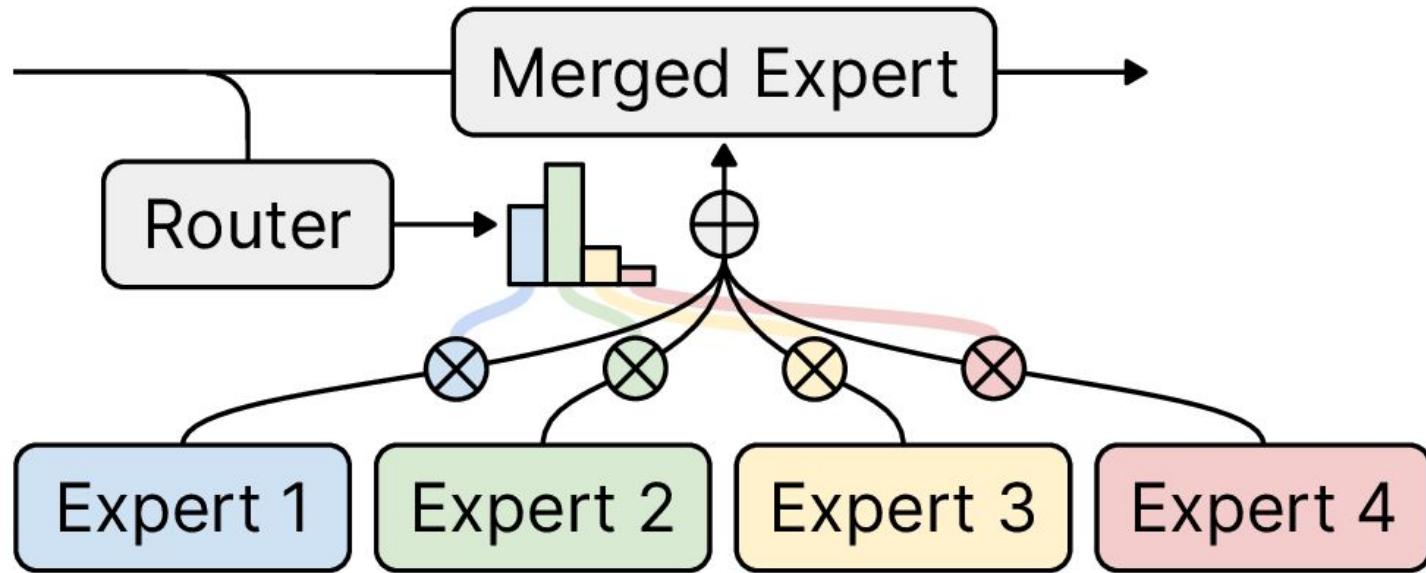
Merging fine-tuned models for better pre-trained models



How can we enable collaborative and continual development of machine learning models?

We need to be able to combine **modular** components to enable new capabilities.

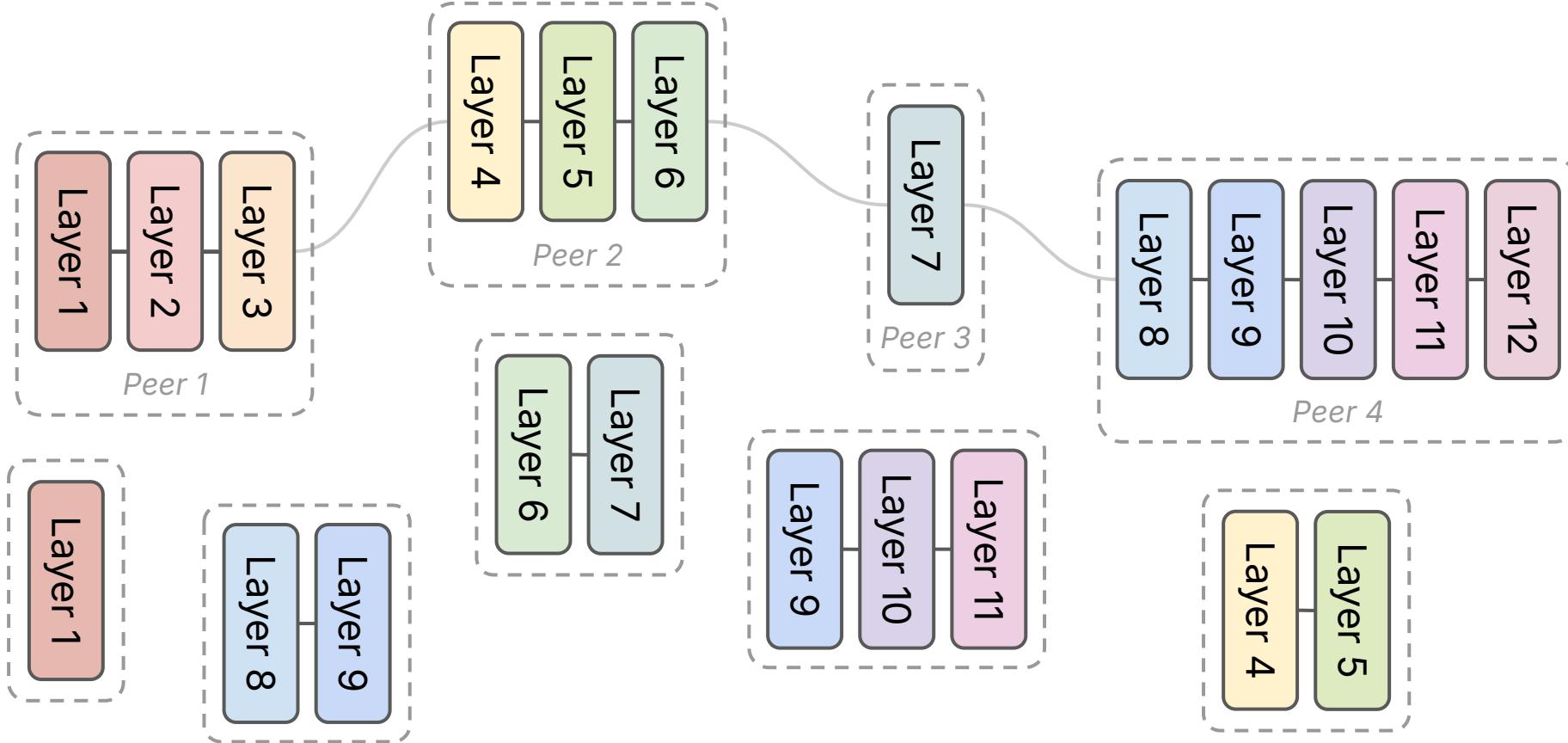
Modularity by merging experts with SMEAR



How can we enable collaborative and continual development of machine learning models?

Users who lack resources need to be able to **train and run** large models.

PETALS enables distributed inference of large models over the internet

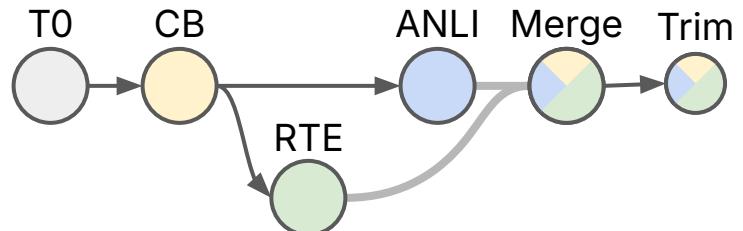


How can we enable collaborative and continual development of machine learning models?

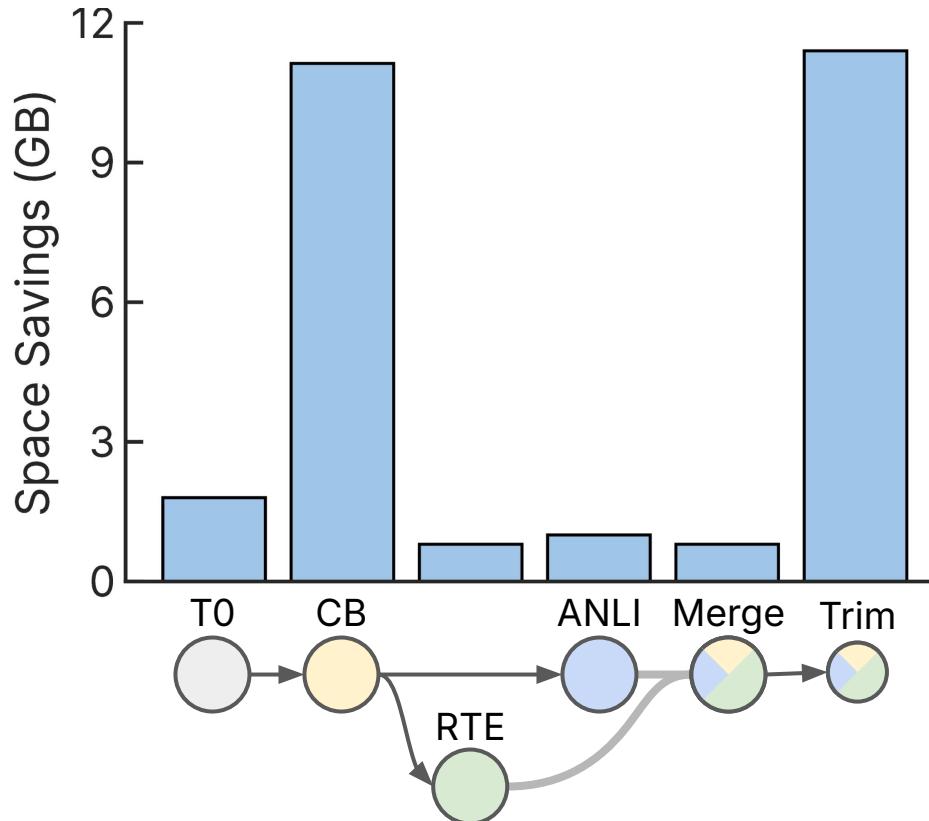
We need a system for **version control** of model parameters.

git-theta tracks, merges, and updates models using the git workflow

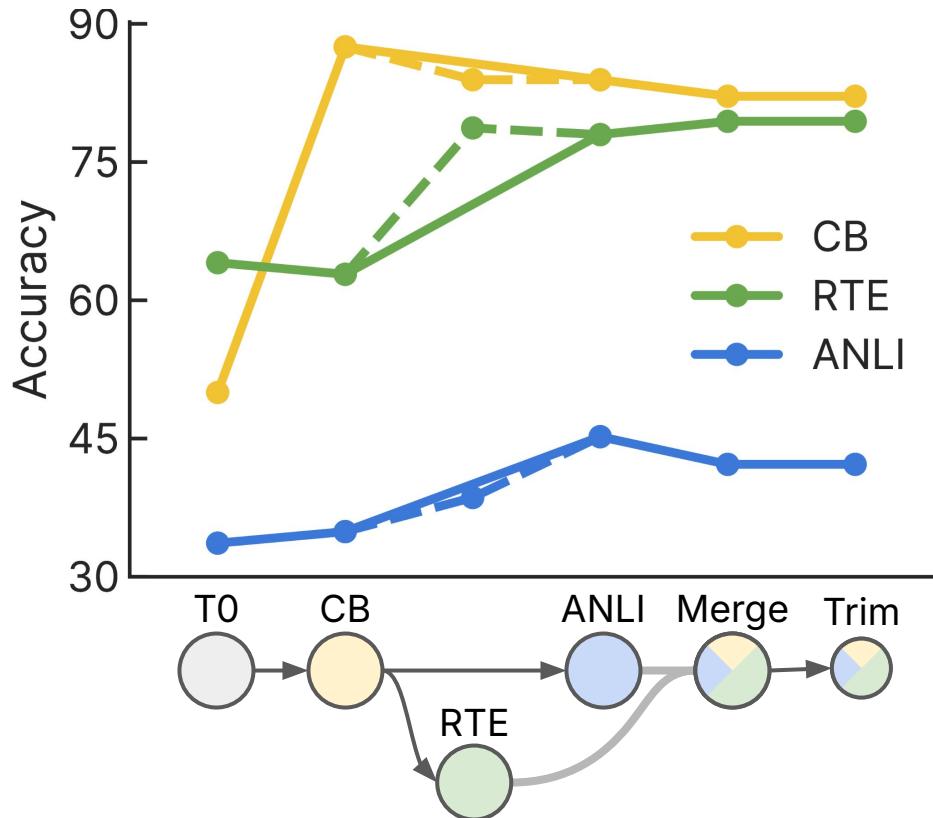
```
$ git-theta track model.pt
$ git commit -am "Add initial model"
$ python finetune.py --dataset="cb" --method="lowrank"
$ git commit -am "Fine-tune on CB dataset with LoRA"
$ git checkout -b rte
$ python finetune.py --dataset="rte" --method="dense"
$ git commit -am "Fine-tune on RTE dataset"
$ git checkout main
$ python finetune.py --dataset="anli" --method="dense"
$ git commit -am "Fine-tune on ANLI dataset"
$ git merge rte
Fixing Merge Conflicts in model.pt
Actions:
  avg) average: Average parameter values.
  tt) take_their: Use their change to the parameter.
  tu) take_us: Use our change to the parameter.
  q) quit
θ avg
$ git commit -am "Merge RTE and ANLI models"
$ python trim_unused_embeddings.py
$ git commit -am "Remove embeddings for unused tokens"
```



Communication-efficient updates result in significant space savings



git-theta allows for continuous and collaborative model development



[Building Machine Learning Models Like Open Source Software](#), Communications of the ACM

Colin Raffel

[Extracting Training Data from Large Language Models](#), USENIX Security 2021

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, Alina Oprea, & Colin Raffel

[Deduplicating Training Data Mitigates Privacy Risks in Language Models](#), ICML 2022

Nikhil Kandpal, Eric Wallace, & Colin Raffel

[Large Language Models Struggle to Learn Long-Tail Knowledge](#), ICML 2023

Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, & Colin Raffel

[Training Neural Networks with Fixed Sparse Masks](#), NeurIPS 2021

Yi-Lin Sung*, Varun Nair*, & Colin Raffel

[Few-Shot Parameter-Efficient Fine-Tuning is Better and Cheaper than In-Context Learning](#), NeurIPS 2022

Haokun Liu*, Derek Tam*, Mohammed Muqeeth*, Jay Mohta, Tenghao Huang, Mohit Bansal, & Colin Raffel

[Merging Models with Fisher-Weighted Averaging](#), NeurIPS 2022

Michael Matena & Colin Raffel

[CoID Fusion: Collaborative Descent for Distributed Multitask Finetuning](#), in submission

Shachar Don-Yehiya, Elad Venezian, Colin Raffel, Noam Slonim, Yoav Katz, & Leshem Choshen

[Soft Merging of Experts with Adaptive Routing](#), in submission

Mohammed Muqeeth, Haokun Liu, & Colin Raffel

[Git-Theta: A Git Extension for Collaborative Development of Machine Learning Models](#), ICML 2023

Nikhil Kandpal*, Brian Lester*, Mohammed Muqeeth, Anisha Mascarenhas, Monty Evans, Vishal Baskaran, Tenghao Huang, Haokun Liu, & Colin Raffel

[Petals: Collaborative Inference and Fine-tuning of Large Models](#), ACL 2023

Alexander Borzunov, Dmitry Baranchuk, Tim Dettmers, Max Ryabinin, Younes Belkada, Artem Chumachenko, Pavel Samygin, & Colin Raffel