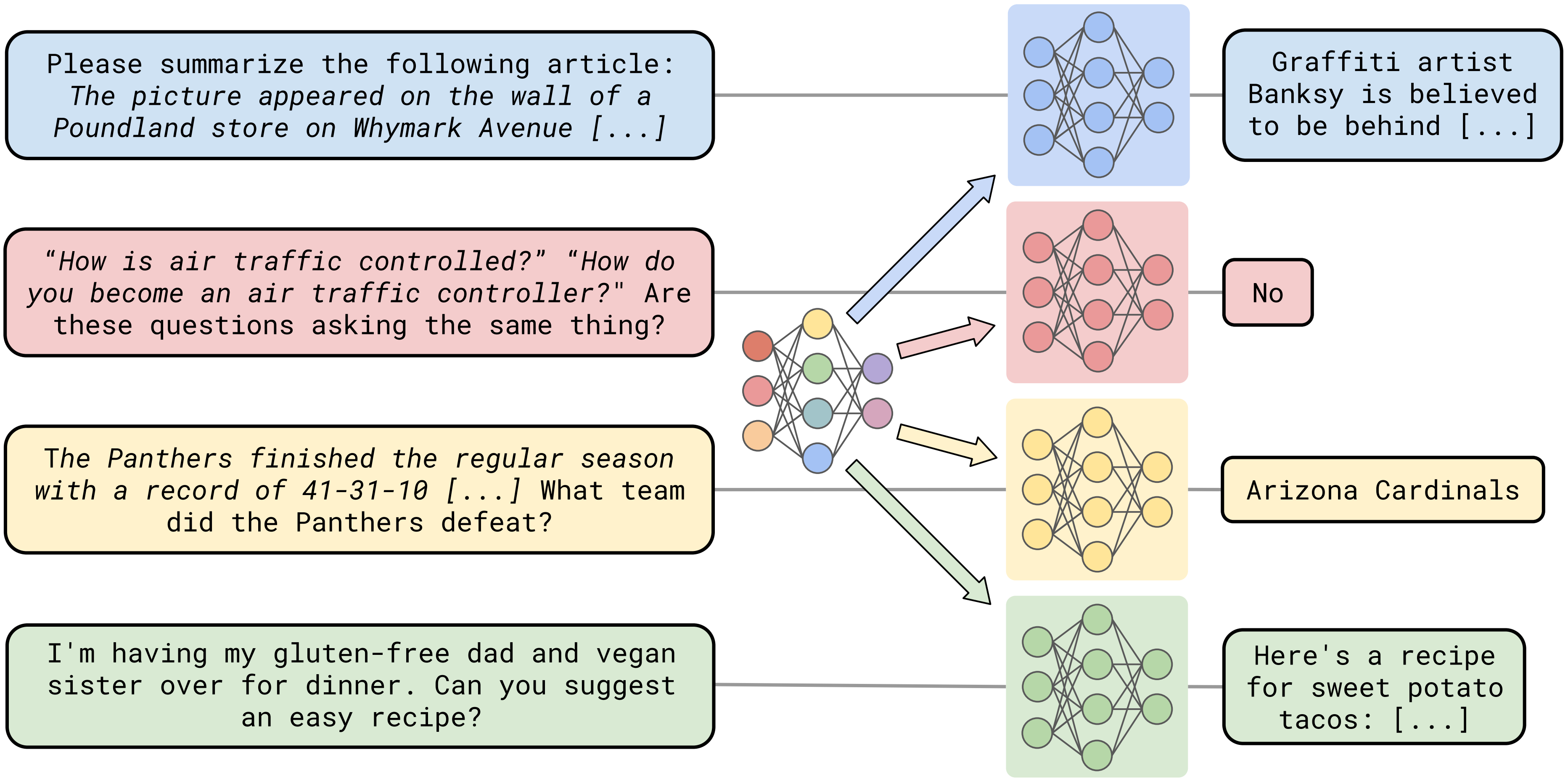
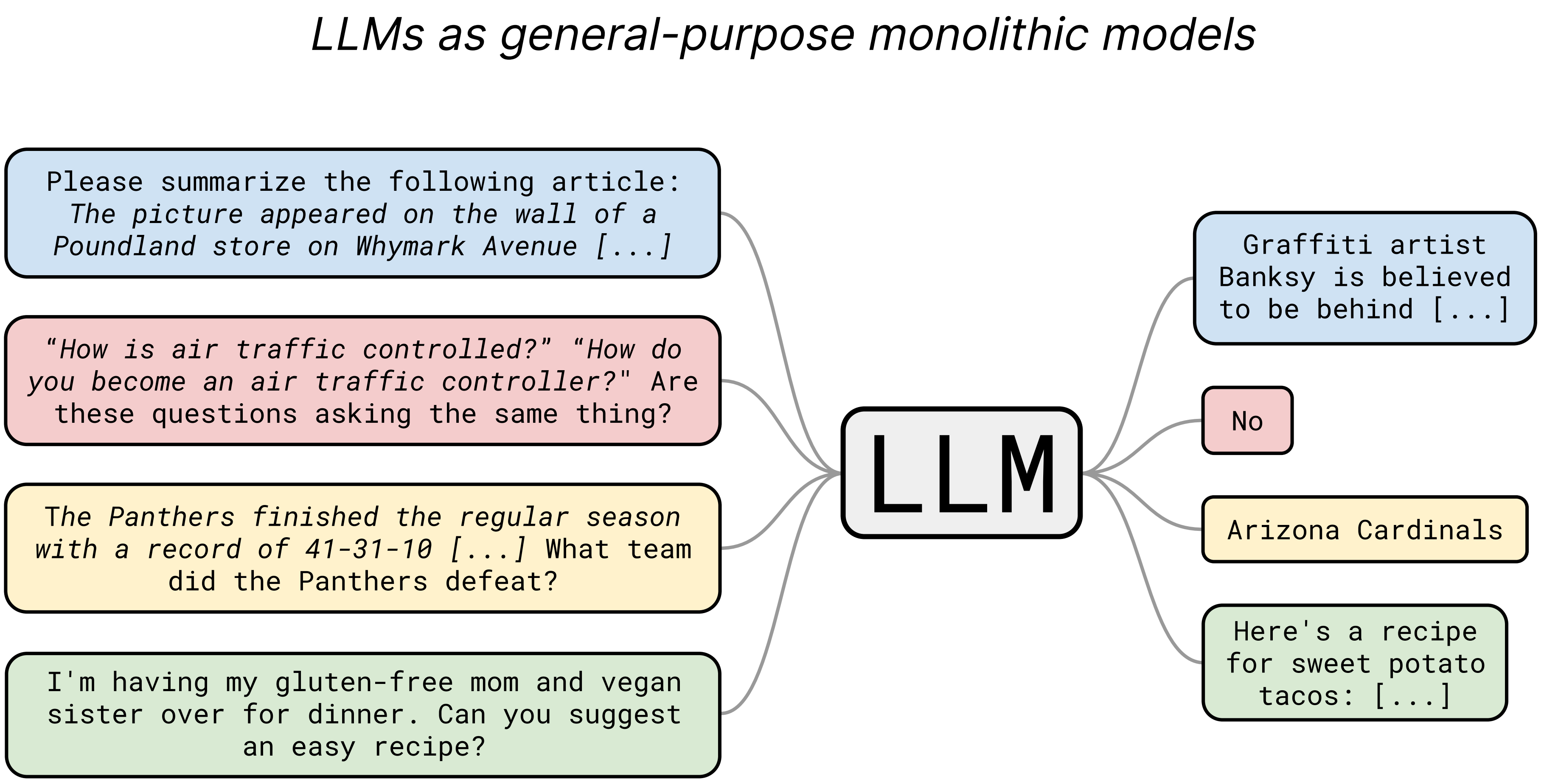
Build an Ecosystem, Not a Monolith

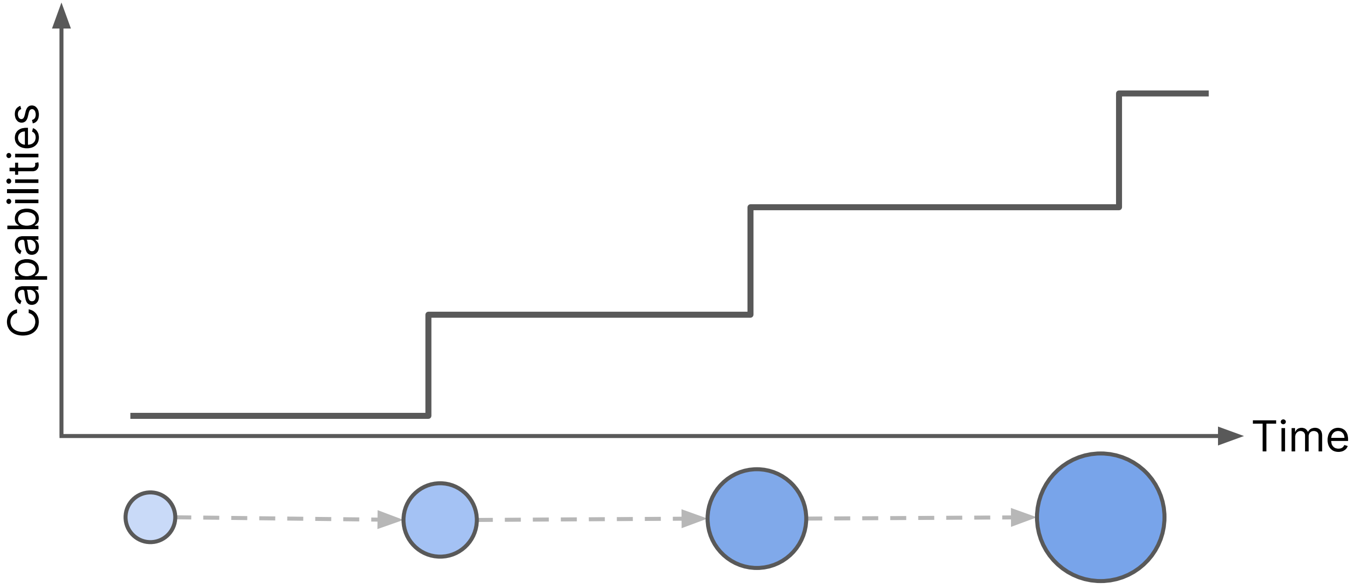
Colin Raffel

Transfer learning: fine-tuning to create specialized models

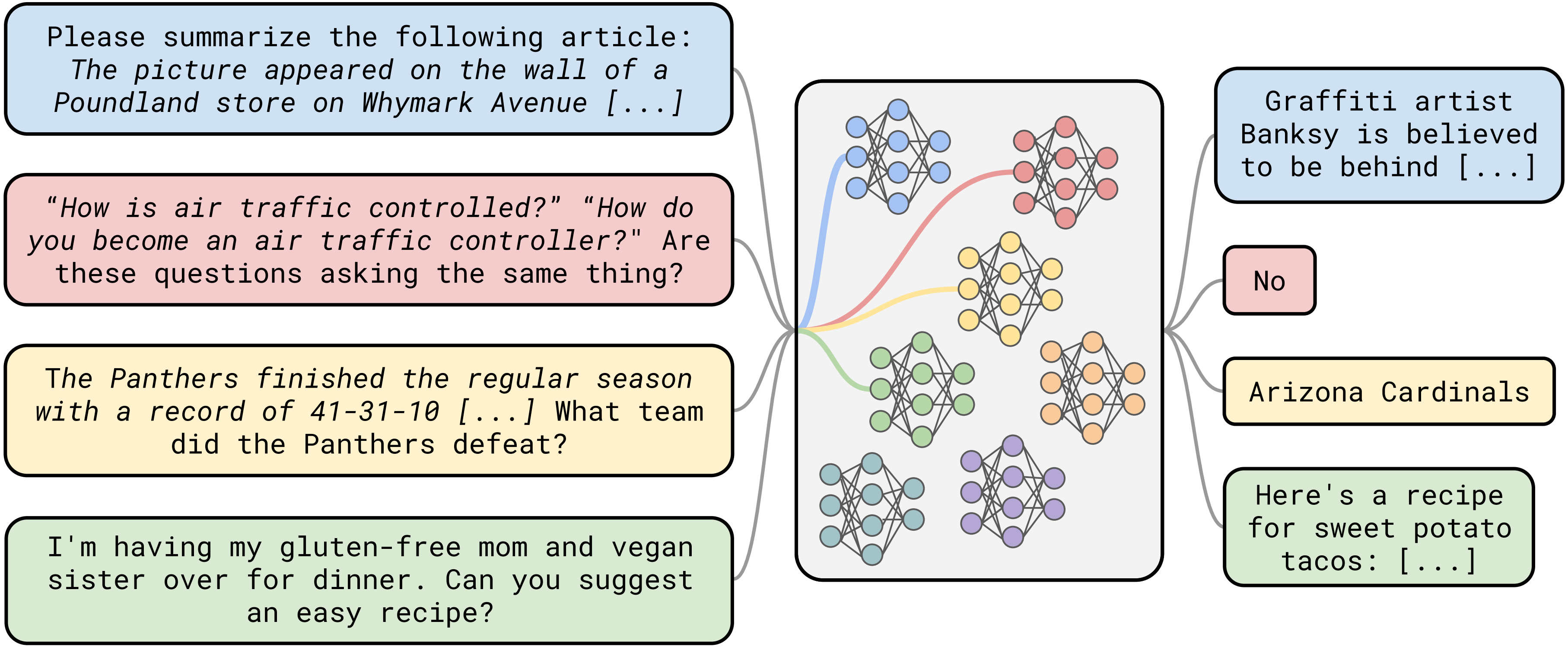



LLMs as general-purpose monolithic models
Please summarize the following article: *The picture appeared on the wall of a Poundland store on Whymark Avenue [...]*
"How is air traffic controlled?" "How do you become an air traffic controller?" Are these questions asking the same thing?
The Panthers finished the regular season with a record of 41-31-10 [...] What team did the Panthers defeat?
I'm having my gluten-free mom and vegan sister over for dinner. Can you suggest an easy recipe?
LLM
Graffiti artist Banksy is believed to be behind [...]
No
Arizona Cardinals
Here's a recipe for sweet potato tacos: [...]

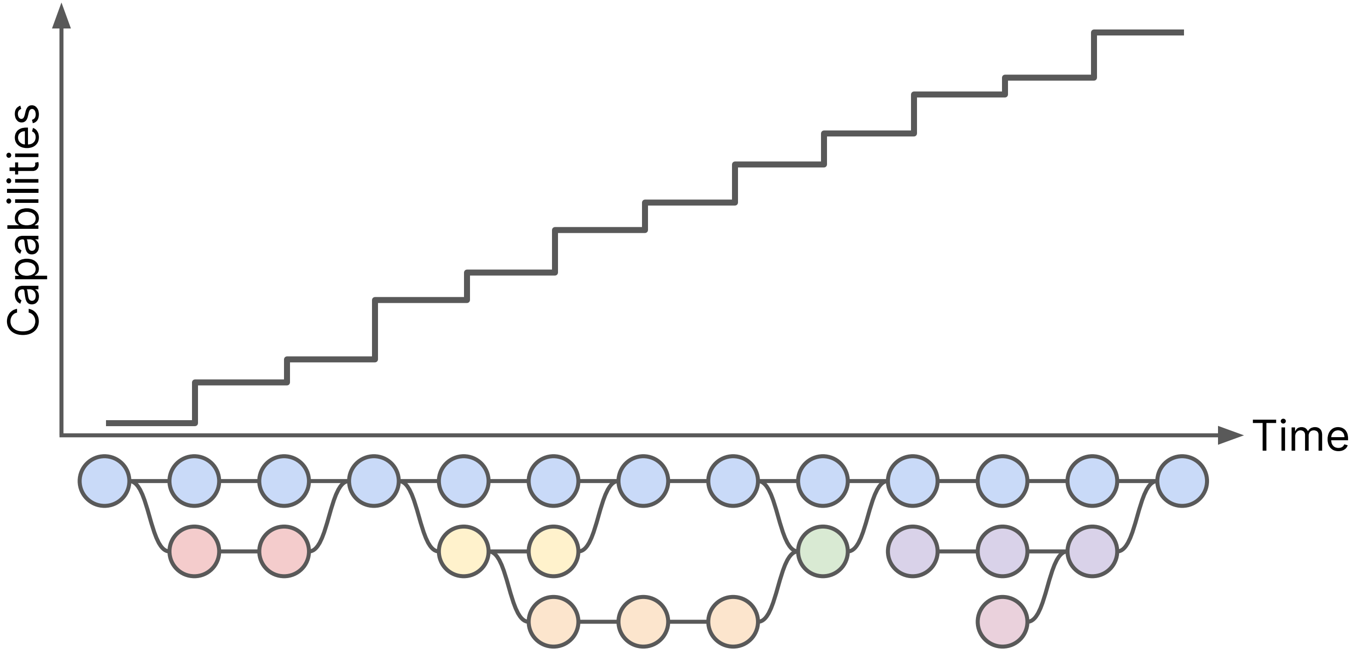
Monolithic model development involves wholesale replacement



Ecosystems of specialist models?



Collaborative ecosystem development will lead to continual improvements

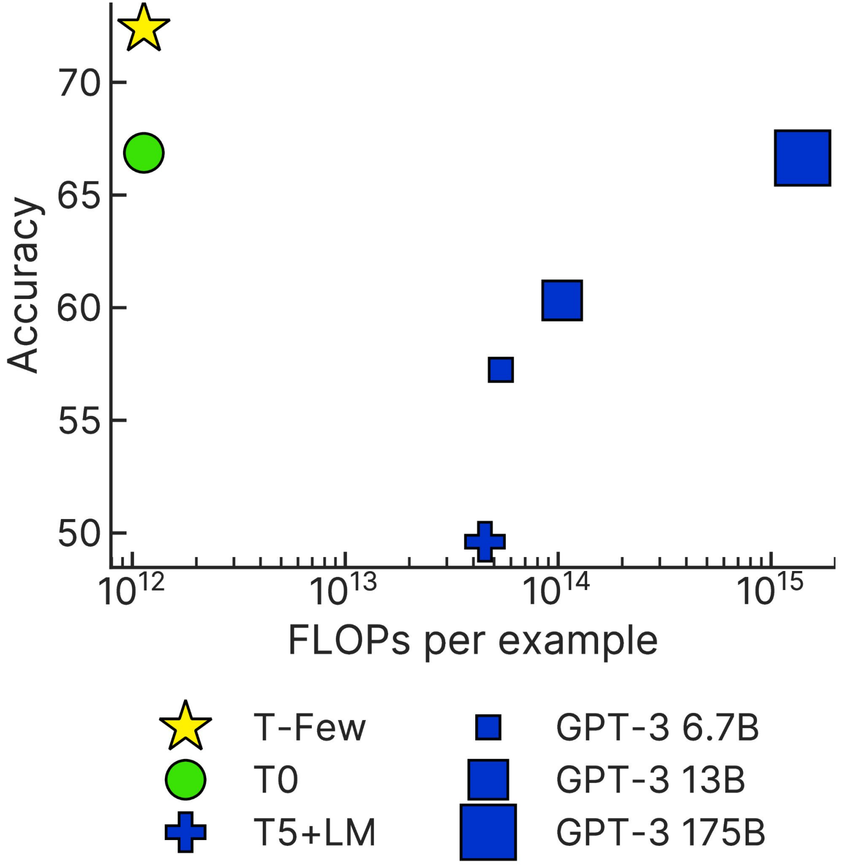


How and why should we build ecosystems of specialist models instead of monolithic models?

How and why should we build ecosystems of specialist models instead of monolithic models?

Specialist models are often **cheaper** and sometimes **better**.

Smaller fine-tuned models often outperform larger generalist models



From “Few-Shot Parameter-Efficient Fine-Tuning is Better and Cheaper than In-Context Learning" by Liu et al.

Smaller fine-tuned models often outperform larger generalist models

Specialist models

	SOTA	GPT-4	PaLM	PaLM 2
WinoGrande	87.5[a]	87.5[a] (5)	85.1[b] (5)	**90.9** (5)
ARC-C	**96.3**[a]	**96.3**[a] (25)	88.7[c] (4)	95.1 (4)
DROP	**88.4**[d]	80.9[a] (3)	70.8[b] (1)	85.0 (3)
StrategyQA	81.6[c]	-	81.6[c] (6)	**90.4** (6)
CSQA	**91.2**[e]	-	80.7[c] (7)	90.4 (7)
XCOPA	89.9[g]	-	89.9[g] (4)	**94.4** (4)
BB Hard	65.2[f]	-	65.2[f] (3)	**78.1** (3)

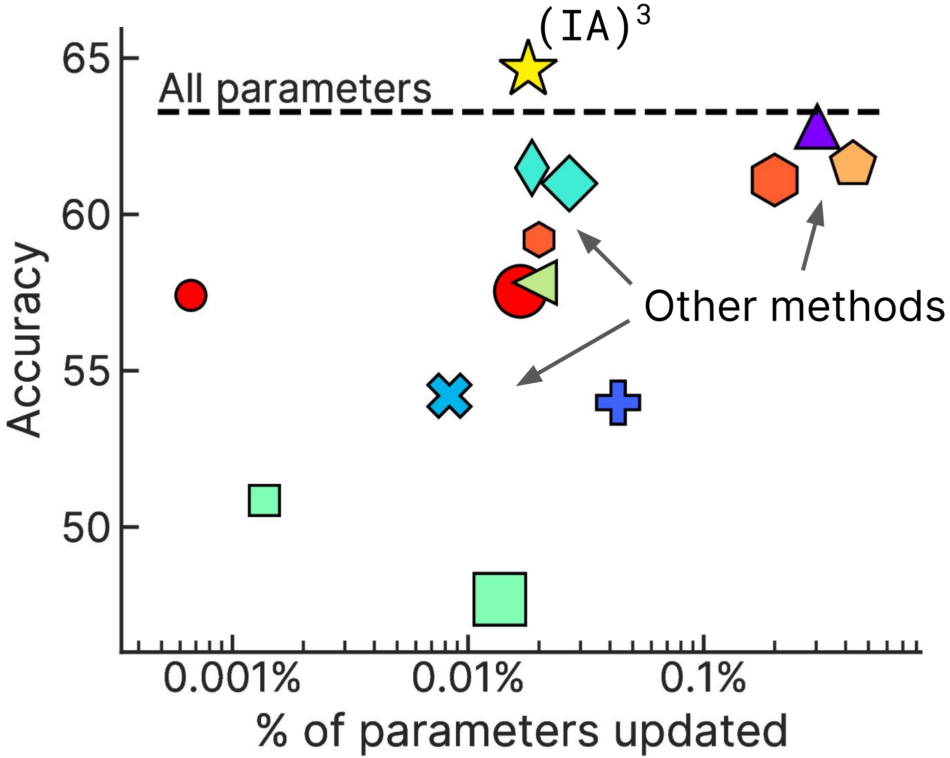
	Chinese⟶English		English⟶German	
	BLEURT ↑	MQM (Human) ↓	BLEURT ↑	MQM (Human) ↓
PaLM	67.4	3.7	71.7	1.2
Google Translate	68.5	3.1	73.0	1.0
PaLM 2	**69.2**	**3.0**	**73.3**	**0.9**

From "PaLM 2 Technical Report" by Google

How and why should we build ecosystems of specialist models instead of monolithic models?

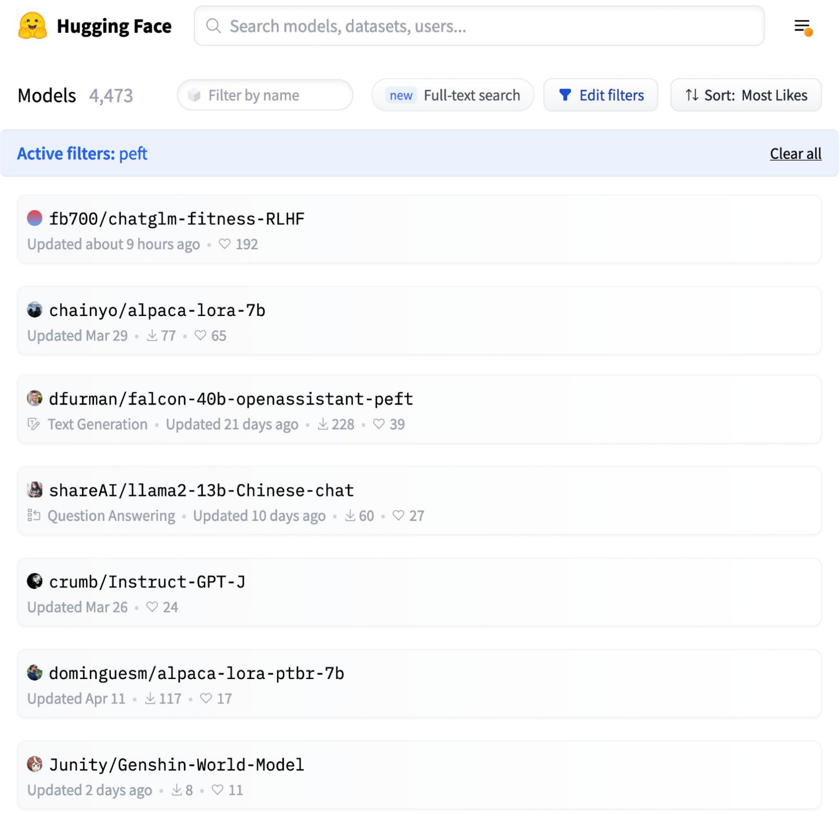
Each specialist model can be a **cheaply communicable** update to a base model.

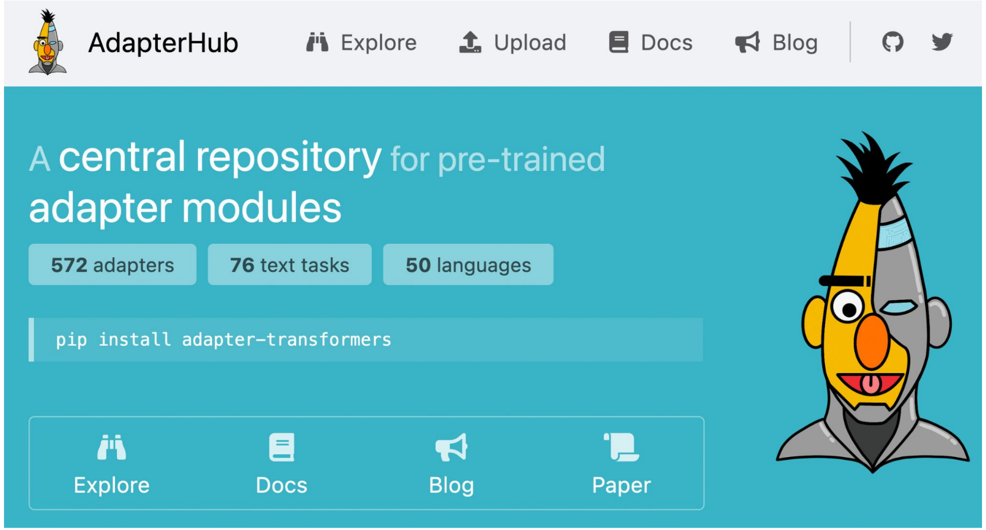
(IA)^3 *outperforms standard training while updating 0.01% of parameters*



From "Few-Shot Parameter-Efficient Fine-Tuning is Better and Cheaper than In-Context Learning" by Liu et al.

Existing "adapter" hubs have thousands of specialized models

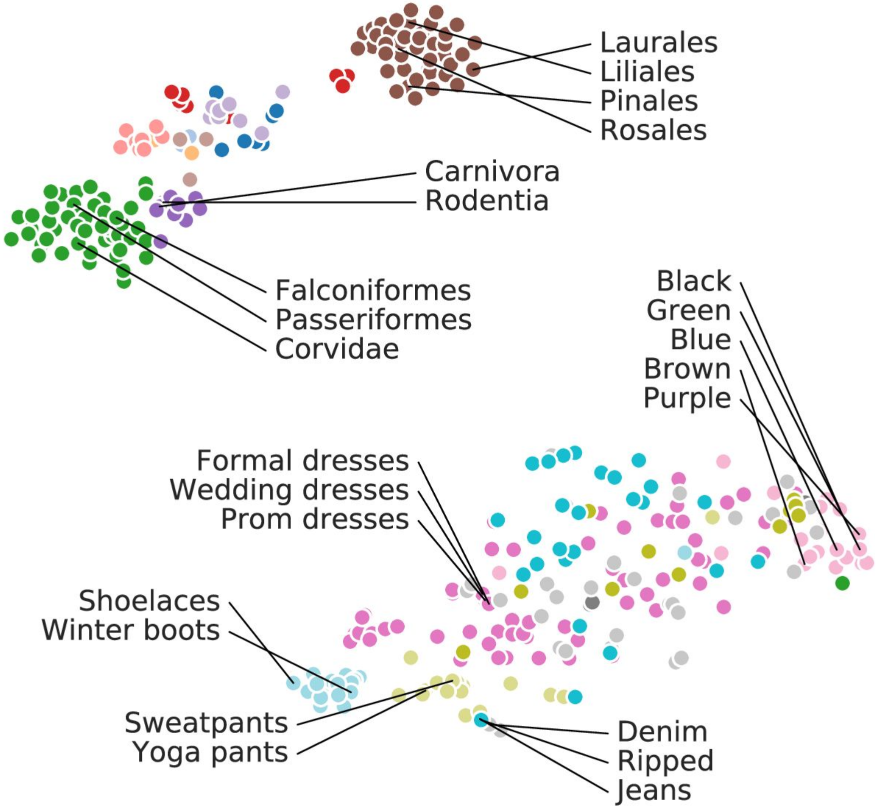




How and why should we build ecosystems of specialist models instead of monolithic models?

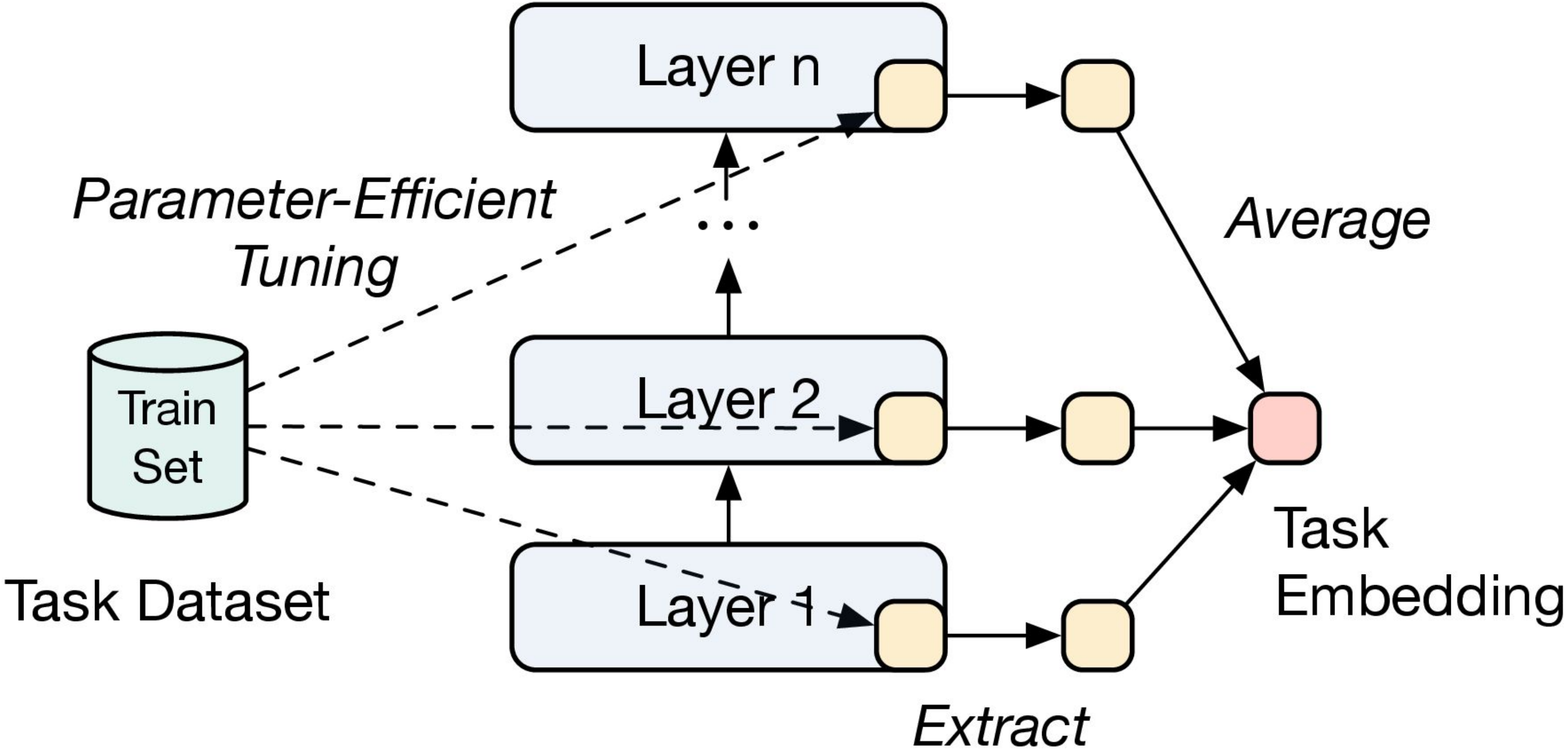
The appropriate model for a query should be **chosen automatically**.

task2vec encodes task similarity via the Fisher information matrix



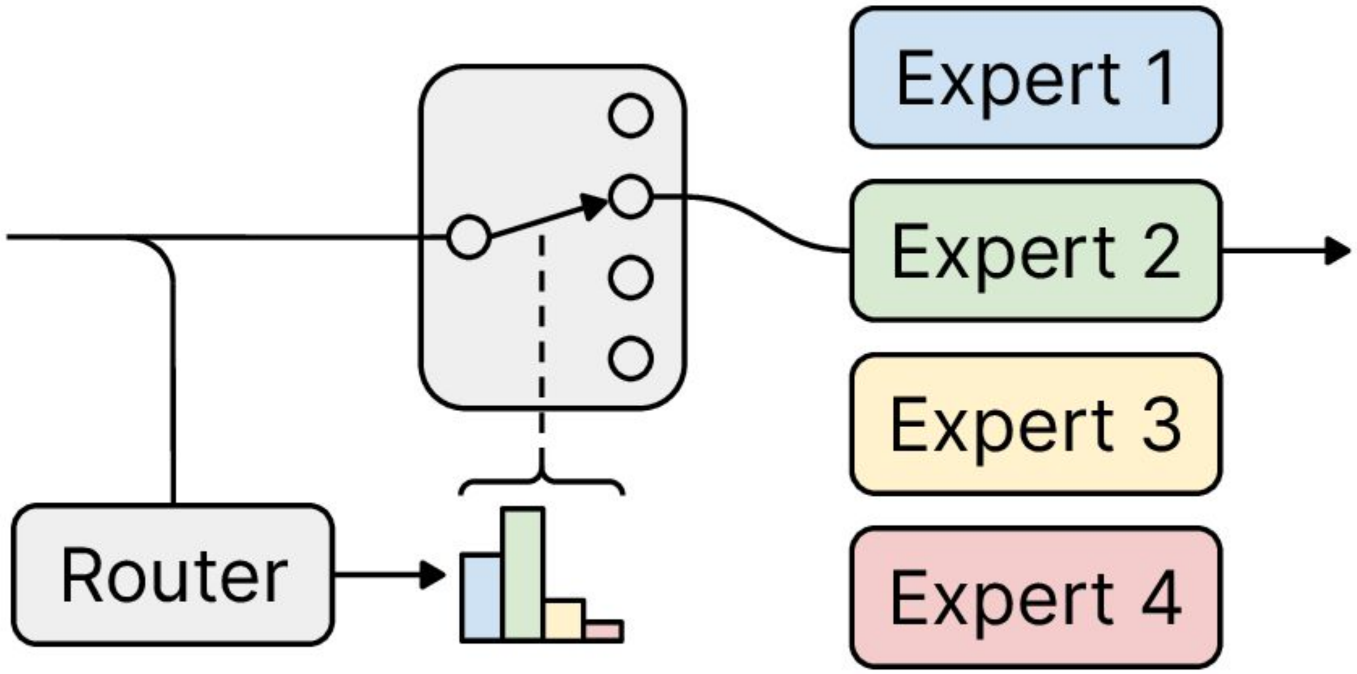
From "task2vec: Task Embedding for Meta-Learning" by Achille et al.

Adapter parameters also encode task similarity



From "Efficiently Tuned Parameters are Task Embeddings" by Zhou et al.

Mixture-of-experts models perform adaptive routing inside the model

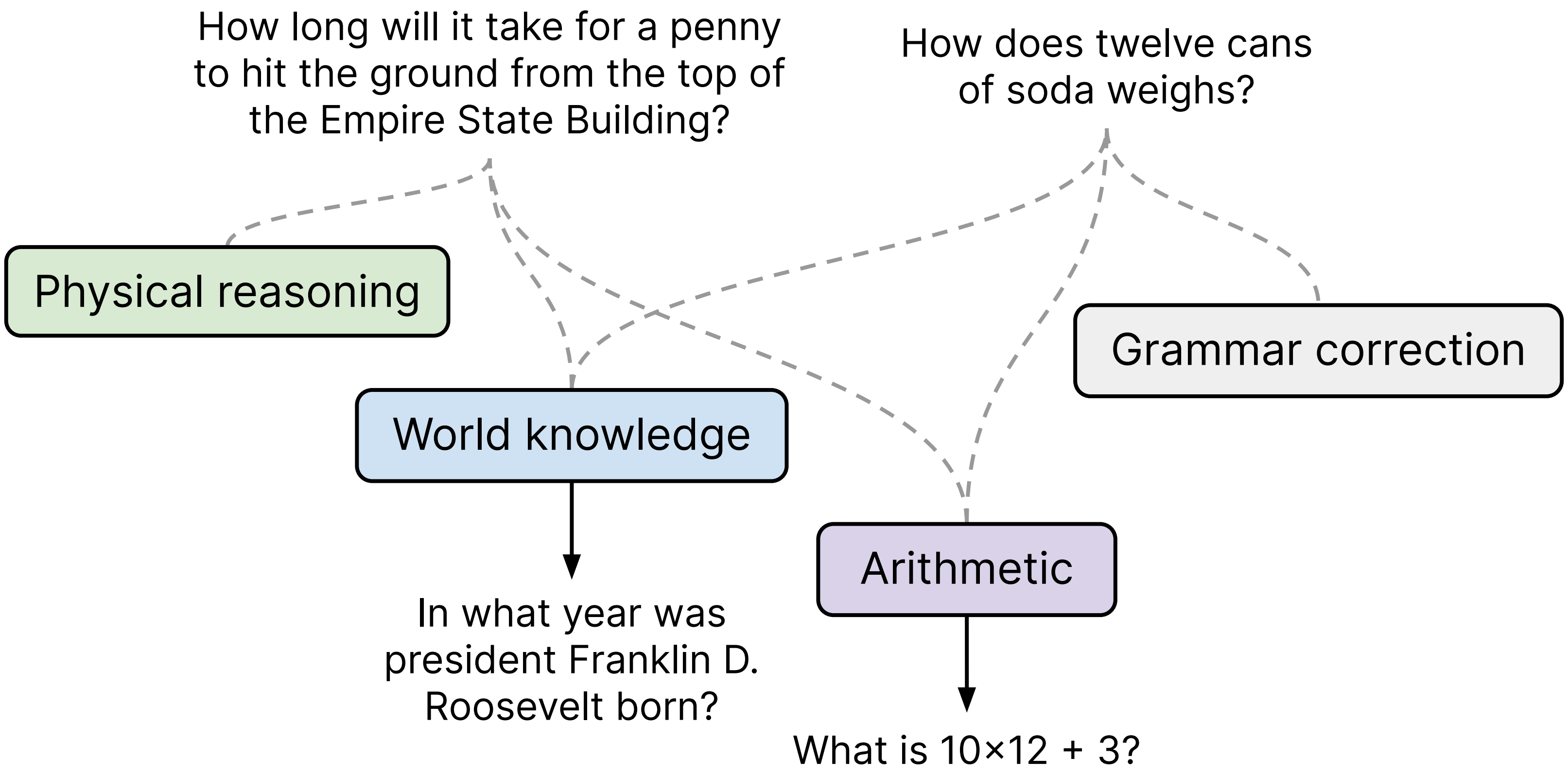


From "Soft Merging of Experts with Adaptive Routing" by Muqeeth et al.

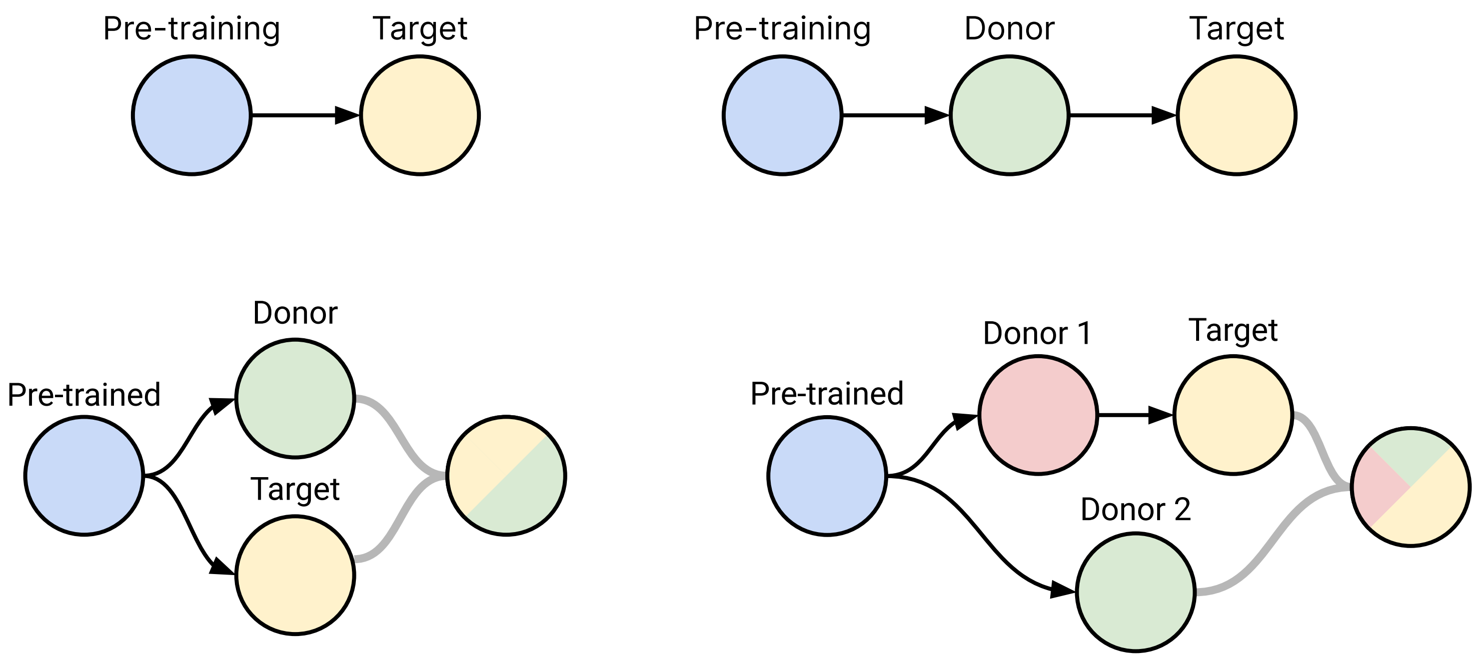
How and why should we build ecosystems of specialist models instead of monolithic models?

Capabilities can be **merged** across models.

Tasks can be considered as a composition of skills

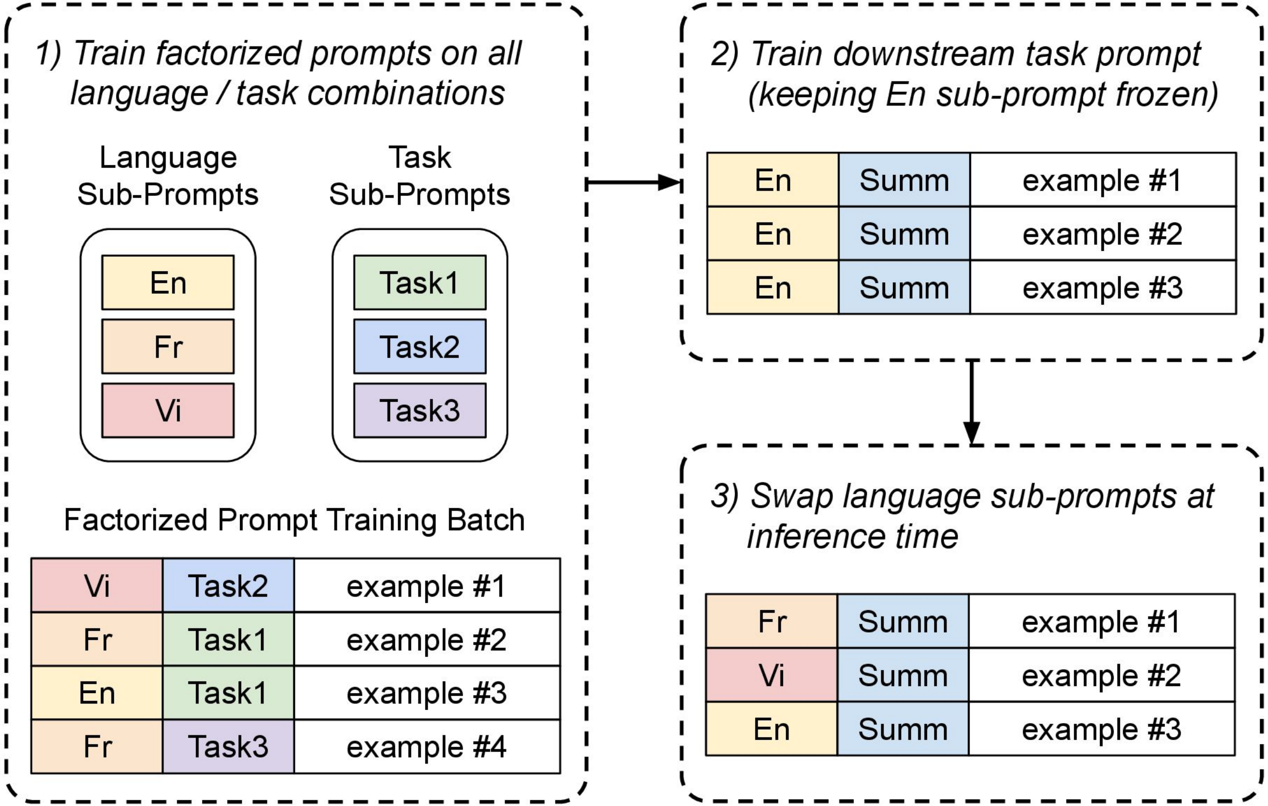


Merging models enables new paths for transferring capabilities



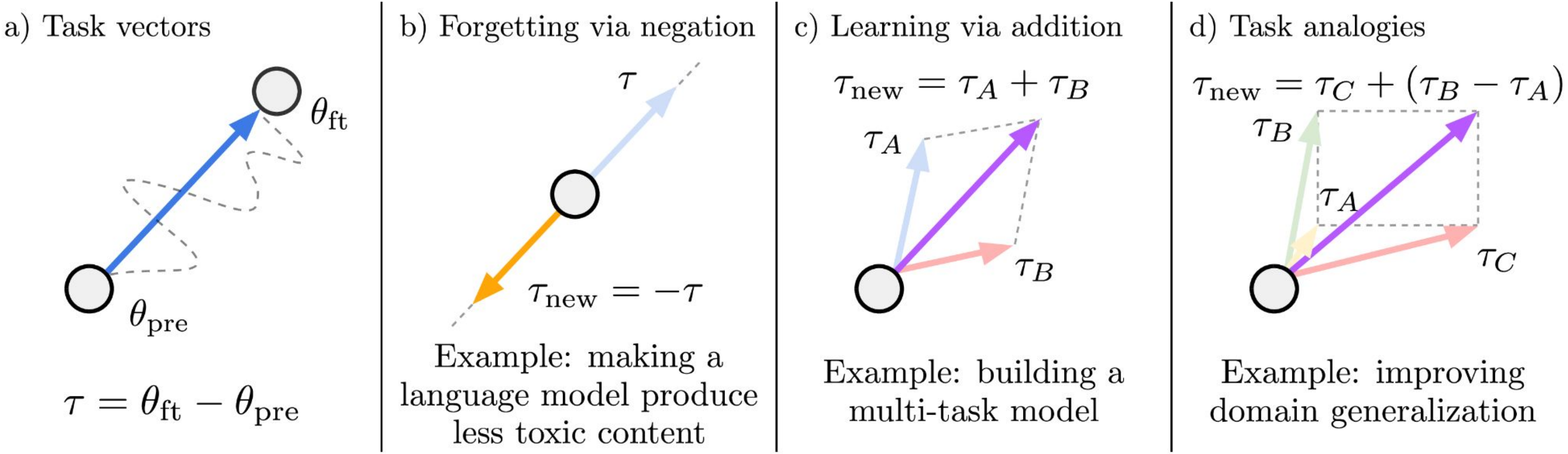
From "Merging Models with Fisher-Weighted Averaging" by Matena et al.

Learning compositional adapters via prompt tuning



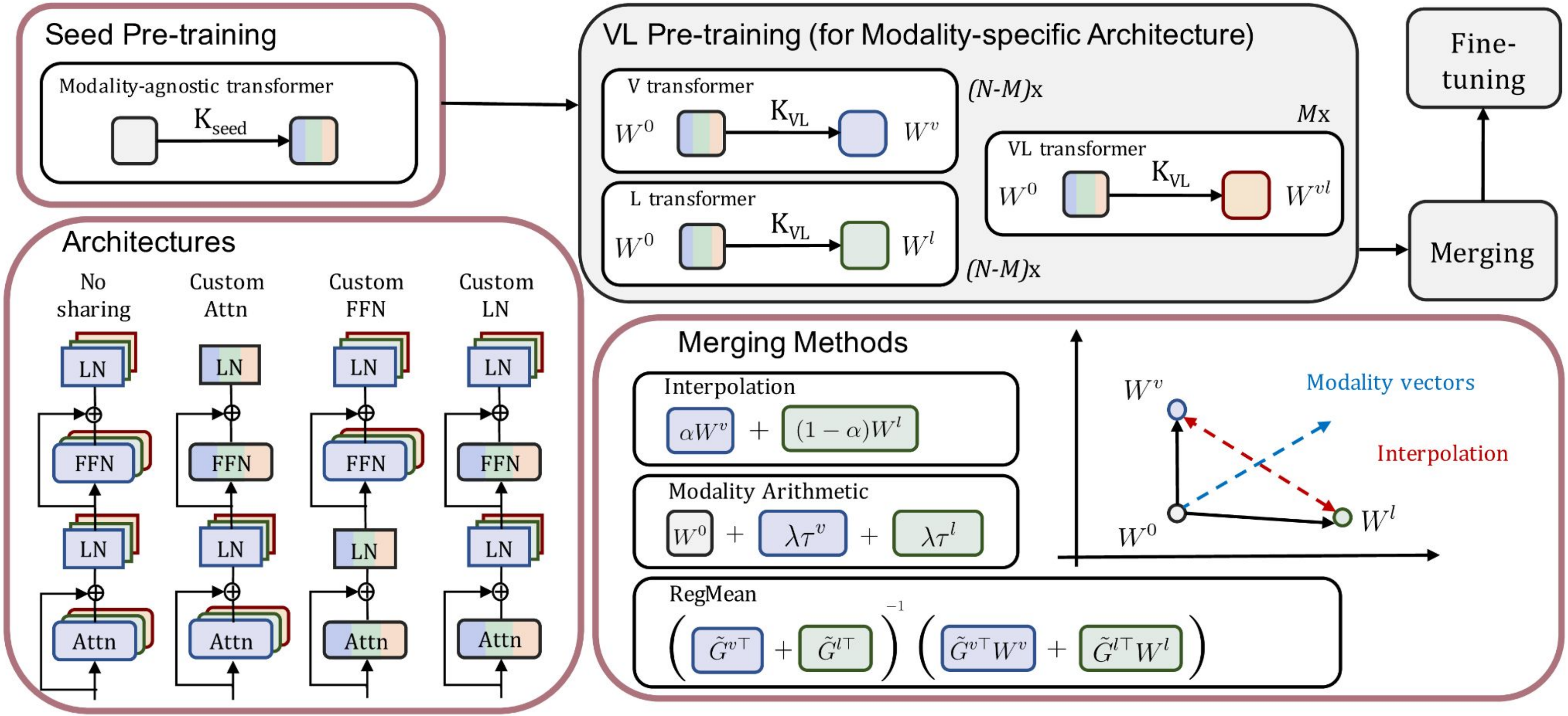
From "Overcoming Catastrophic Forgetting in Zero-Shot Cross-Lingual Generation" by Vu et al.

Editing models with task vectors



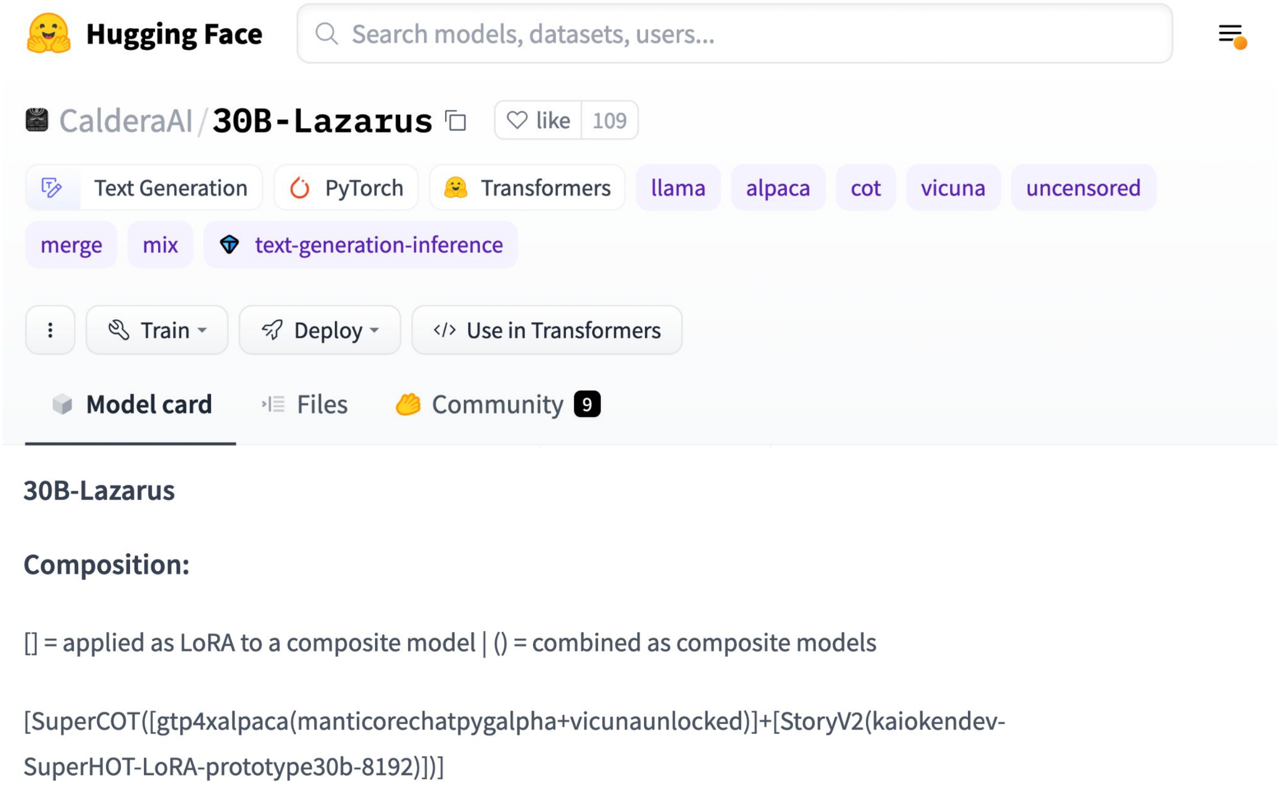
From "Editing Models with Task Arithmetic" by Ilharco et al.

Merging can create multimodal models from unimodal models



From "An Empirical Study of Multimodal Model Merging" by Sung et al.

Recent community-developed models are built via merging



From https://huggingface.co/CalderaAI/30B-Lazarus

Model merging as an optimization problem

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

From "Merging Models with Fisher-Weighted Averaging" by Matena et al.

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

Hyperparameter controlling the importance of model i

From “Merging Models with Fisher-Weighted Averaging" by Matena et al.

Fisher merging uses the Laplace approximation

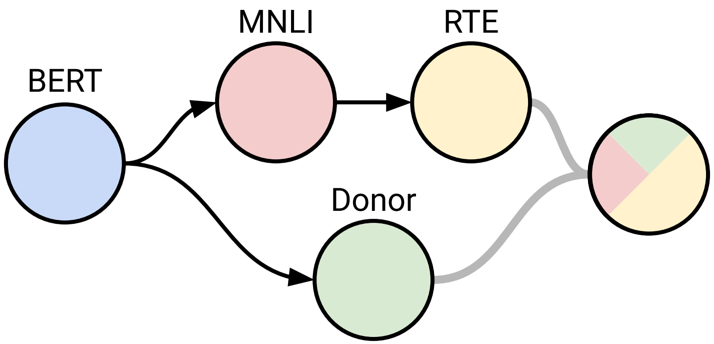
$$\arg\max_{\theta} \sum_{i=1}^{M} \lambda_i \log p(\theta|\mathcal{D}_i)$$

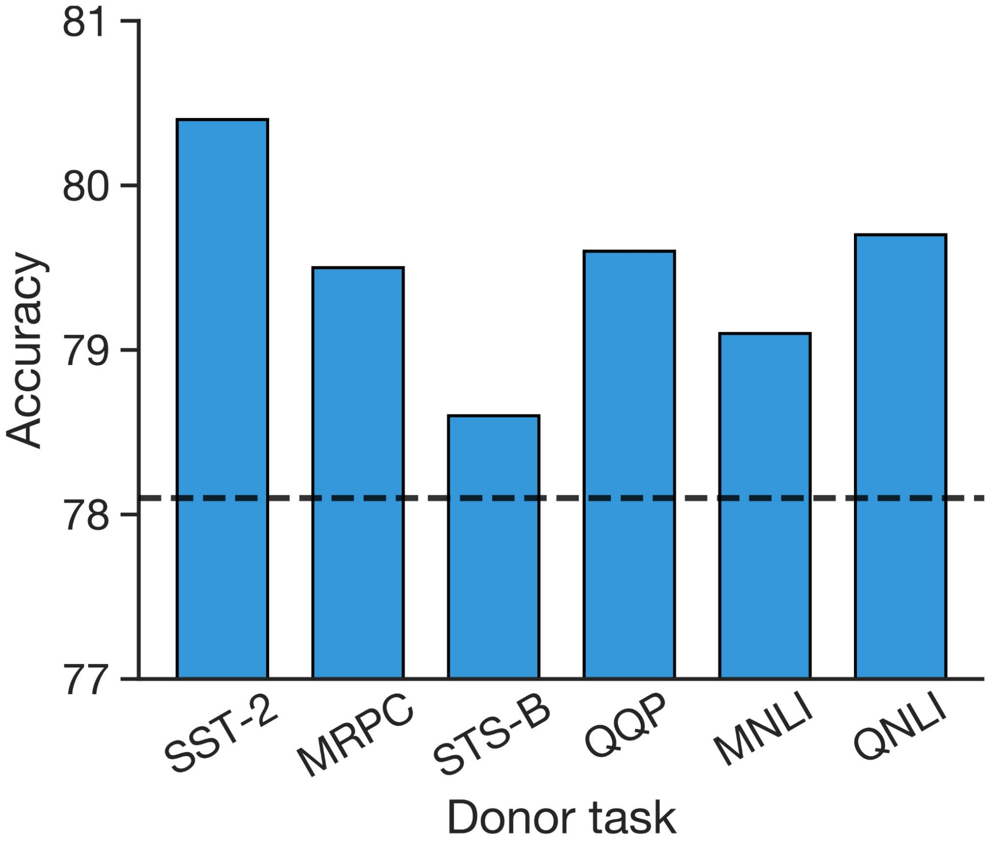
$$\Big\downarrow \; \theta \sim \mathcal{N}(\theta_i, \hat{F}_i^{-1})$$

$$\theta^* = \frac{\sum_i \lambda_i \hat{F}_i \odot \theta_i}{\sum_i \lambda_i \hat{F}_i}$$

From "Merging Models with Fisher-Weighted Averaging" by Matena et al.

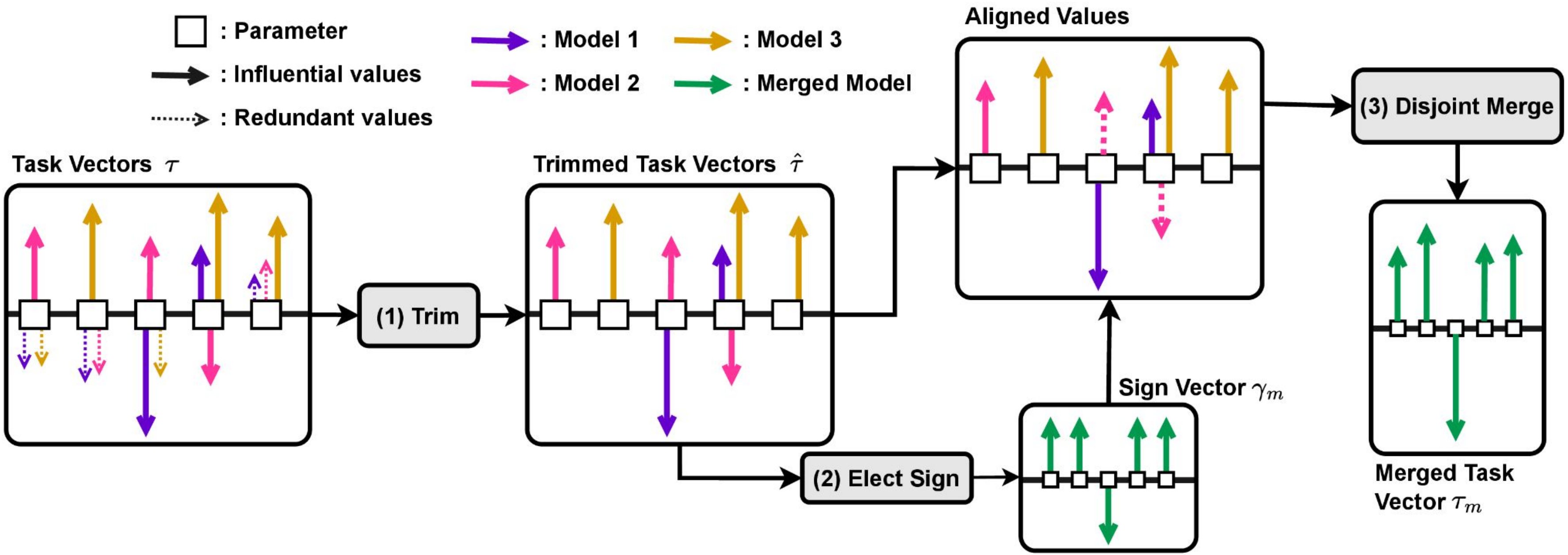
Fisher merging can combine the capabilities of different models





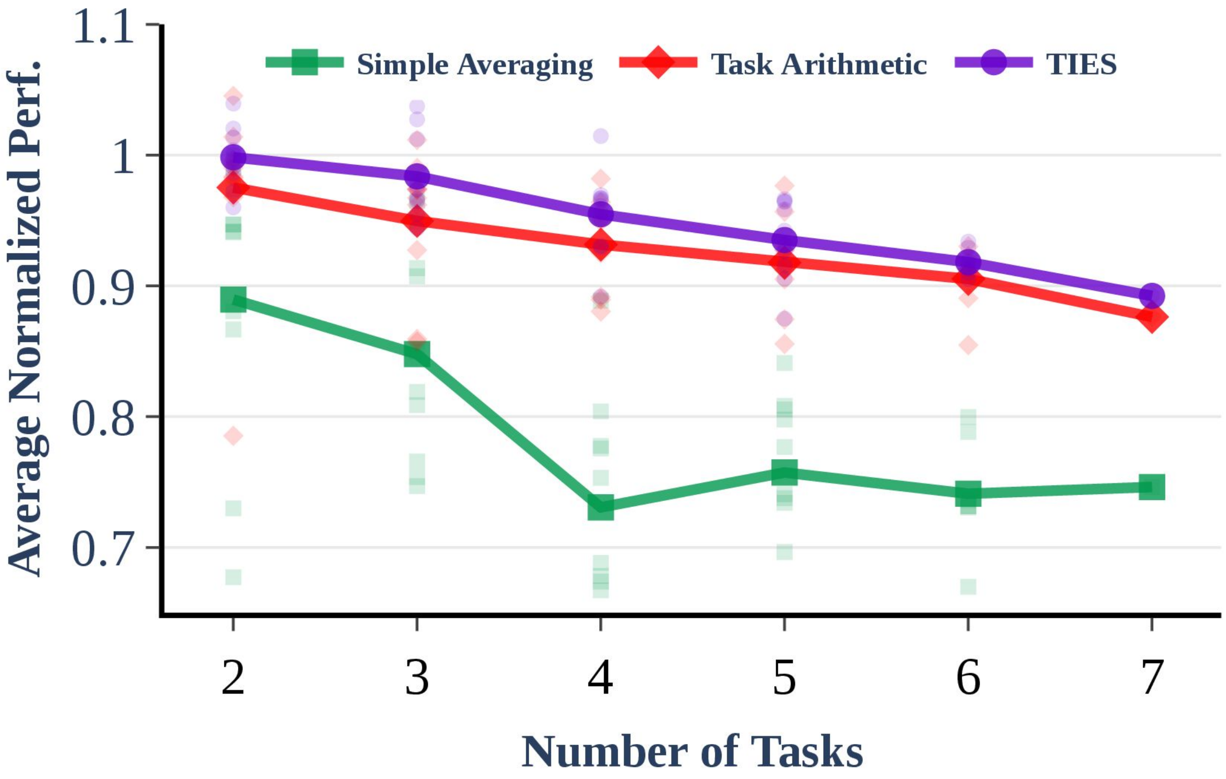
From "Merging Models with Fisher-Weighted Averaging" by Matena et al.

TIES Merging resolves interference when merging models



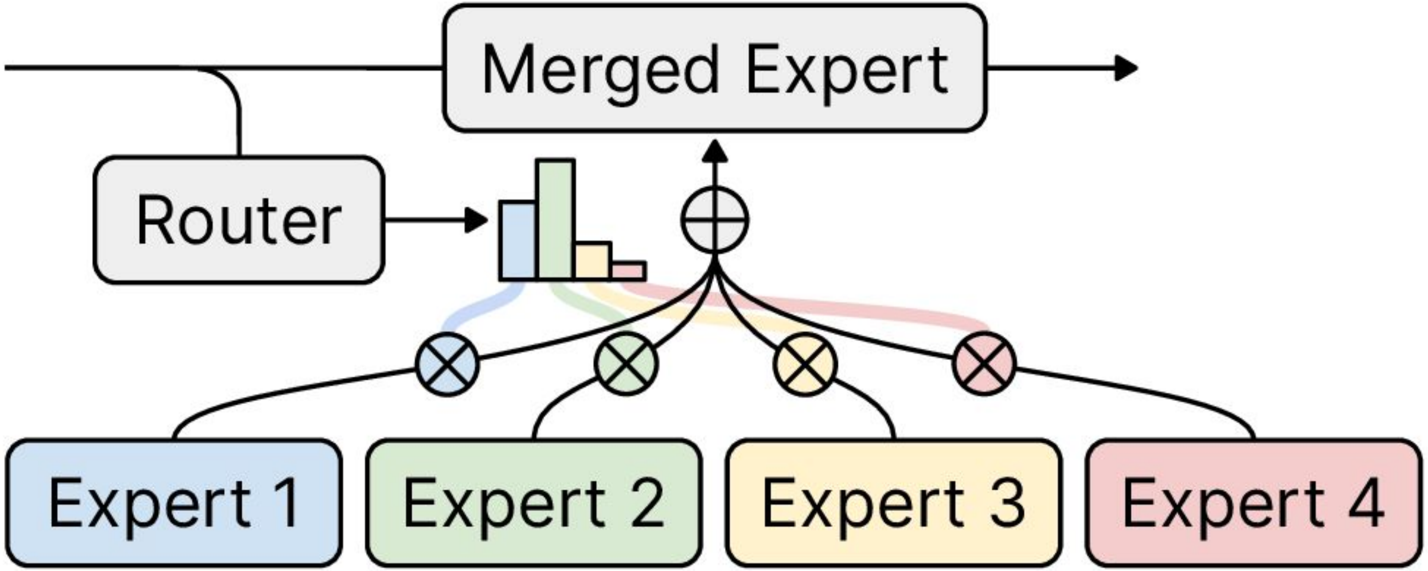
From "Resolving Interference When Merging Models" by Yadav et al.

TIES helps retain specialist model performance



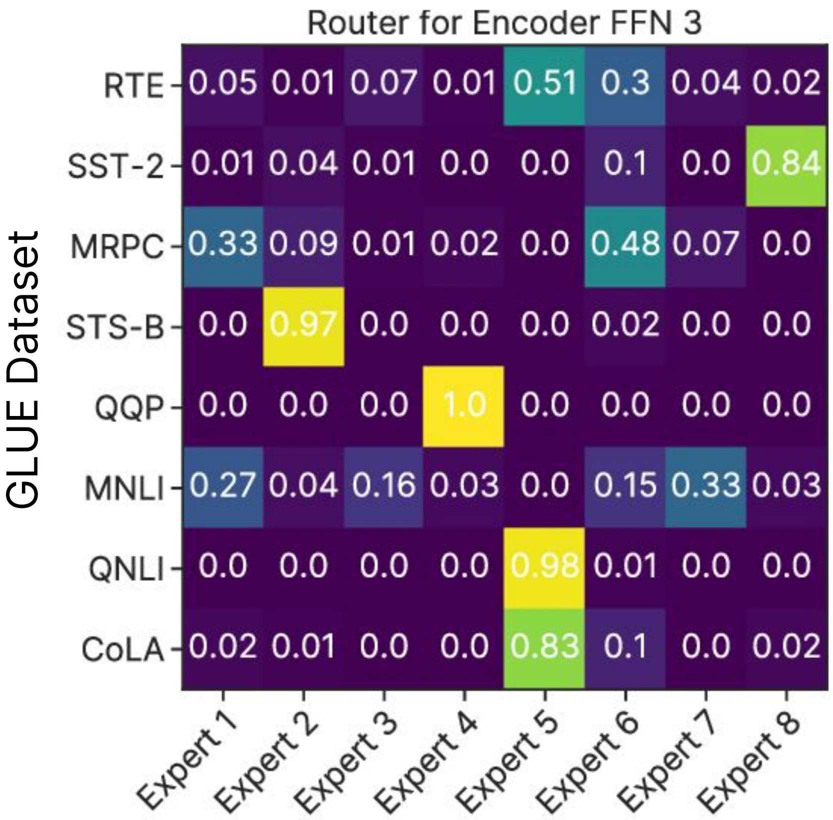
From “Resolving Interference When Merging Models" by Yadav et al.

Differentiable routing between specialist submodels with SMEAR



From "Soft Merging of Experts with Adaptive Routing" by Muqeeth et al.

Experts specialize and are shared across different tasks



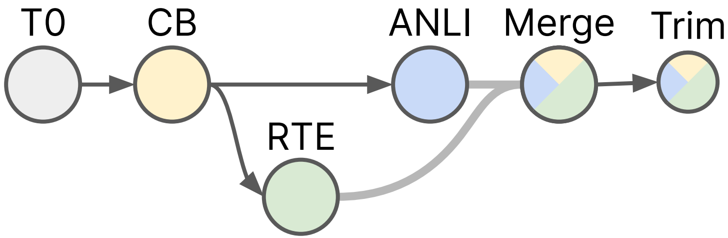
From "Soft Merging of Experts with Adaptive Routing" by Muqeeth et al.

How and why should we build ecosystems of specialist models instead of monolithic models?

An ecosystem can be built and used collaboratively with the right **systems**.

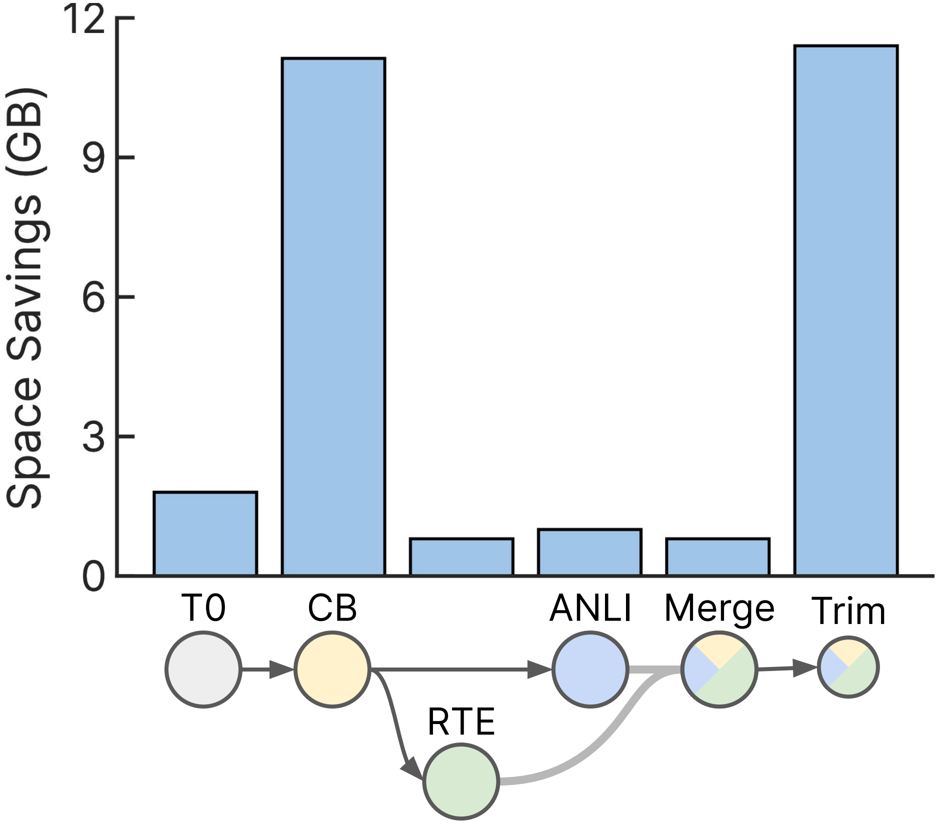
git-theta tracks, merges, and updates models using the git workflow

```
$ 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_them: 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"
```



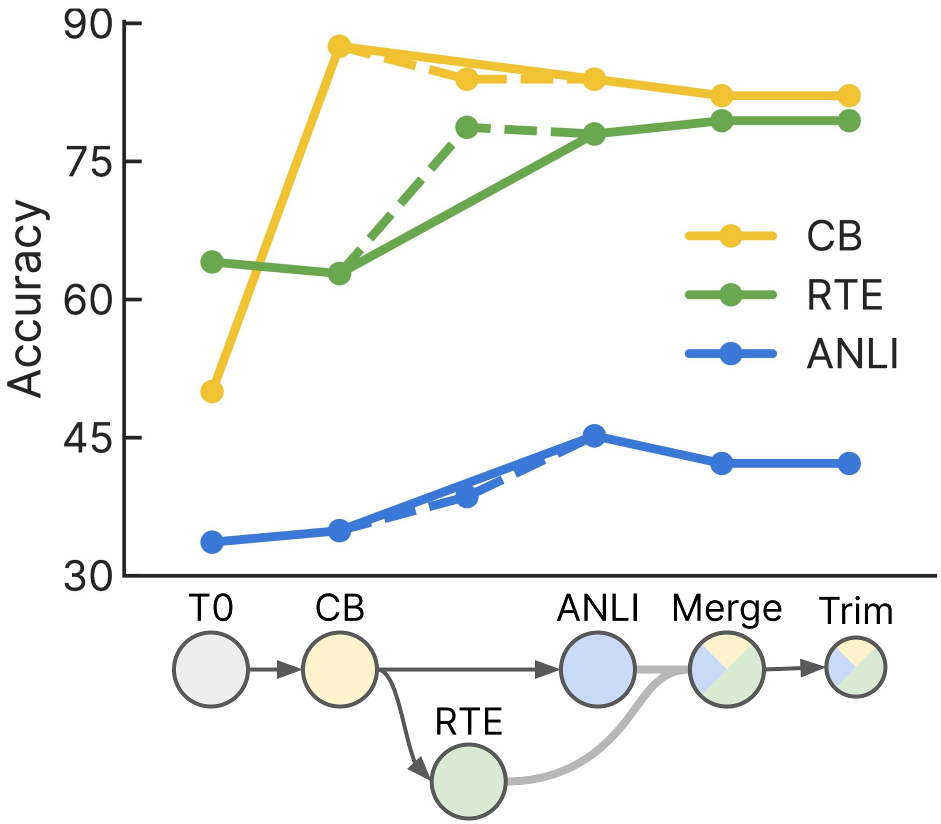
From "Git-Theta: A Git Extension for Collaborative Development of Machine Learning Models" by Kandpal et al.

Communication-efficient updates result in significant space savings



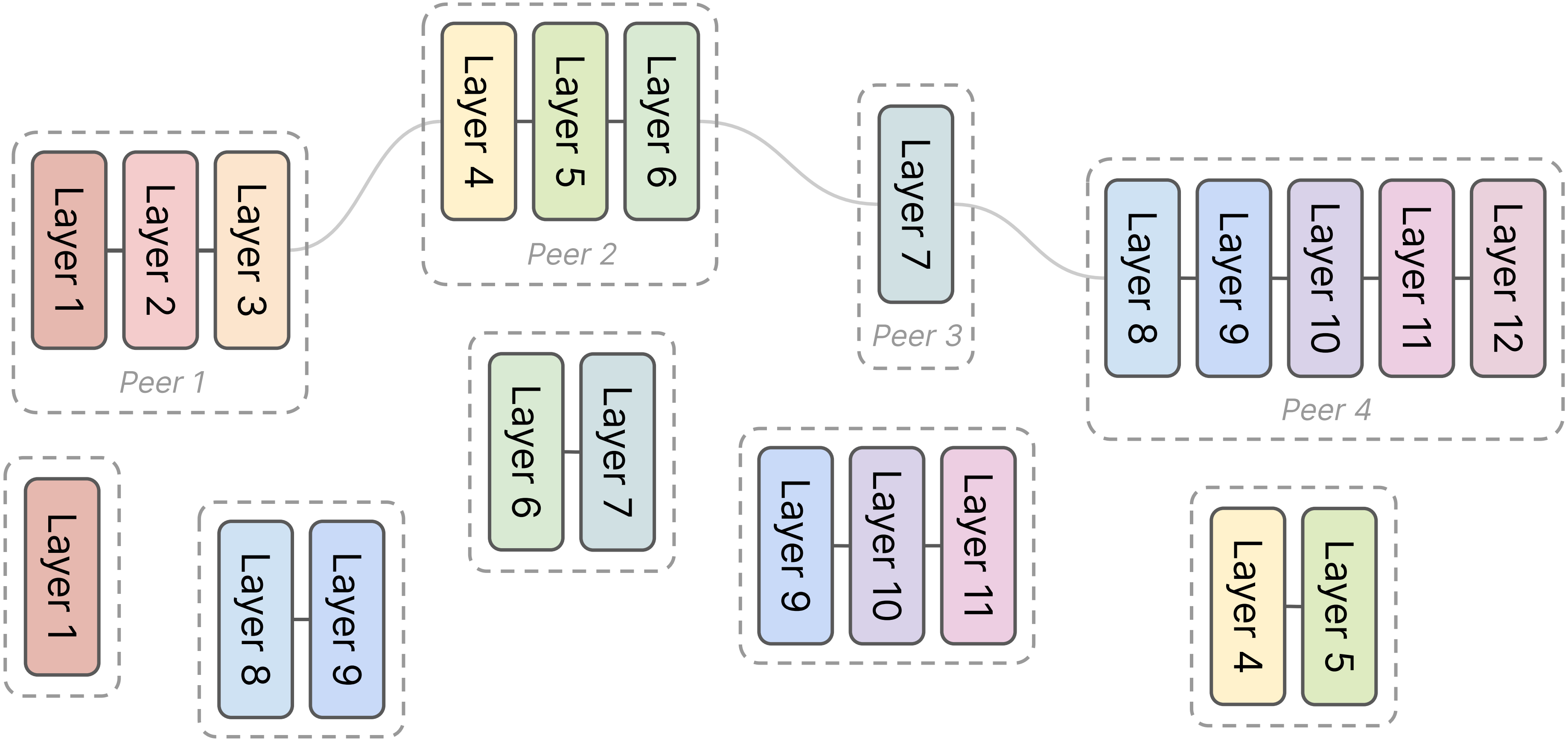
From "Git-Theta: A Git Extension for Collaborative Development of Machine Learning Models" by Kandpal et al.

git-theta allows for continuous and collaborative model development



From "Git-Theta: A Git Extension for Collaborative Development of Machine Learning Models" by Kandpal et al.

Petals enables distributed inference and fine-tuning over the internet



From "Petals: Collaborative Inference and Fine-tuning of Large Models" by Borzunov et al.

Current Petals swarm status

Petals Health Monitor

Bootstrap peers: ●●

Model stabilityai/StableBeluga2 (healthy):

Server ID »	Contributor ?	Version	Throughput »	Precision ?	Adapters ?	Cache ?	Avl. ?	Pings ?	Served blocks
...vYA3Rn		2.0.1.post2	1333 tok/s	bf16 (nf4)		32768	Direct	Show	0:40
...TfceK7		2.0.1.post2	1333 tok/s	bf16 (nf4)		32768	Direct	Show	40:80
...uF3WXf	FYY	2.0.1.post2	1015 tok/s	bf16 (nf4)		24576	Direct	Show	57:76
...sQzHZf	FYY	2.0.1.post1	905 tok/s	f16 (nf4)		24576	Direct	Show	0:19
...yRjqCy	Zetta	2.0.1.post2	1001 tok/s	bf16 (nf4)		30720	Relay	Show	19:38
...m3Vfh7	FYY	2.0.1.post1	324 tok/s	bf16 (nf4)		29044	Direct	Show	69:80
...5D4AAQ		2.0.1.post1	100 tok/s	bf16 (nf4)		32768	Relay	Show	75:80
...dSkeys	FYY	2.0.1.post2	1015 tok/s	bf16 (nf4)		22528	Direct	Show	38:57

Model meta-llama/Llama-2-70b-chat-hf (healthy):

Server ID »	Contributor ?	Version	Throughput »	Precision ?	Adapters ?	Cache ?	Avl. ?	Pings ?	Served blocks
...kDJVjh	jobs.trelent.com	2.0.1	8107 tok/s	f16 (nf4)		32768	Relay	Show	20:46
...bsDGGc		2.0.1.post2	17 tok/s	f16 (nf4)		32768	Relay	Show	0:3
...MWaAxr	nora	2.0.1.post1	670 tok/s	f16 (nf4)		12288	Direct	Show	46:66
...rgNAo9	nora	2.0.1.post1	670 tok/s	f16 (nf4)		12288	Direct	Show	60:80
...RPFSct	nora	2.0.1.post1	670 tok/s	f16 (nf4)		12288	Direct	Show	0:20

From https://health.petals.dev/

Thanks.

Please give me feedback:

http://bit.ly/colin-talk-feedback

craffel@gmail.com