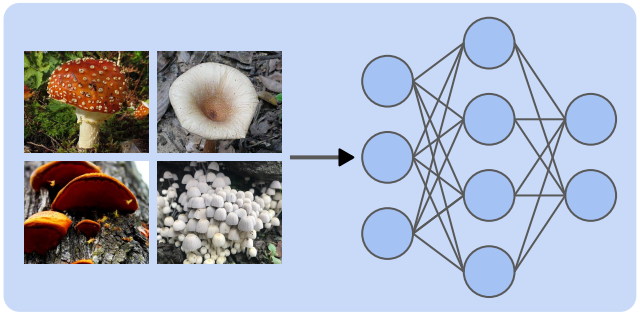
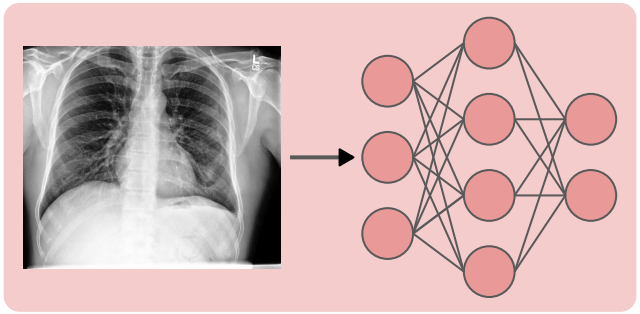
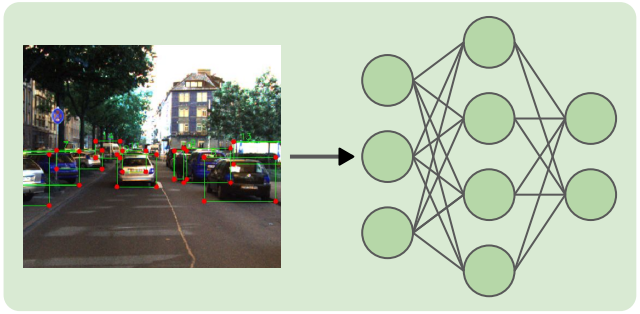
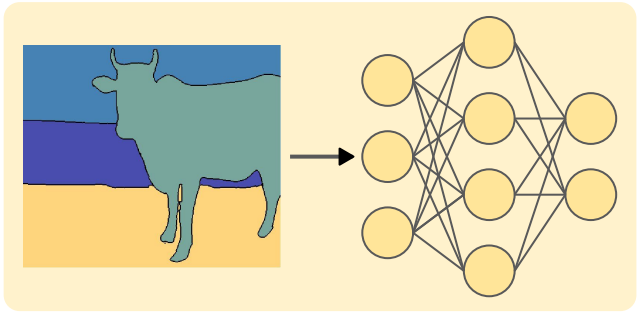
Building Machine Learning Models like Open-Source Software

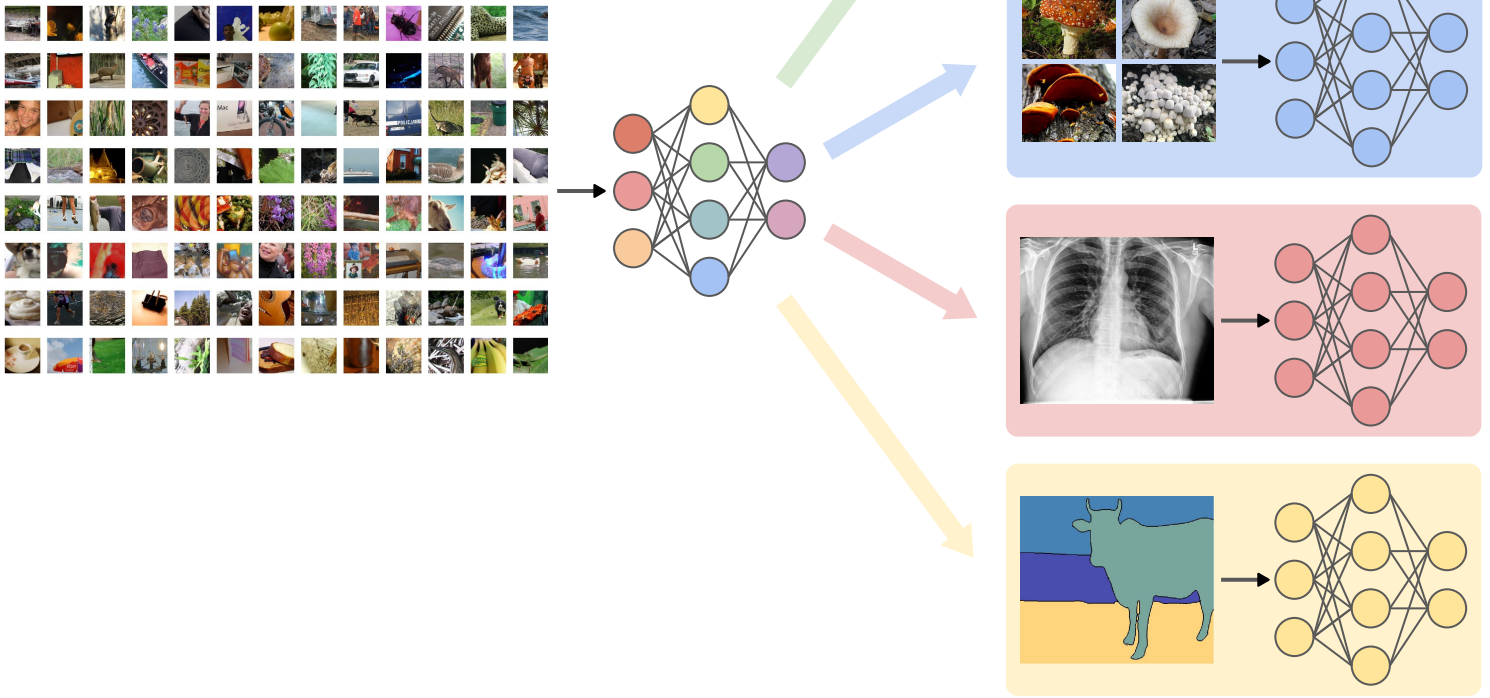
Colin Raffel

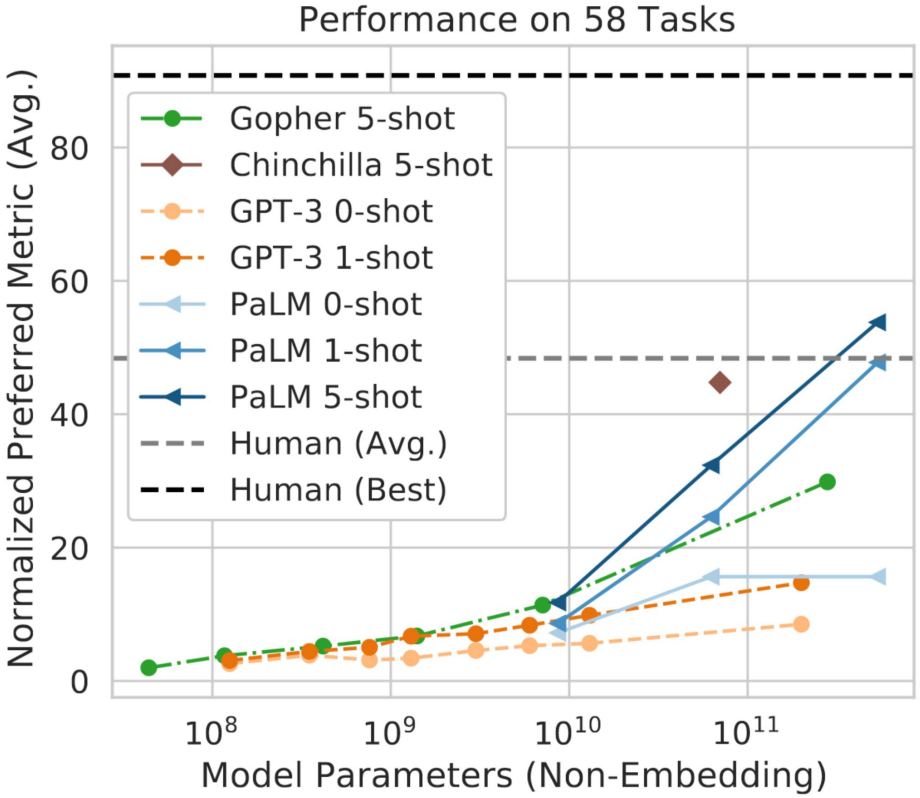


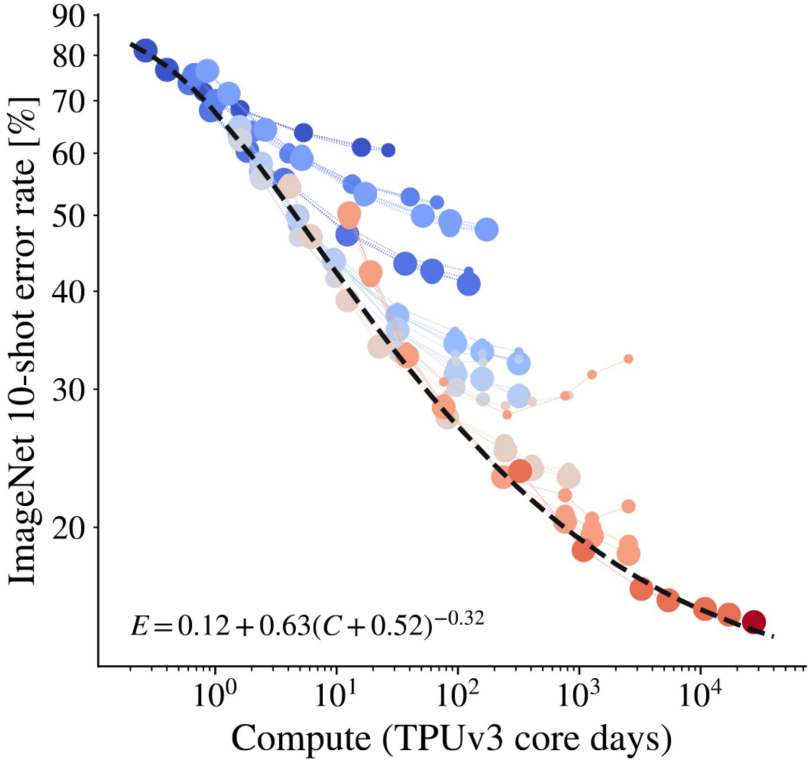




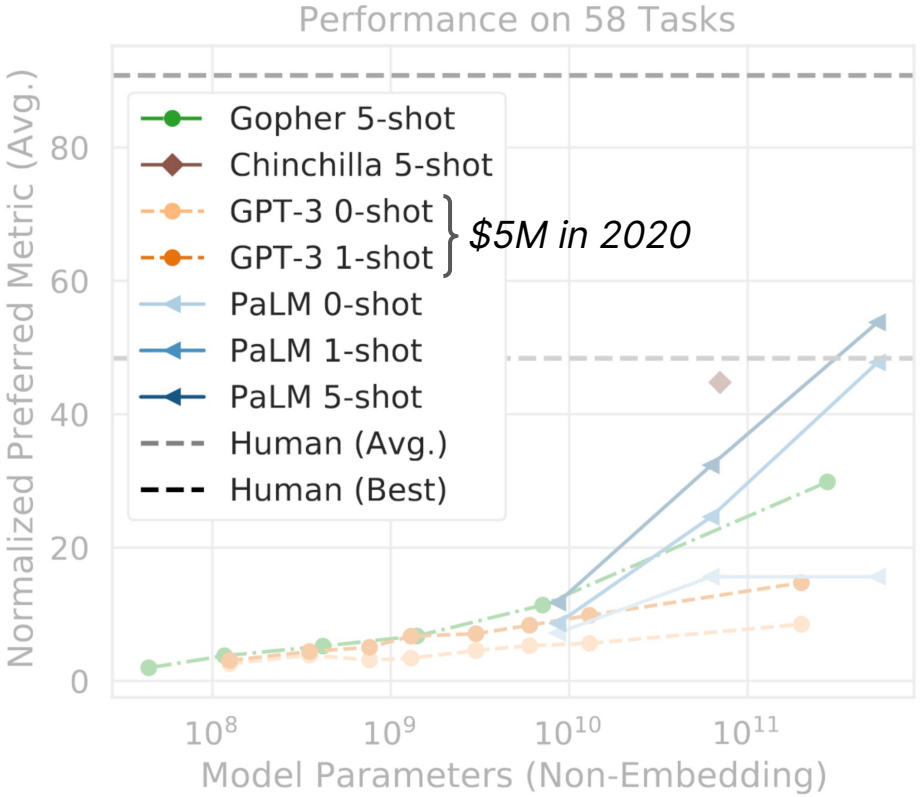


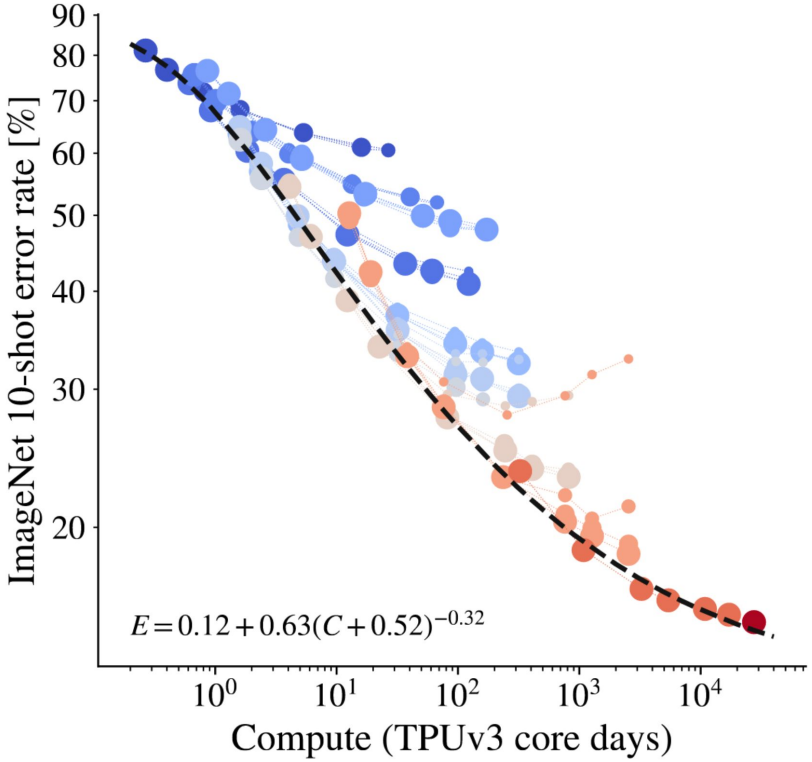




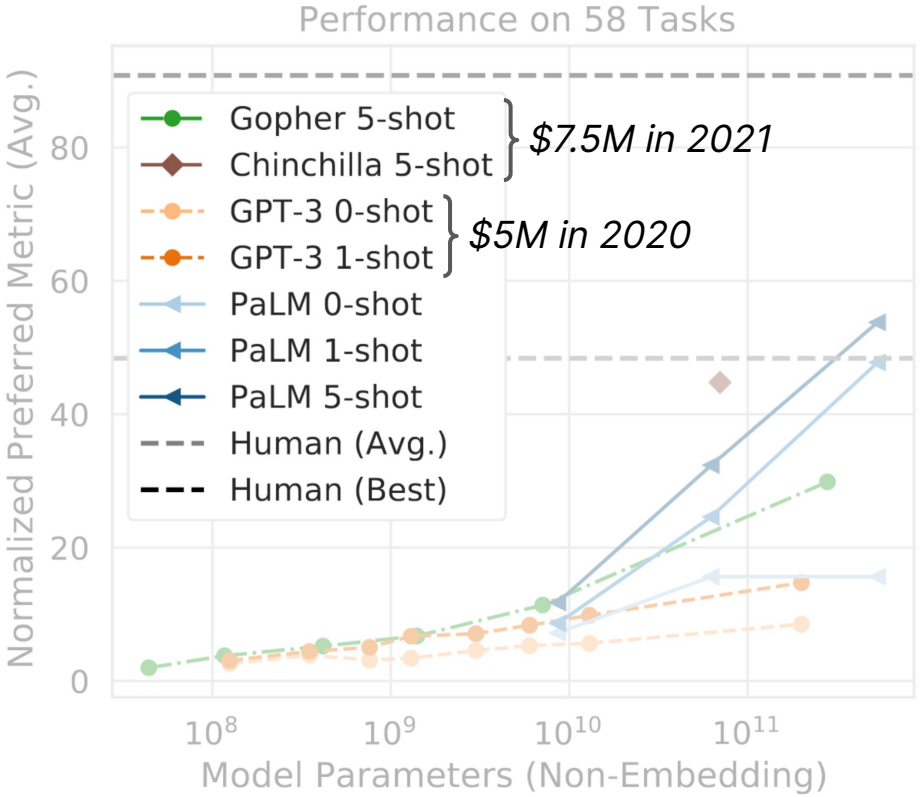


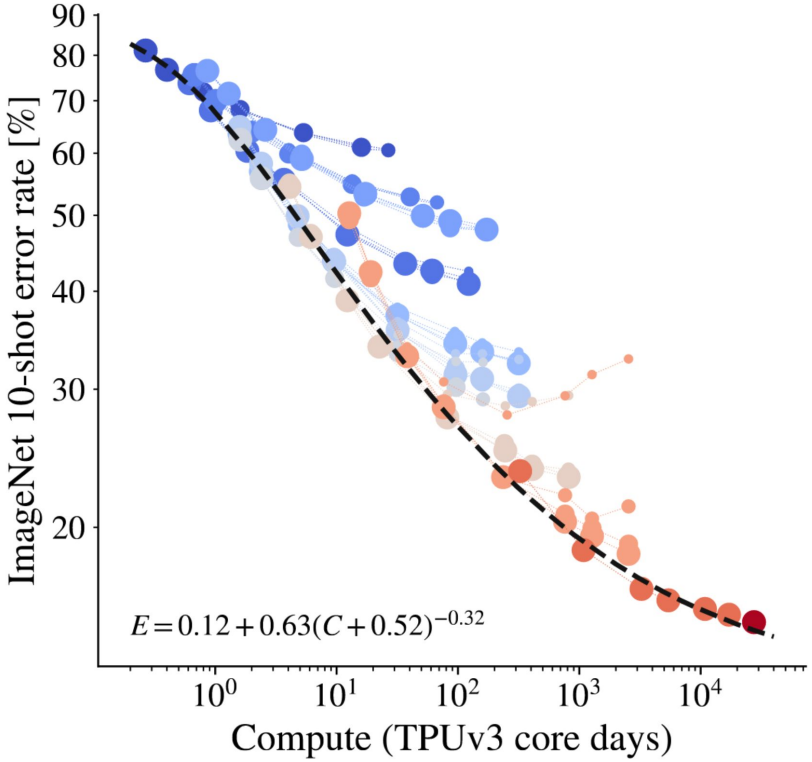
From "PaLM: Scaling Language Modeling with Pathways" by Chowdhery et al. and "Scaling Vision Transformers" by Zhai et al.



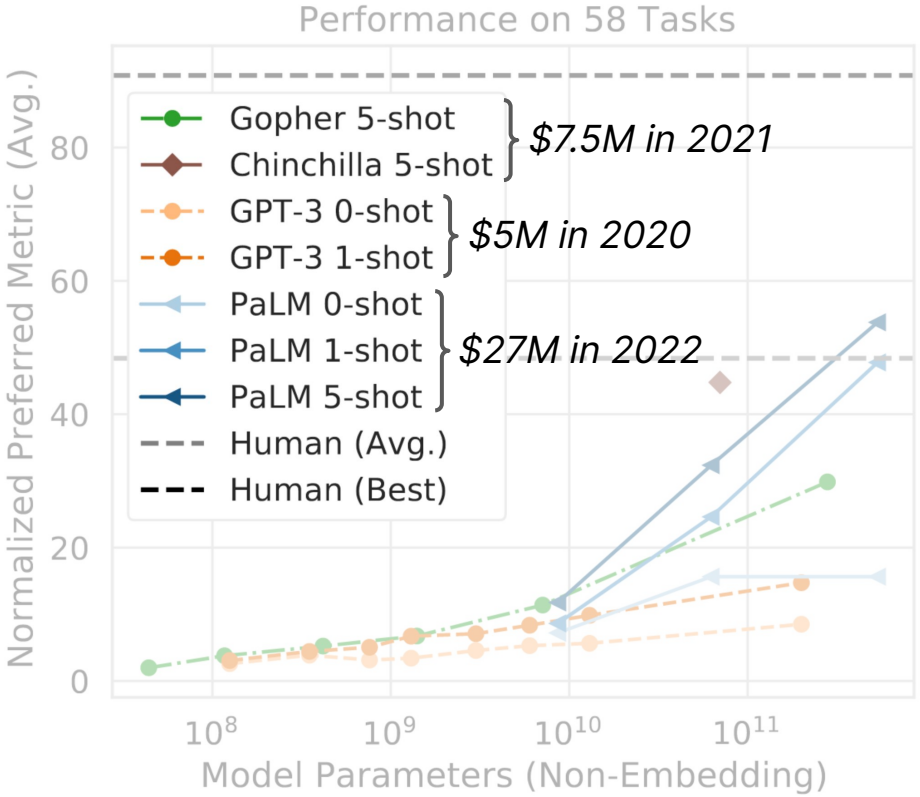


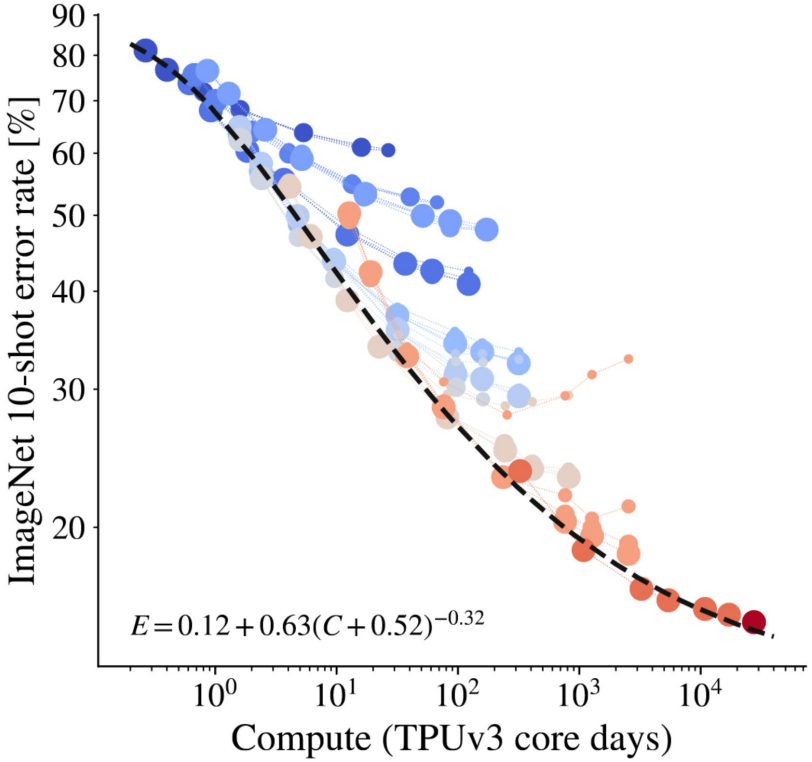
From "PaLM: Scaling Language Modeling with Pathways" by Chowdhery et al. and "Scaling Vision Transformers" by Zhai et al.



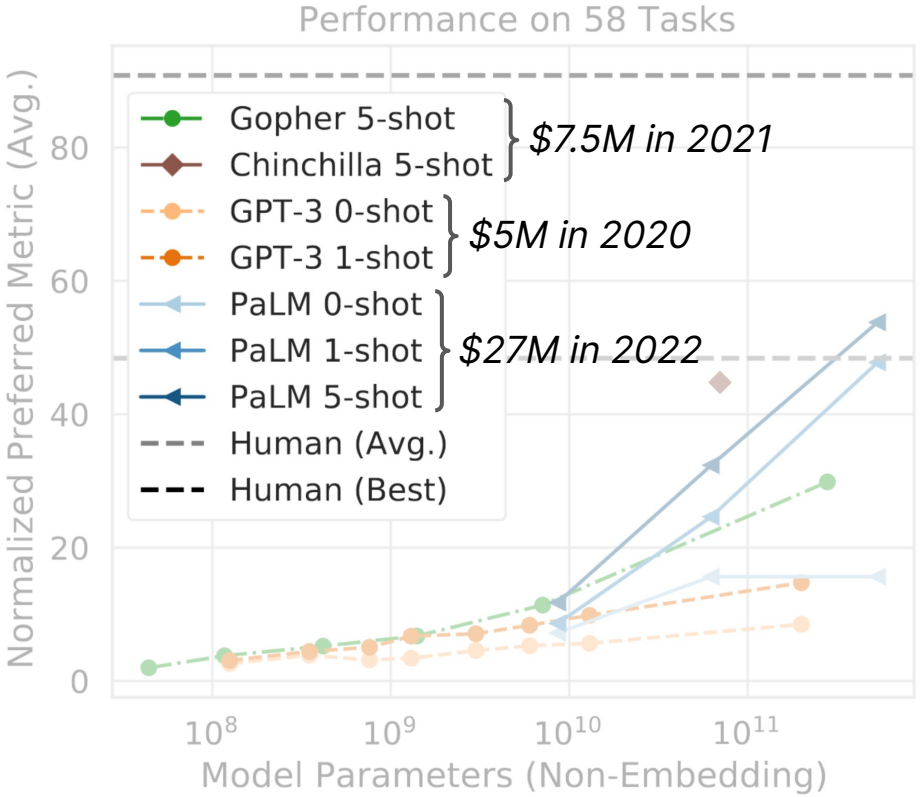


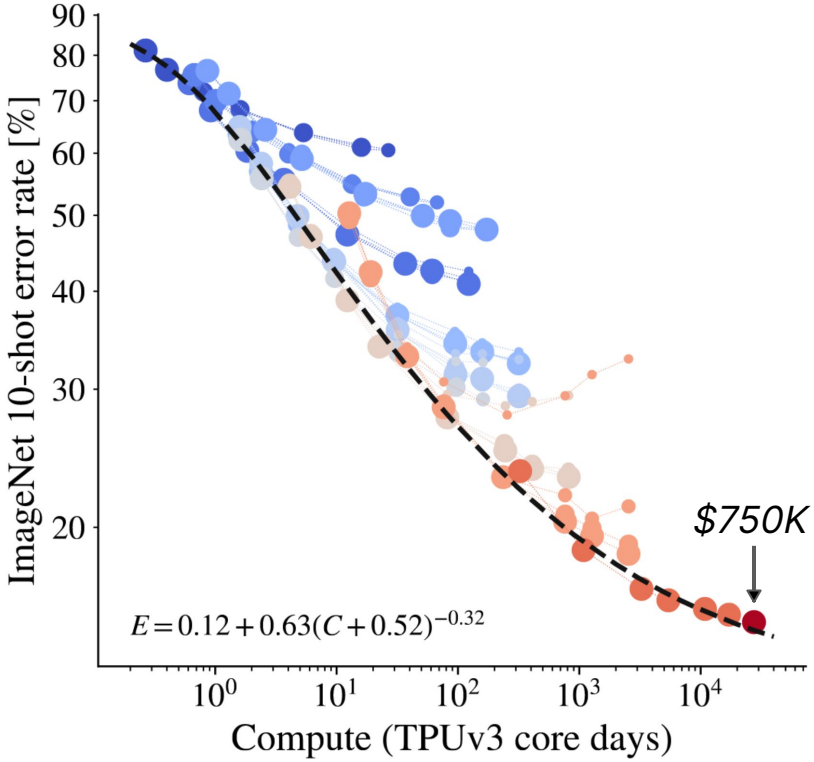
From "PaLM: Scaling Language Modeling with Pathways" by Chowdhery et al. and "Scaling Vision Transformers" by Zhai et al.



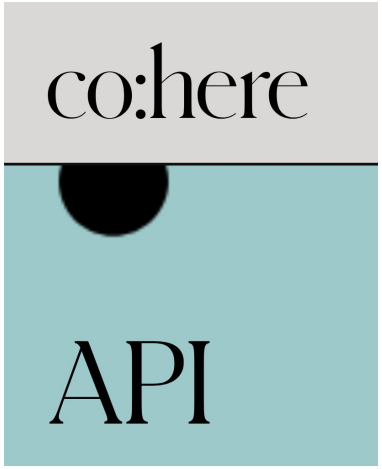


From "PaLM: Scaling Language Modeling with Pathways" by Chowdhery et al. and "Scaling Vision Transformers" by Zhai et al.

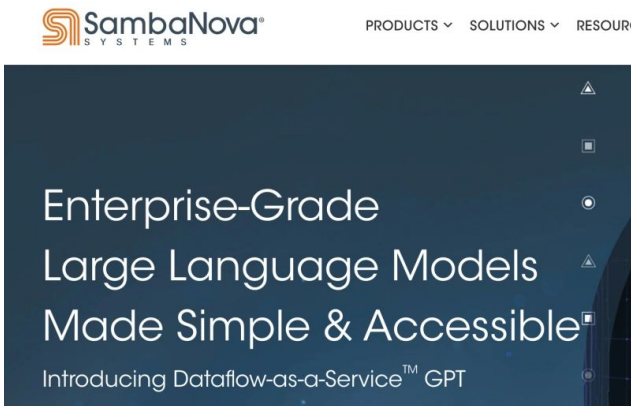




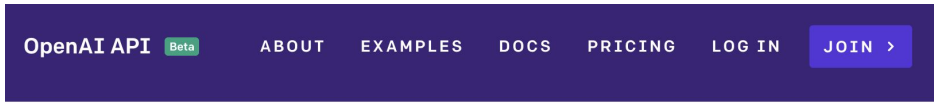
From "PaLM: Scaling Language Modeling with Pathways" by Chowdhery et al. and "Scaling Vision Transformers" by Zhai et al.

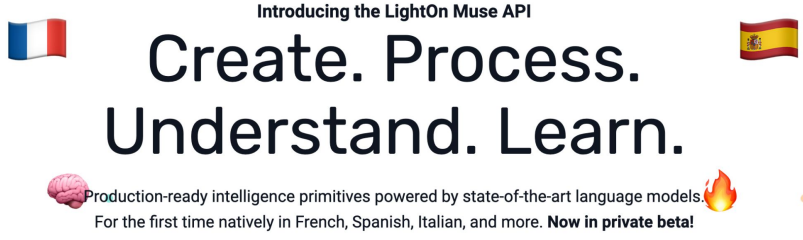

co:here
API

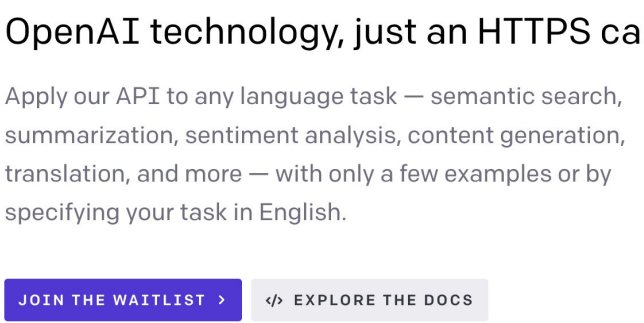

AI21 studio
Custom language models built for scale
Build sophisticated language applications on top of AI21's language models

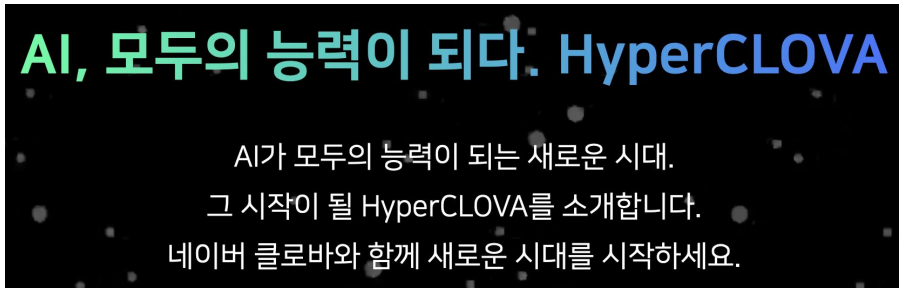

SambaNova SYSTEMS
PRODUCTS
SOLUTIONS
RESOUR
Enterprise-Grade
Large Language Models
Made Simple & Accessible
Introducing Dataflow-as-a-Service™ GPT

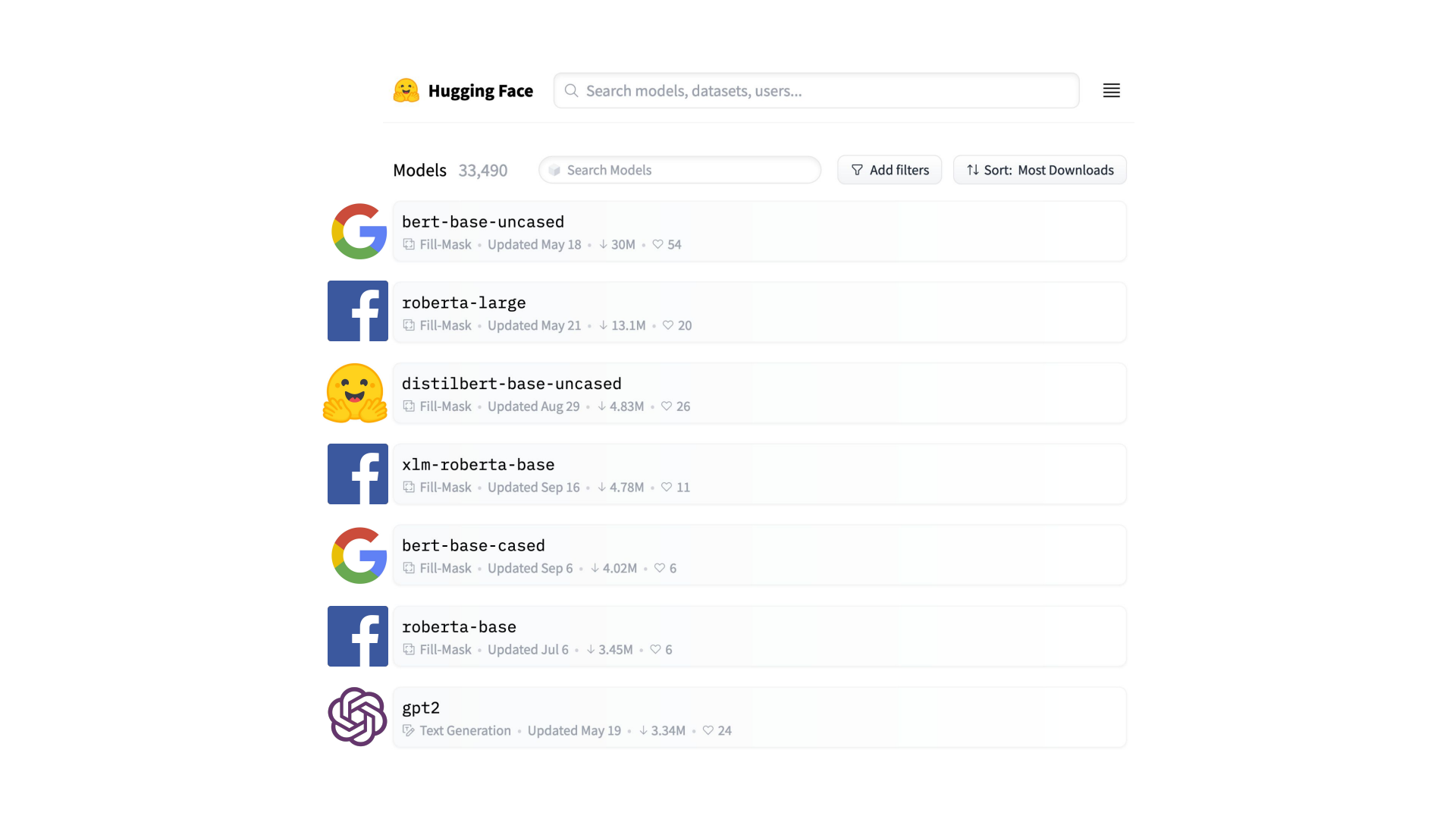

Microsoft Megatron-Turing NLG 530B
The World's Largest and Most Powerful Generative Language Model

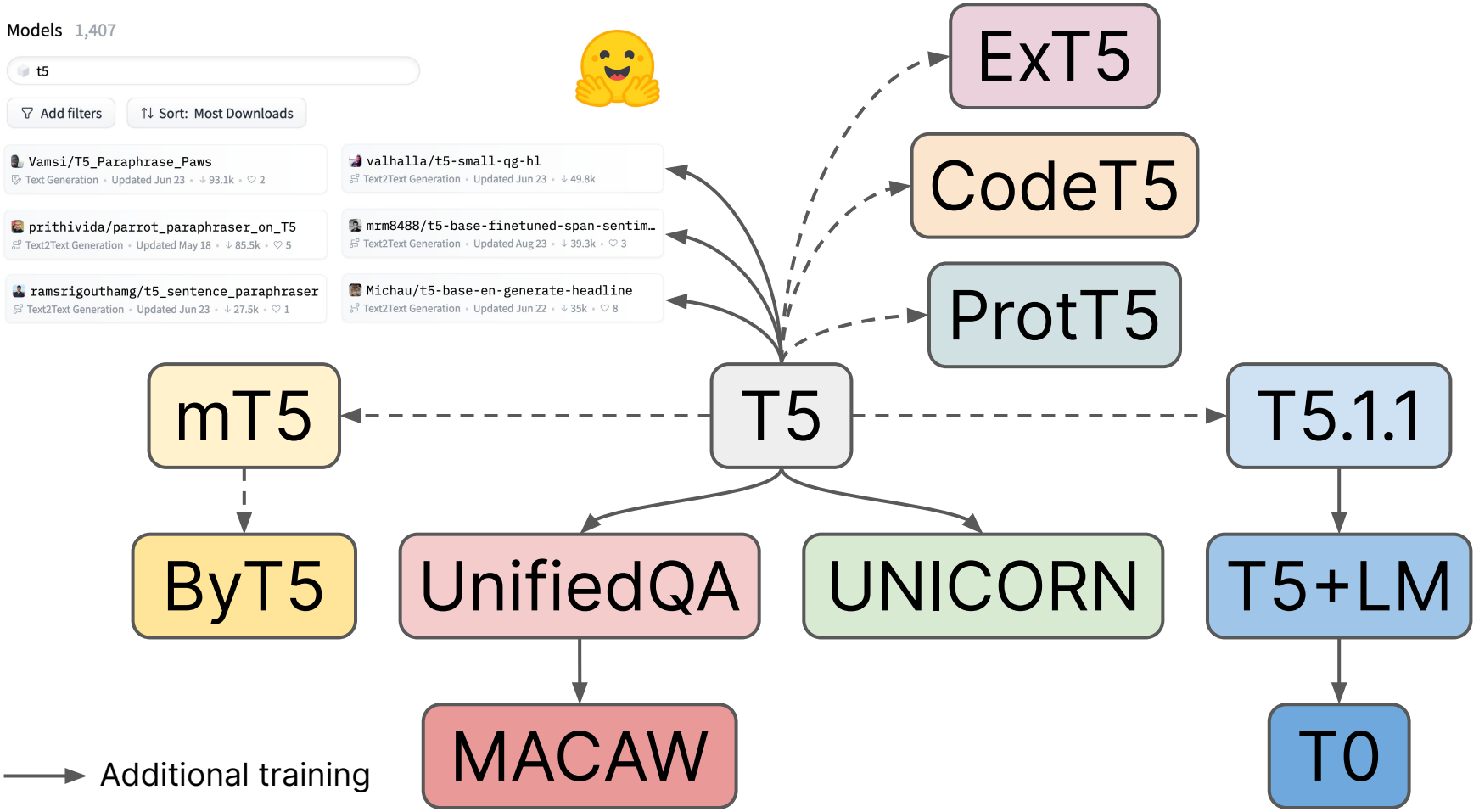

OpenAI API Beta
ABOUT
EXAMPLES
DOCS
PRICING
LOG IN
JOIN ›


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!

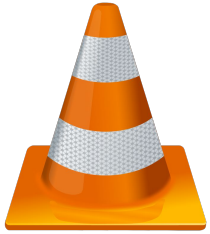
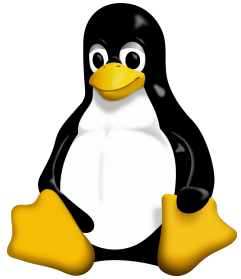

OpenAI technology, just an HTTPS call away
Apply our API to any language task — semantic search, summarization, sentiment analysis, content generation, translation, and more — with only a few examples or by specifying your task in English.
JOIN THE WAITLIST ›
</> EXPLORE THE DOCS


AI, 모두의 능력이 되다. HyperCLOVA
AI가 모두의 능력이 되는 새로운 시대.
그 시작이 될 HyperCLOVA를 소개합니다.
네이버 클로바와 함께 새로운 시대를 시작하세요.


Hugging Face
Search models, datasets, users...
Models 33,490
Search Models
Add filters
Sort: Most Downloads
bert-base-uncased
Fill-Mask • Updated May 18 • ↓ 30M • ♡ 54
roberta-large
Fill-Mask • Updated May 21 • ↓ 13.1M • ♡ 20
distilbert-base-uncased
Fill-Mask • Updated Aug 29 • ↓ 4.83M • ♡ 26
xlm-roberta-base
Fill-Mask • Updated Sep 16 • ↓ 4.78M • ♡ 11
bert-base-cased
Fill-Mask • Updated Sep 6 • ↓ 4.02M • ♡ 6
roberta-base
Fill-Mask • Updated Jul 6 • ↓ 3.45M • ♡ 6
gpt2
Text Generation • Updated May 19 • ↓ 3.34M • ♡ 24


Models 1,407
t5
Add filters
Sort: Most Downloads
Vamsi/T5_Paraphrase_Paws
Text Generation • Updated Jun 23 • ↓ 93.1k • ♡ 2
valhalla/t5-small-qg-hl
Text2Text Generation • Updated Jun 23 • ↓ 49.8k
prithivida/parrot_paraphraser_on_T5
Text2Text Generation • Updated May 18 • ↓ 85.5k • ♡ 5
mrm8488/t5-base-finetuned-span-sentim…
Text2Text Generation • Updated Aug 23 • ↓ 39.3k • ♡ 3
ramsrigouthamg/t5_sentence_paraphraser
Text2Text Generation • Updated Jun 23 • ↓ 27.5k • ♡ 1
Michau/t5-base-en-generate-headline
Text2Text Generation • Updated Jun 22 • ↓ 35k • ♡ 8
ExT5
CodeT5
ProtT5
mT5
T5
T5.1.1
ByT5
UnifiedQA
UNICORN
T5+LM
MACAW
T0
Additional training
New model










GCC

7ZIP

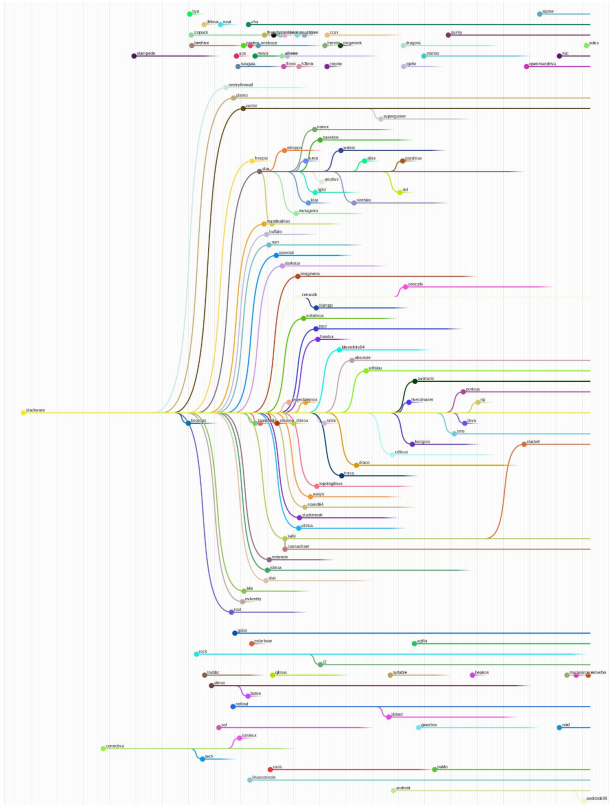
MySQL®

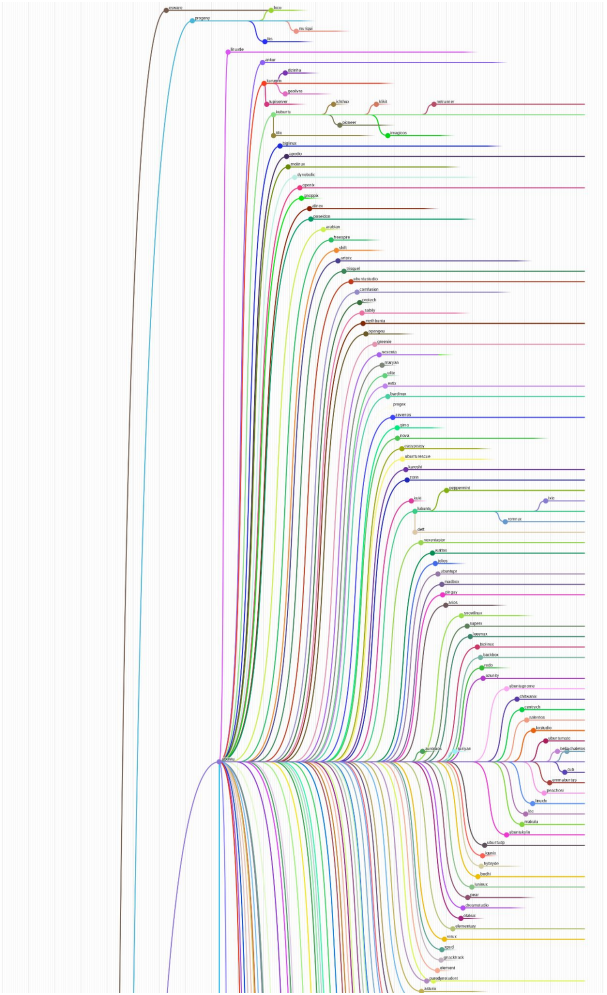
RAILS


Open
SSL

TOP
SECRET

R



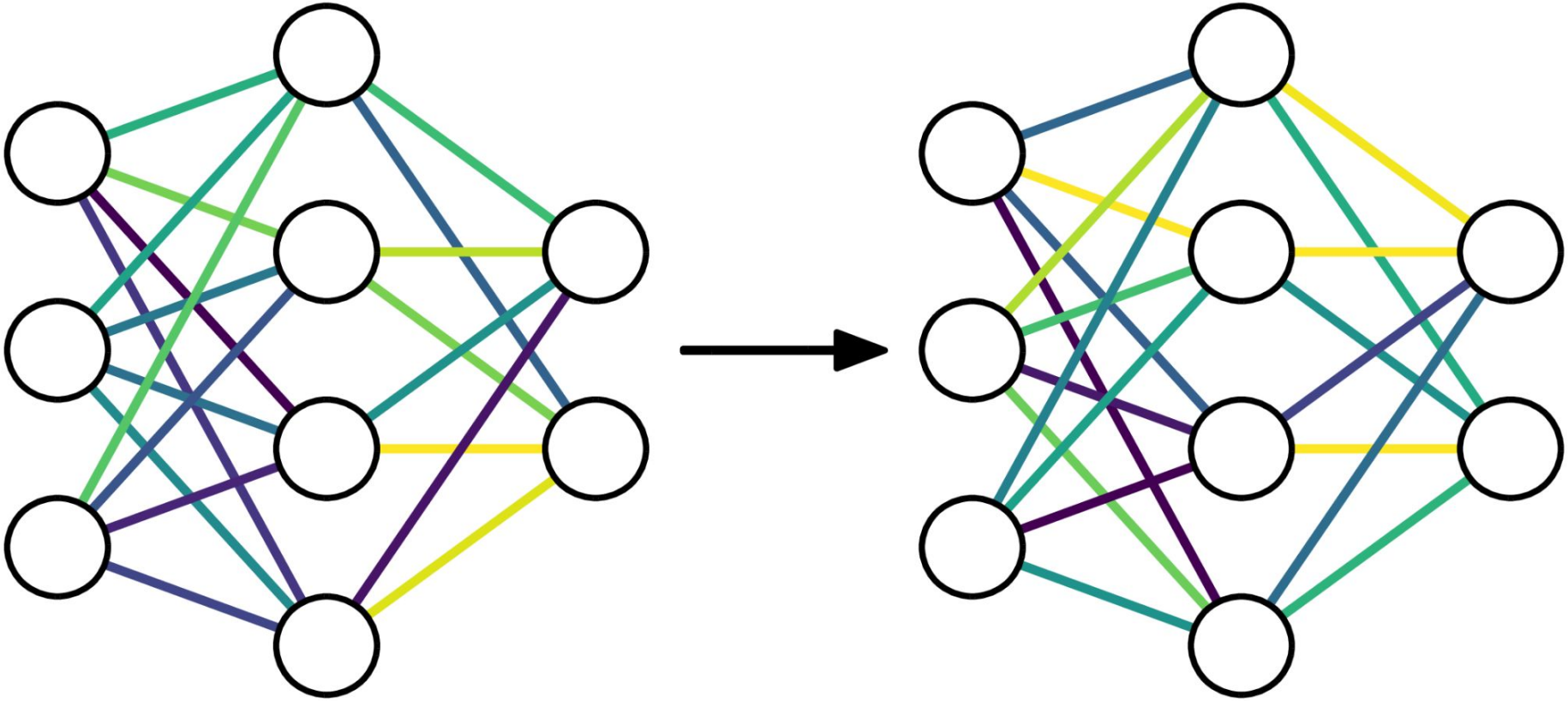


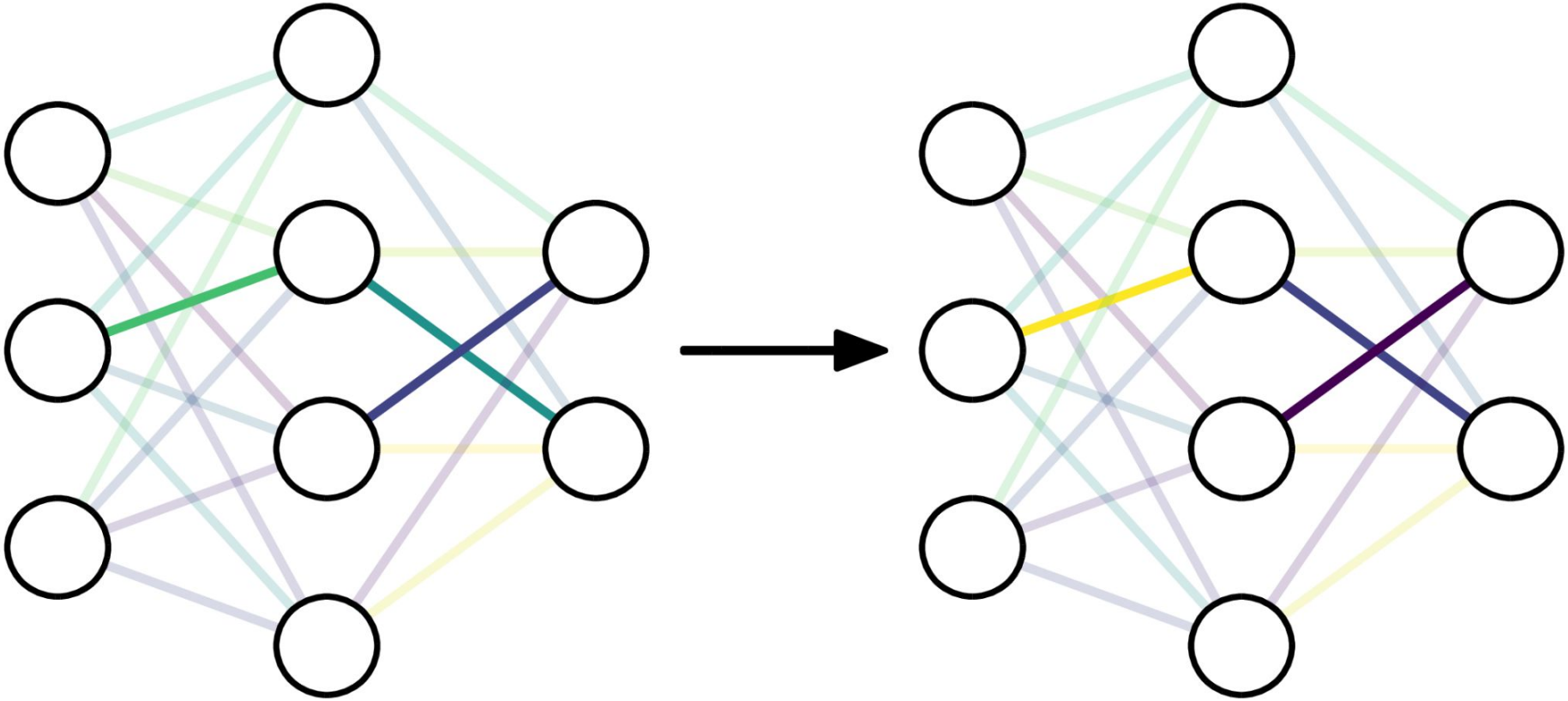
from https://distrowatch.com/

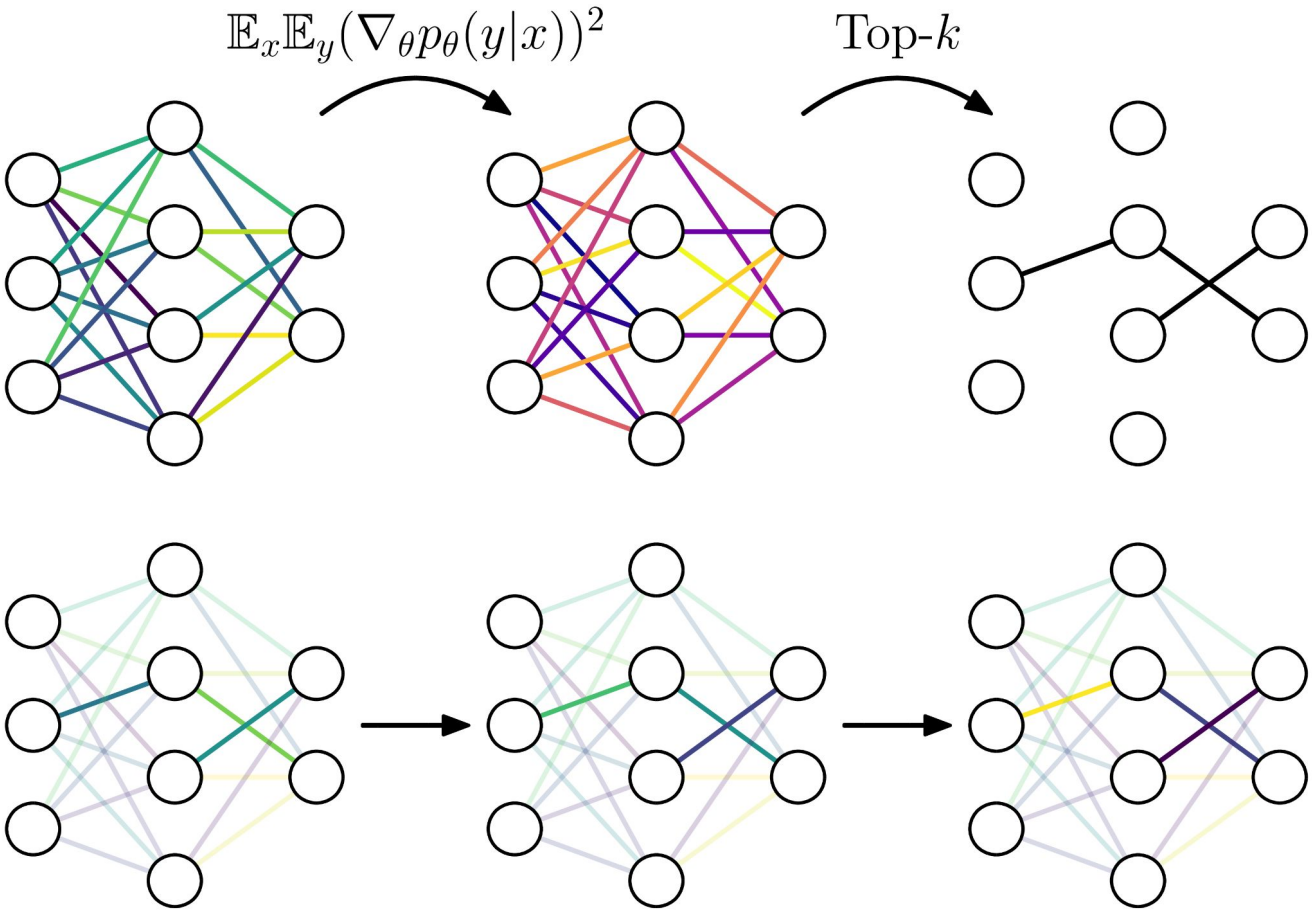
How can we enable collaborative and continual development of machine learning models?

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

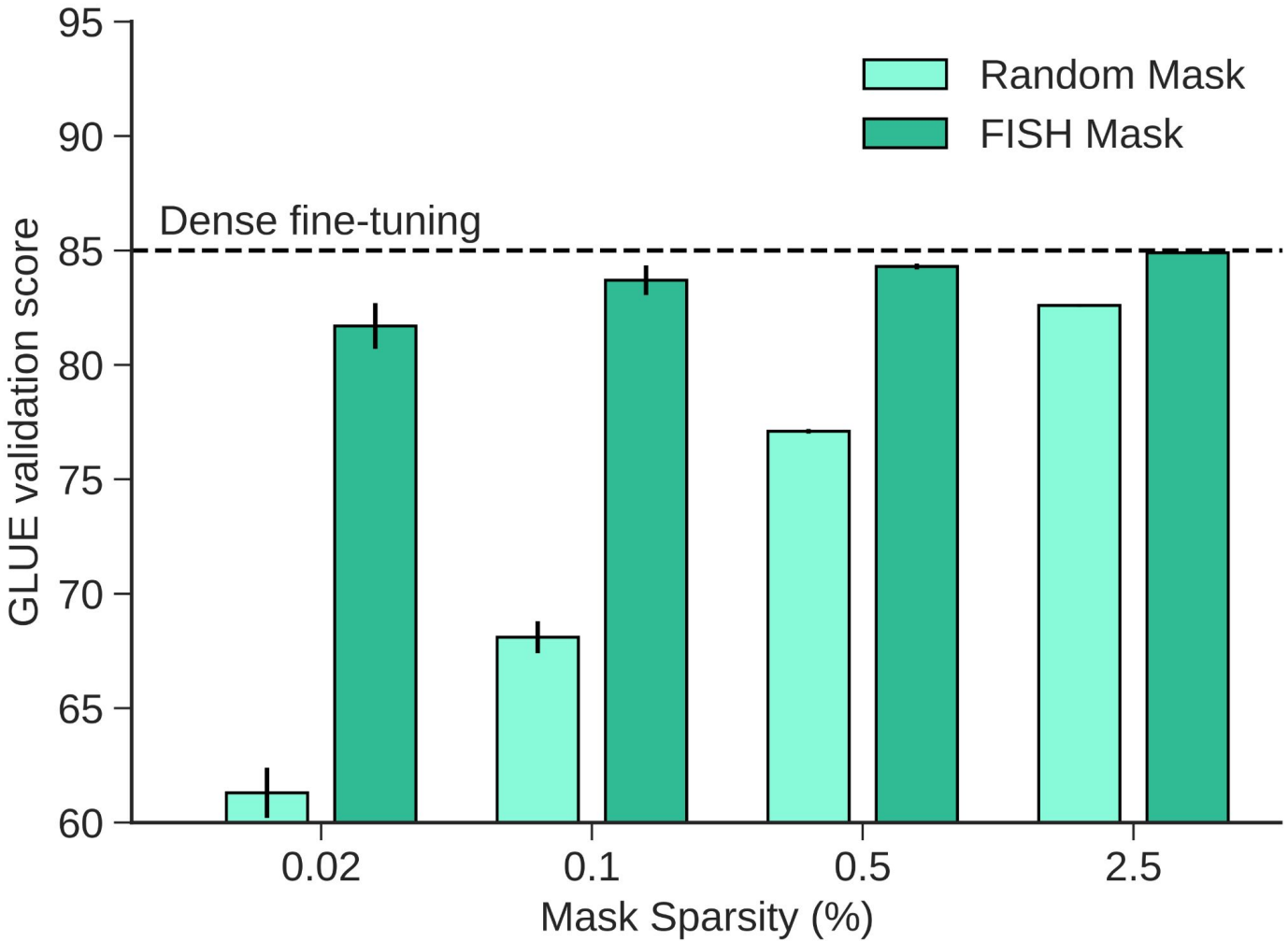
Contributors need to be able to cheaply communicate **patches** to a model.

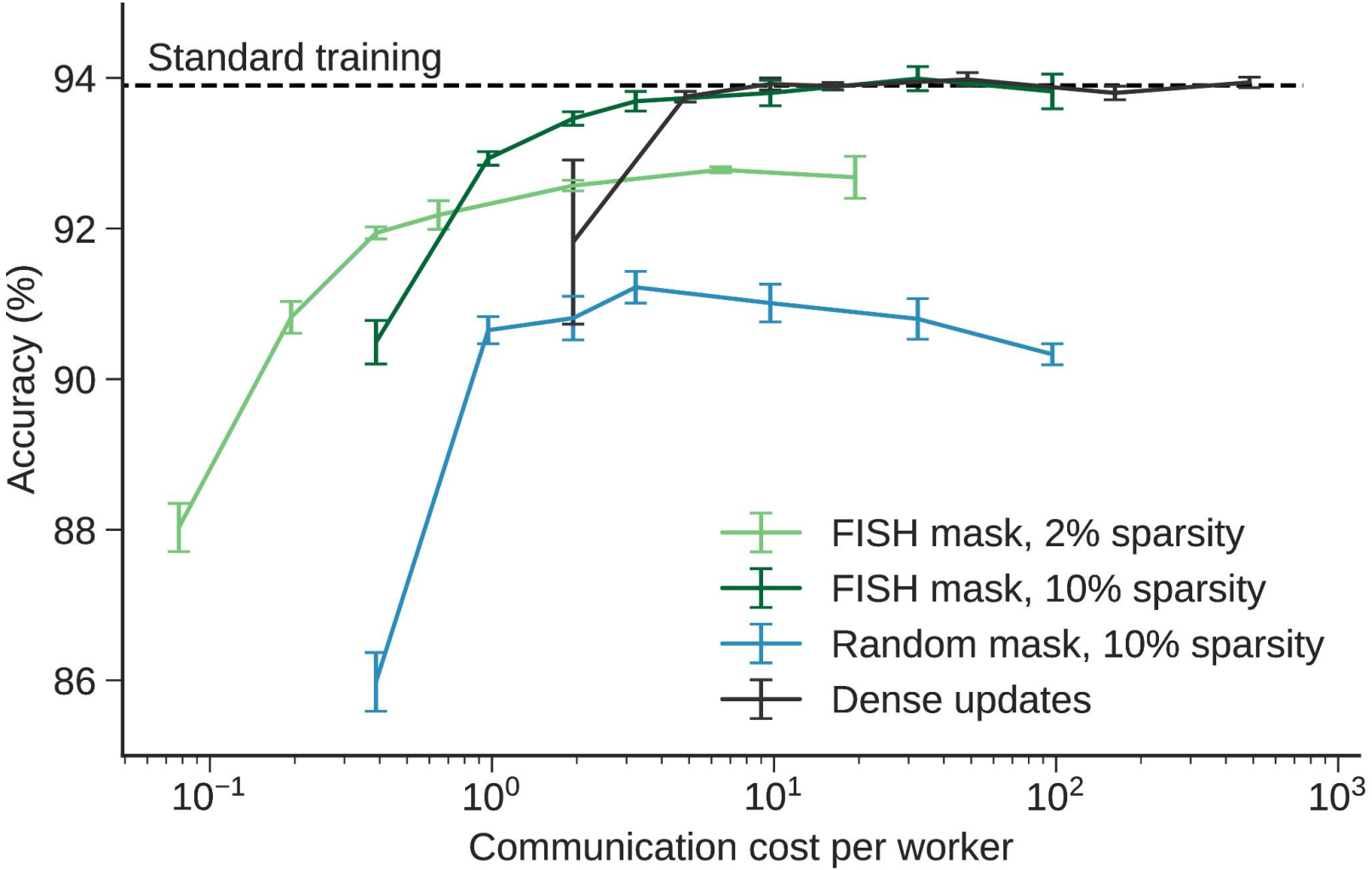


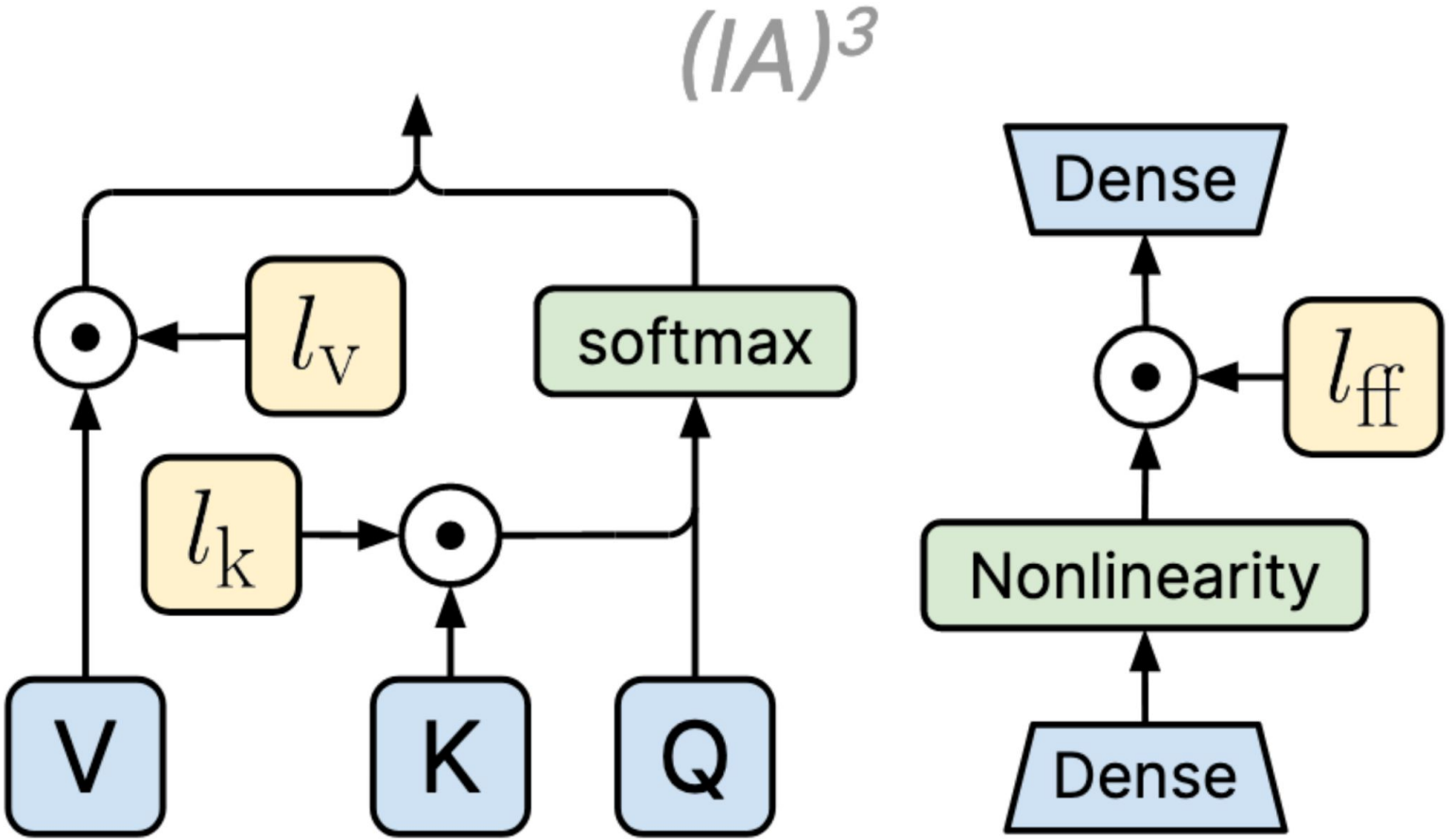


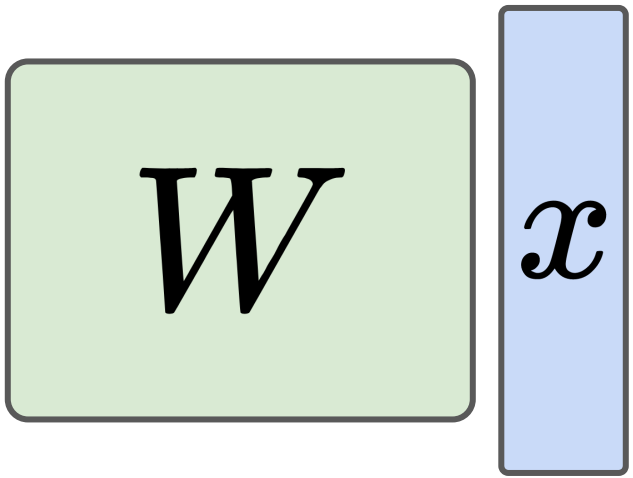


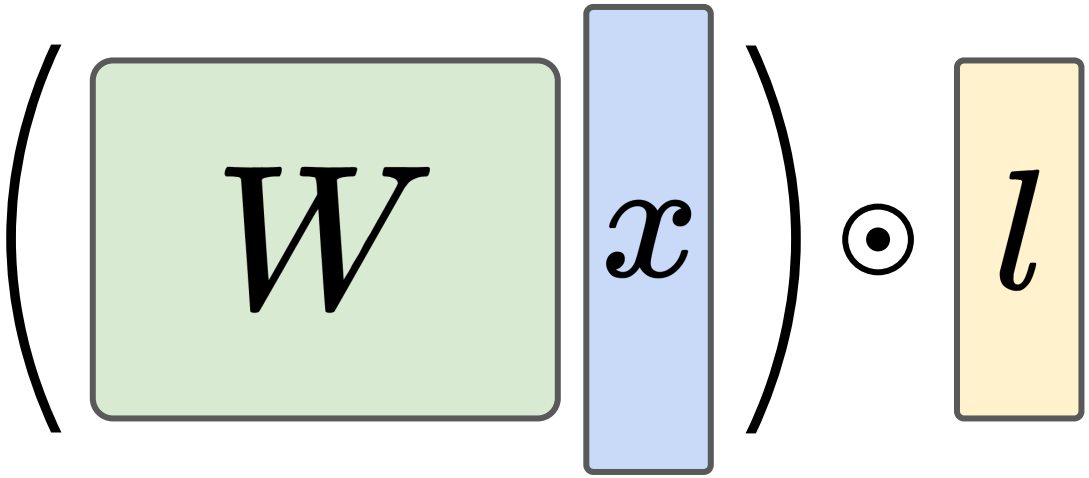
***F**isher-**I**nduced **S**parse Unc**h**anging (FISH) Mask*

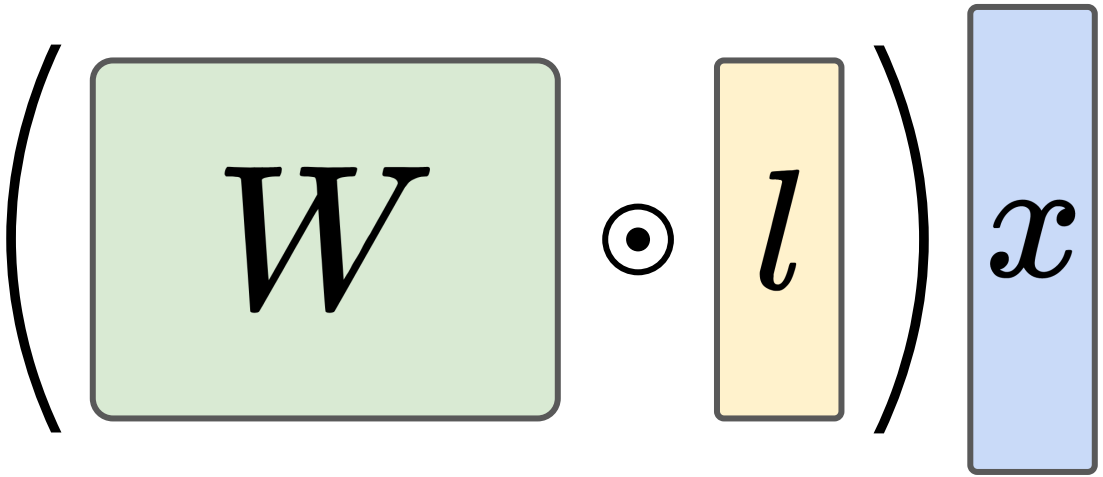

95
90
85
80
75
70
65
60
GLUE validation score
Dense fine-tuning
Random Mask
FISH Mask
0.02
0.1
0.5
2.5
Mask Sparsity (%)

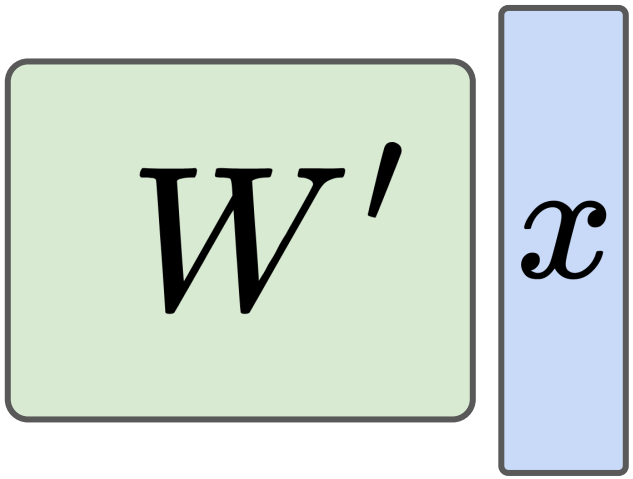

Standard training
94
92
90
88
86
Accuracy (%)
10^{-1}
10^{0}
10^{1}
10^{2}
10^{3}
Communication cost per worker
FISH mask, 2% sparsity
FISH mask, 10% sparsity
Random mask, 10% sparsity
Dense updates

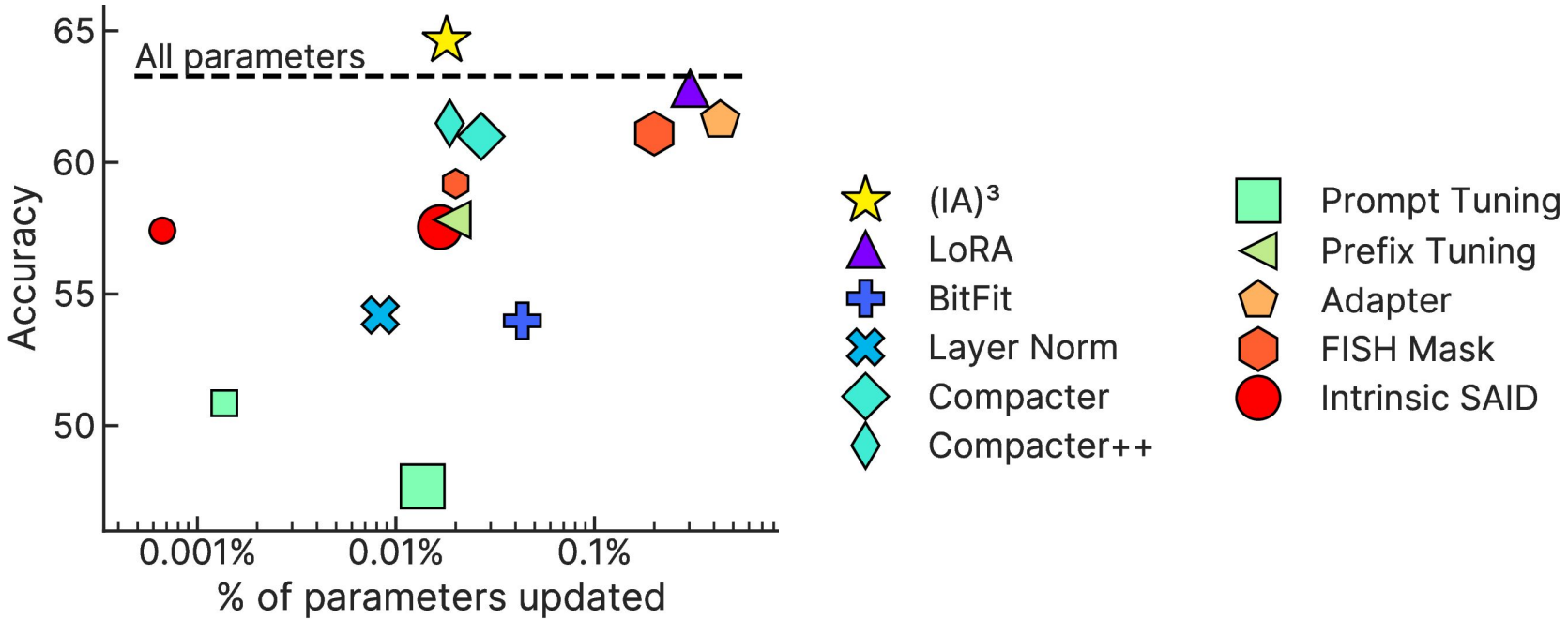

(IA)³
l_{v}
softmax
l_{k}
V
K
Q
Dense
l_{ff}
Nonlinearity
Dense


W
x


$\left(W \right| x \left) \odot \right| l$

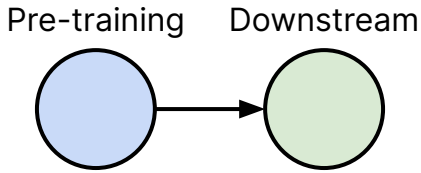

W
l
x

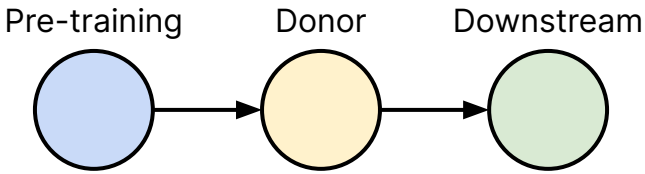

W'
x

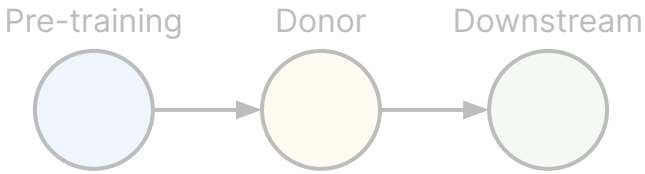

65
60
55
50
All parameters
Accuracy
0.001%
0.01%
0.1%
% of parameters updated
(IA)³
LoRA
BitFit
Layer Norm
Compacter
Compacter++
Prompt Tuning
Prefix Tuning
Adapter
FISH Mask
Intrinsic SAID

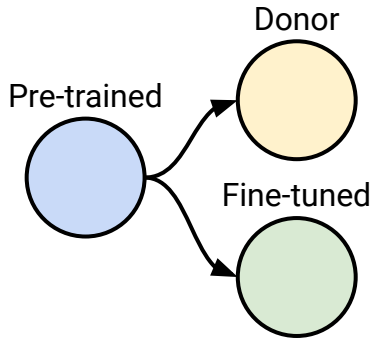
How can we enable collaborative and continual development of machine learning models?

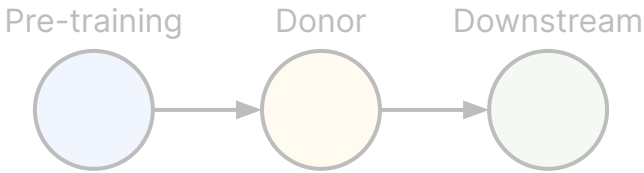
Maintainers need to be able to **merge** updates from different contributors.

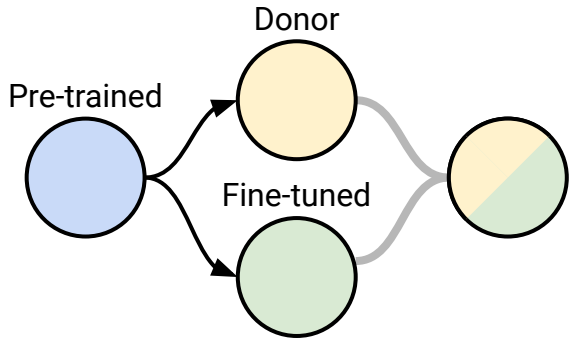

Pre-training
Downstream

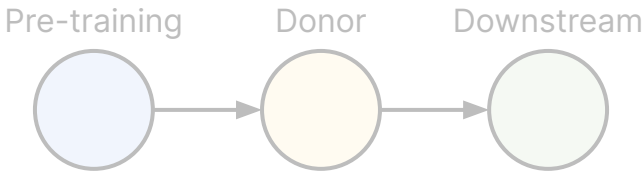

Pre-training
Donor
Downstream

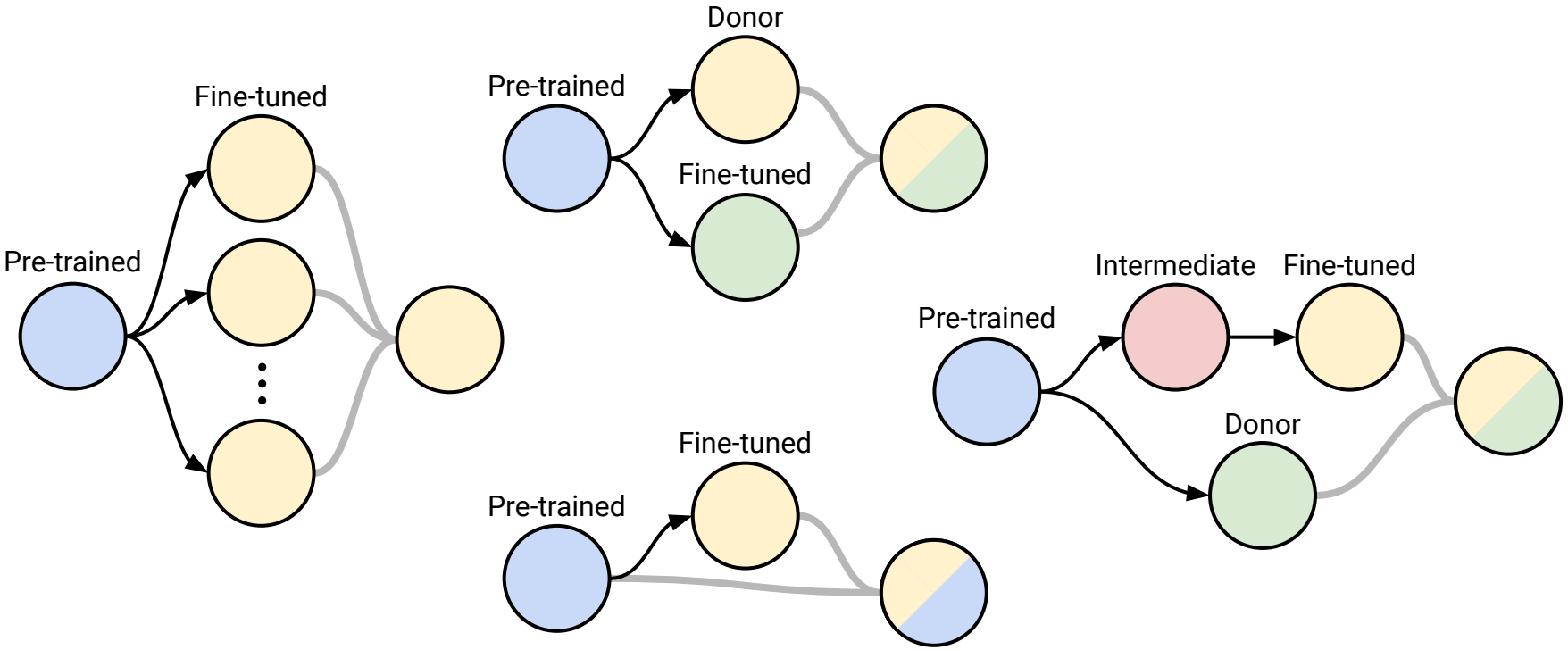

Pre-training
Donor
Downstream


Donor
Pre-trained
Fine-tuned


Pre-training
Donor
Downstream


Donor
Pre-trained
Fine-tuned


Pre-training
Donor
Downstream


Pre-trained
Fine-tuned
Pre-trained
Donor
Fine-tuned
Pre-trained
Intermediate
Fine-tuned
Donor
Pre-trained
Fine-tuned

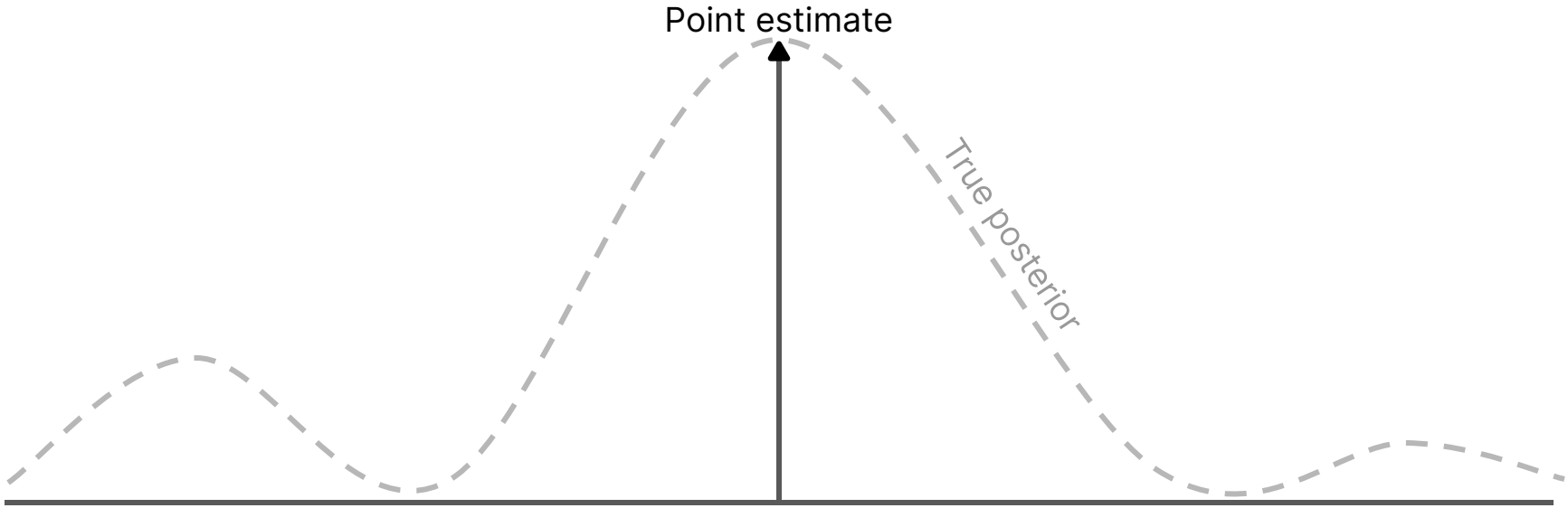
$$\arg\max_{\theta} \sum_{i=1}^{M} \lambda_i \log p(\theta|\mathcal{D}_i)$$

$$\arg\max_{\theta} \sum_{i=1}^{M} \lambda_i \underbrace{\log p(\theta|\mathcal{D}_i)}_{\text{Log posterior for model i}}$$

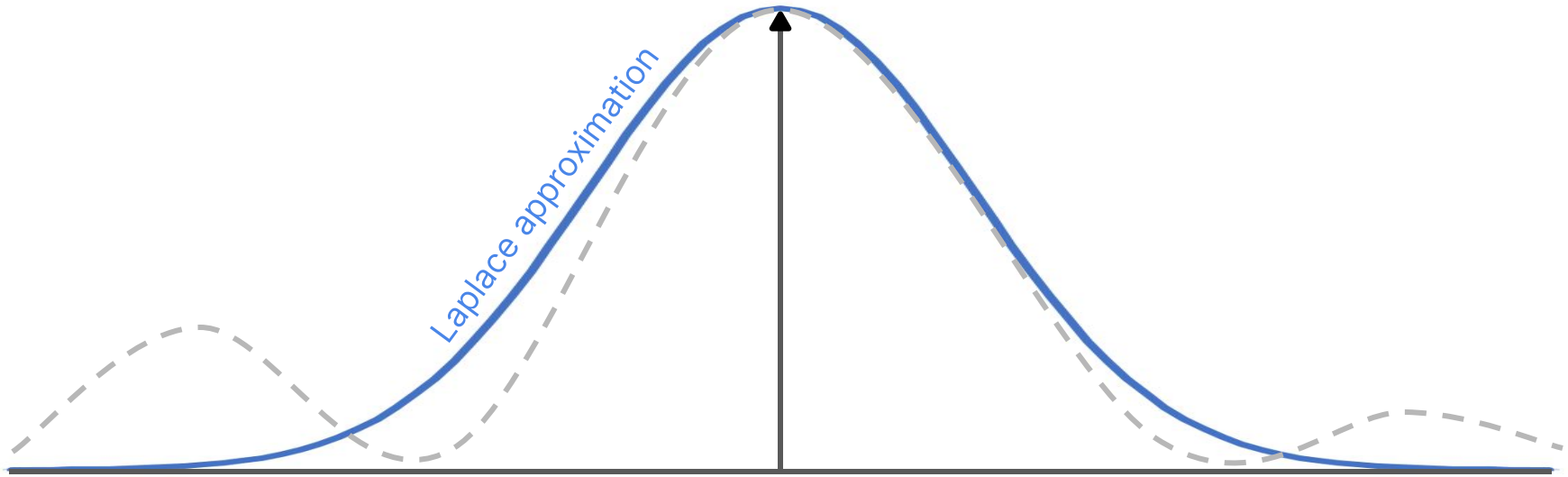
$$\arg\max_{\theta} \sum_{i=1}^{M} \lambda_i \log p(\theta|\mathcal{D}_i)$$

Hyperparameter controlling the importance of model i

$$\arg\max_{\theta} \sum_{i=1}^{M} \lambda_i \log p(\theta|\mathcal{D}_i)$$



$$\arg\max_{\theta} \sum_{i=1}^{M} \lambda_i \log \mathcal{N}(\theta|\theta_i, \hat{F}_i^{-1})$$



$$\arg\max_{\theta} \sum_{i=1}^{M} \lambda_i \log \mathcal{N}(\theta | \theta_i, \hat{F}_i^{-1})$$

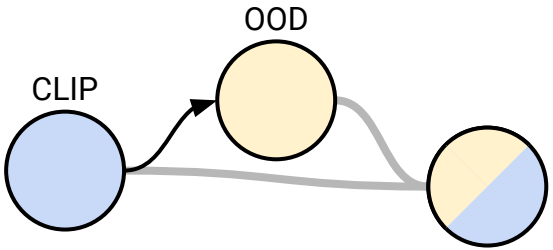
$$\theta^{*(j)} = \frac{\sum_{i=1}^{M} \lambda_i \hat{F}_i^{(j)} \theta_i^{(j)}}{\sum_{i=1}^{M} \lambda_i \hat{F}_i^{(j)}}$$

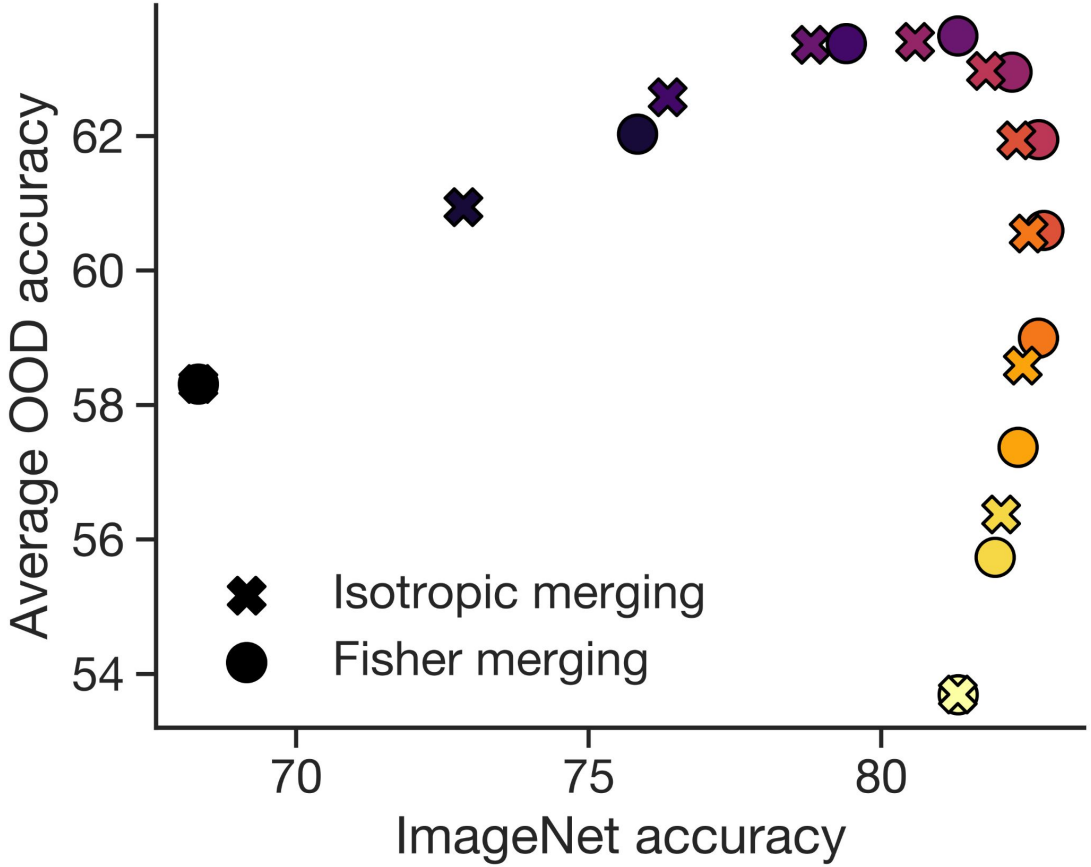
Fisher Merging

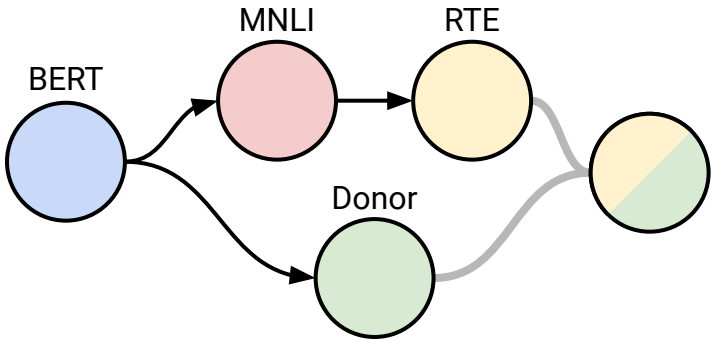
$$\arg\max_{\theta} \sum_{i=1}^{M} \lambda_i \log \mathcal{N}(\theta|\theta_i, \boldsymbol{I})$$

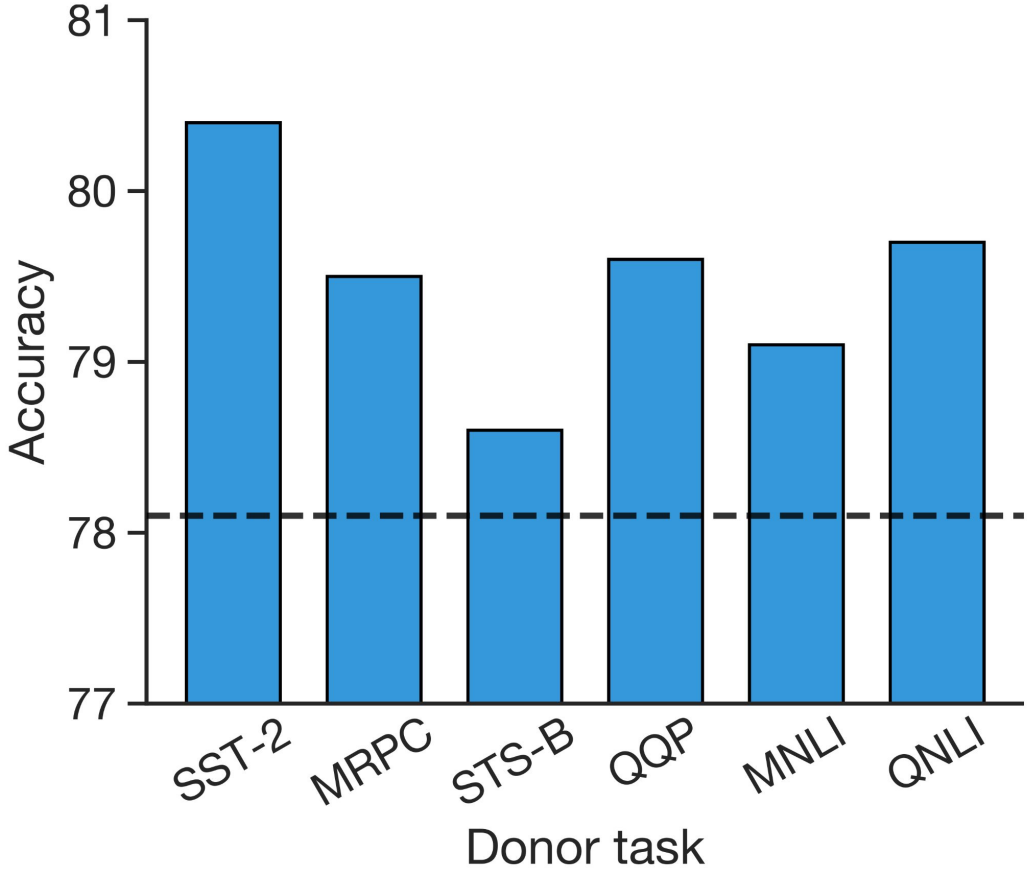
$$\theta^{*(j)} = \frac{\sum_{i=1}^{M} \lambda_i \theta_i^{(j)}}{\sum_{i=1}^{M} \lambda_i}$$

Isotropic Merging


OOD
CLIP

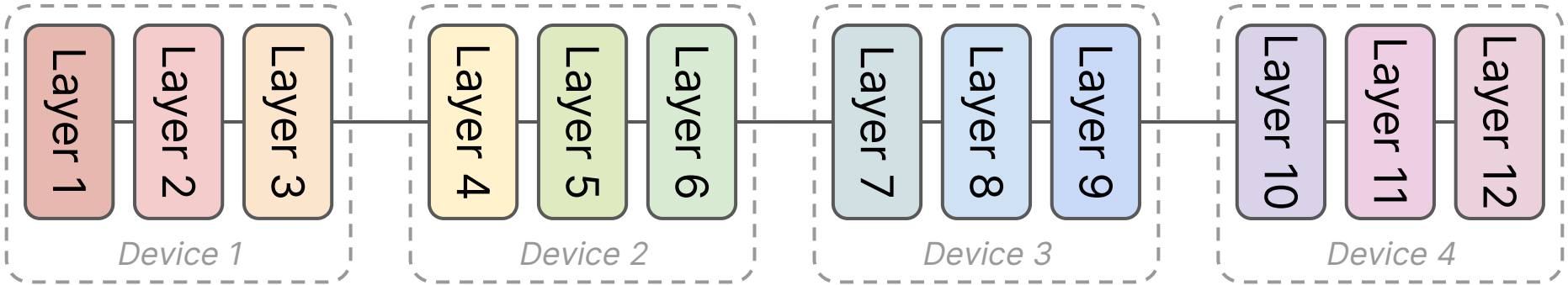

Average OOD accuracy
62
60
58
56
54
Isotropic merging
Fisher merging
70
75
80
ImageNet accuracy

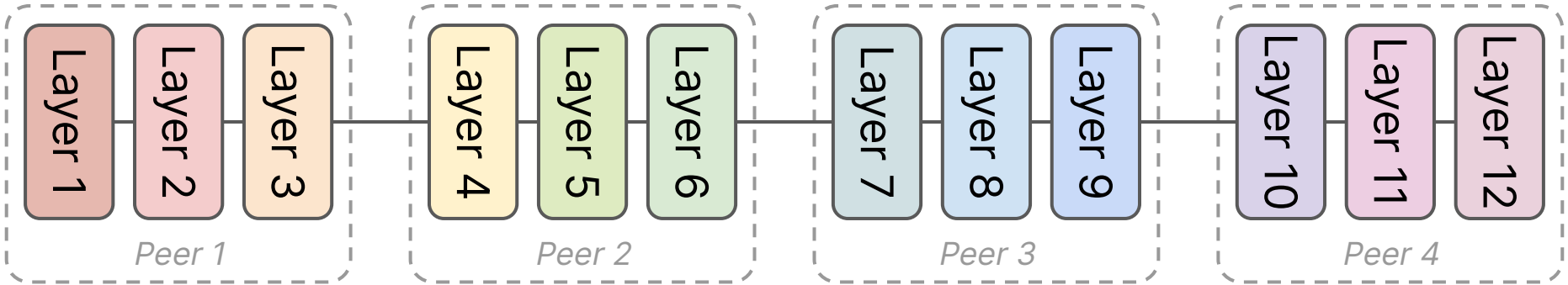

BERT
MNLI
RTE
Donor


81
80
79
78
77
Accuracy
SST-2
MRPC
STS-B
QQP
MNLI
QNLI
Donor task

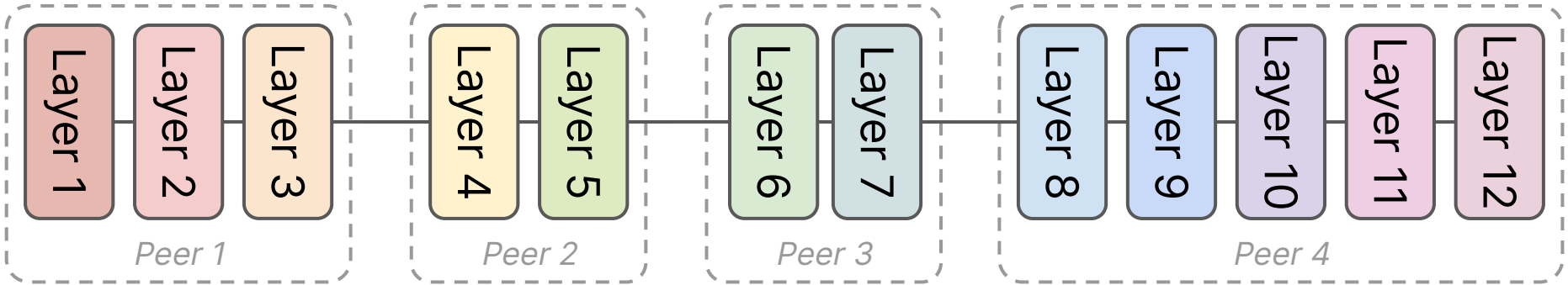
How can we enable collaborative and continual development of machine learning models?

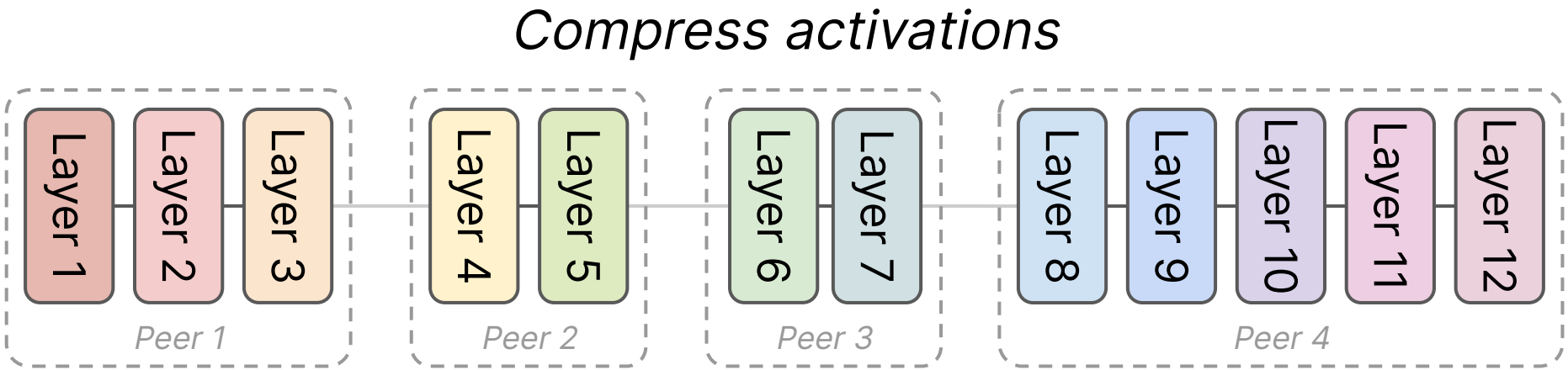
Users who lack resources need to be able to **train and run** large models.


Layer 1
Layer 2
Layer 3
Device 1
Layer 4
Layer 5
Layer 6
Device 2
Layer 7
Layer 8
Layer 9
Device 3
Layer 10
Layer 11
Layer 12
Device 4

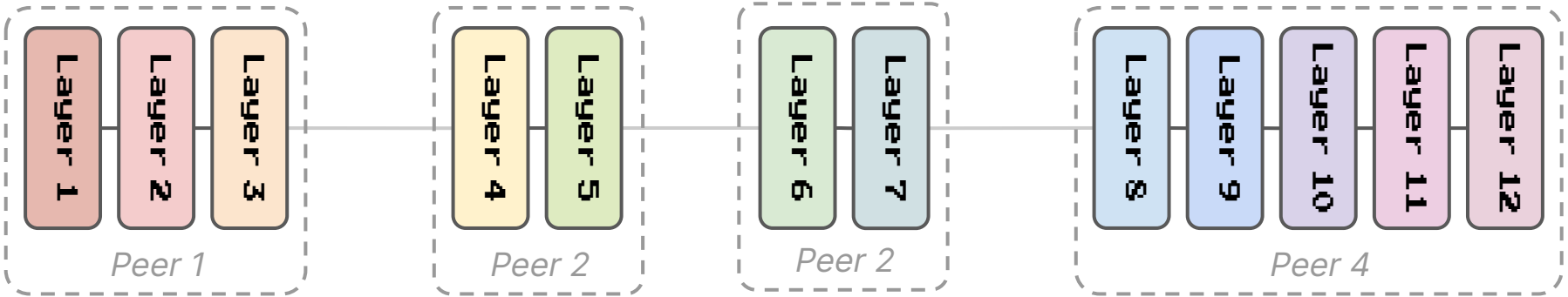

Layer 1
Layer 2
Layer 3
Peer 1
Layer 4
Layer 5
Layer 6
Peer 2
Layer 7
Layer 8
Layer 9
Peer 3
Layer 10
Layer 11
Layer 12
Peer 4

Distribute across peers with different compute budgets




Compress activations
Layer 1
Layer 2
Layer 3
Peer 1
Layer 4
Layer 5
Peer 2
Layer 6
Layer 7
Peer 3
Layer 8
Layer 9
Layer 10
Layer 11
Layer 12
Peer 4

Compress weights



Find the most efficient chain of peers

Network		Inference (steps / s)		Parallel (tokens / s)	
		Sequence length		Batch size	
Bandwidth	Latency	128	2048	1	128
3 local servers on 3xA100					
1 Gbit/s	< 5 ms	1.22	1.11	70.0	253.6
100 Mbit/s	< 5 ms	1.19	1.08	56.4	182.0
100 Mbit/s	100 ms	0.89	0.8	19.7	112.2
8 desktops & servers in Europe and North America					
Real world		0.63	0.57	28.3	135.4

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

Maintainers need to be able to quickly **vet** community contributions.

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.

A Call to Build Machine Learning Models like Open-Source Software
Colin Raffel

Training Neural Networks with Fixed Sparse Masks
Yi-Lin Sung*, Varun Nair*, and Colin Raffel

Few-Shot Parameter-Efficient Fine-Tuning is Better and Cheaper than In-Context Learning
Haokun Liu*, Derek Tam*, Mohammed Muqeeth*, Jay Mohta, Tenghao Huang, Mohit Bansal, and Colin Raffel

Merging Models with Fisher-Weighted Averaging
Michael Matena and Colin Raffel

PETALS: Collaborative Inference of Large Models
Alexander Borzunov*, Dmitry Baranchuk*, Tim Dettmers*, Max Ryabinin*, Younes Belkada*, Artem Chumachenko, Pavel Samygin, and Colin Raffel

Please give me feedback:
http://bit.ly/colin-talk-feedback