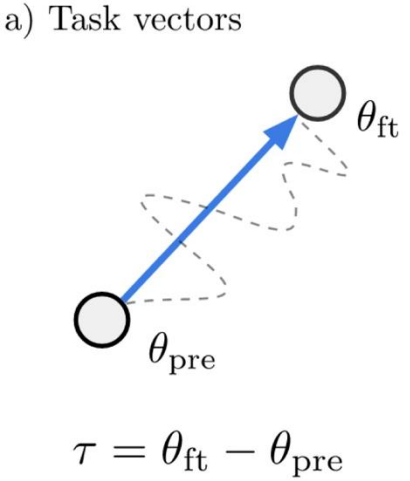
MergeBench: A Benchmark for Merging Domain-Specialized LLMs

Yifei He, Siqi Zeng, Yuzheng Hu, Rui Yang, Tong Zhang, Han Zhao

University of Illinois Urbana-Champaign

Background



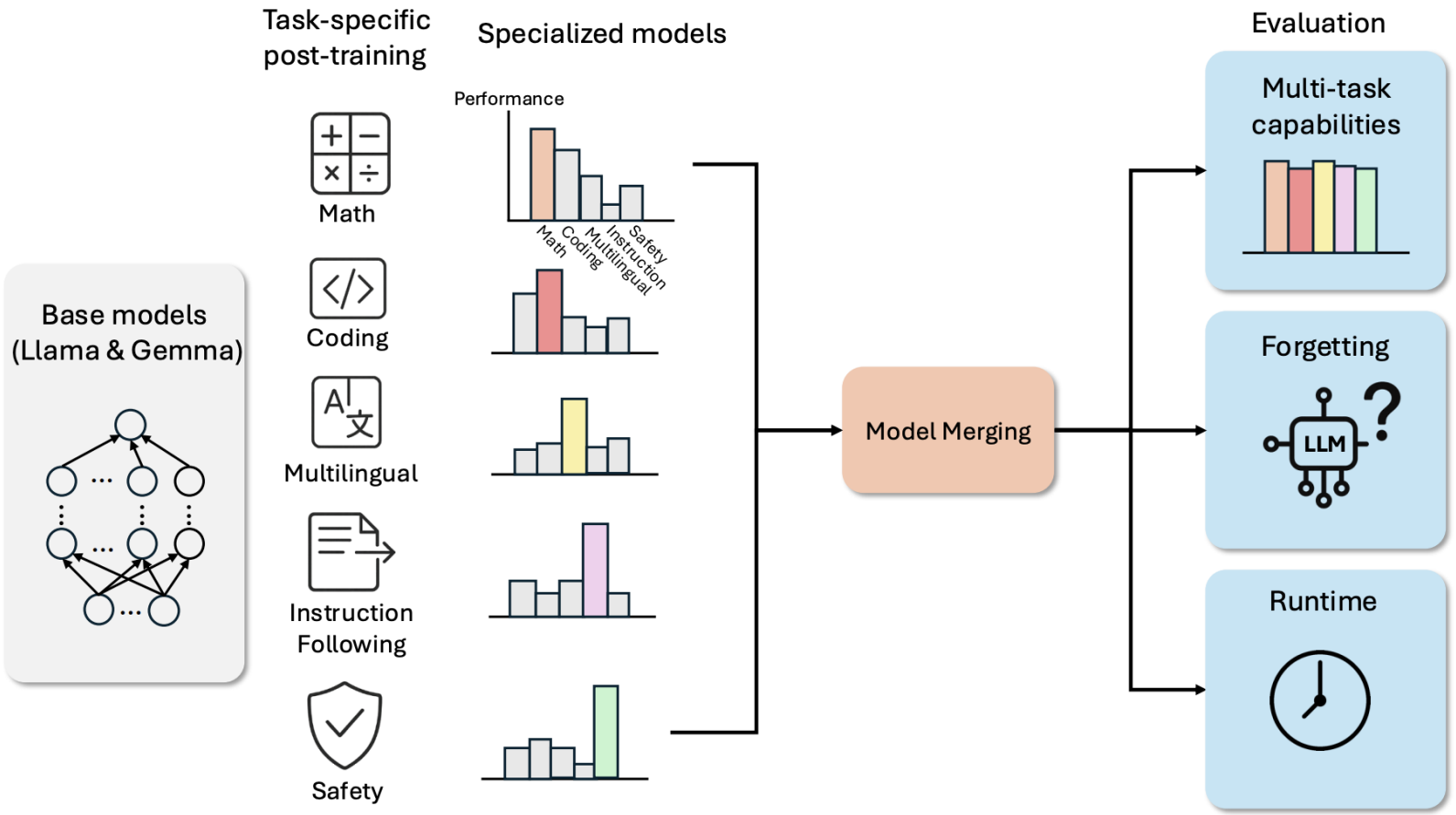
- Pretrained model weight: $\theta_{pre} \in R^d$
- Finetuned model weight: $\theta_{ft}^{(t)} \in R^d$, where t is the task index
- Task vector: $\tau_t = \theta_{ft}^{(t)} - \theta_{pre}$
- Task vectors define a direction in the weight space, along which the task performance can be improved
- Task arithmetic: Update model θ via element-wise addition of task vectors
  - $\theta_{MTL} = \theta_{pre} + \sum_t \lambda_t \tau_t$, where λ_t is an optional scaling term

Criteria

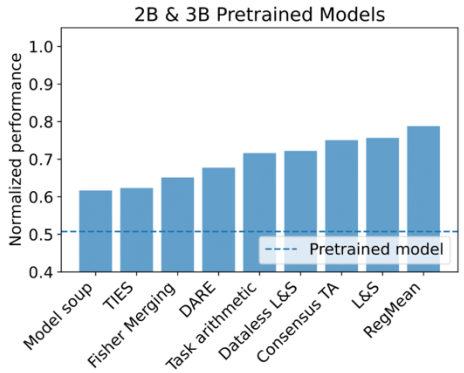
- **Diverse model**: evaluates models from different model families.
- **Large model**: includes models larger than 7B.
- **Domain task**: focuses on real-world, general-domain tasks beyond conventional NLP tasks.
- **Gradient-based methods**: supports merging methods requiring gradient information or training statistics.
- **Open-source**: provides open access to both evaluation pipelines and constituent specialized models.

Evaluation	Diverse model	Large model	Domain task	Gradient-based methods	Open-source
FusionBench [63]	✗	✗	✗	✓	✓
Compositional eval [61]	✗	✗	✗	✓	✓
Merging at scale [76]	✗	✓	✗	✗	✗
Model-GLUE [87]	✗	✓	✓	✗	✓
MergeBench	✓	✓	✓	✓	✓

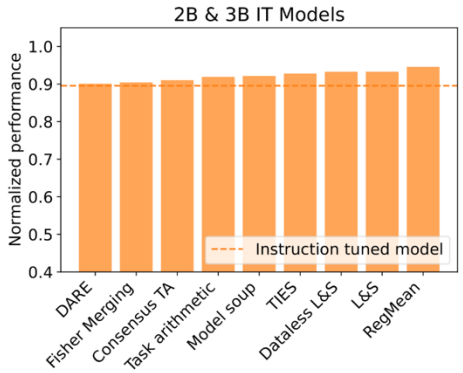
Overview


Base models
(Llama & Gemma)
Task-specific
post-training
Math
Coding
Multilingual
Instruction
Following
Safety
Specialized models
Performance
Math
Coding
Multilingual
Instruction
Safety
Model Merging
Evaluation
Multi-task
capabilities
Forgetting
LLM
Runtime

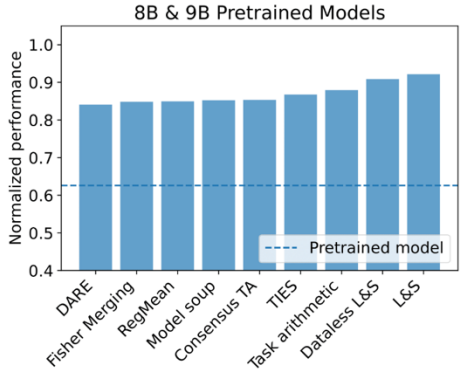
Main results



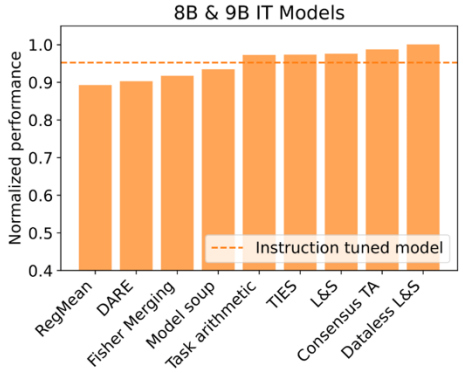
Performance on 2B Pretrained Models



Performance on 2B Instruction-Tuned Models



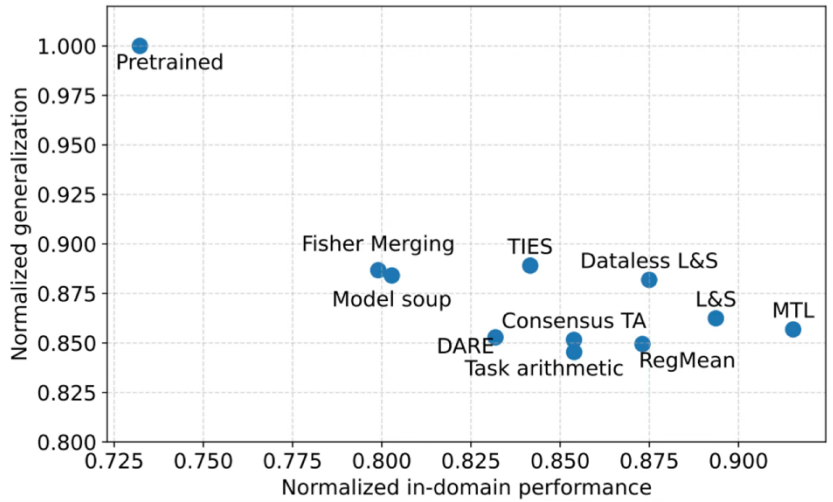
Performance on 8B Pretrained Models

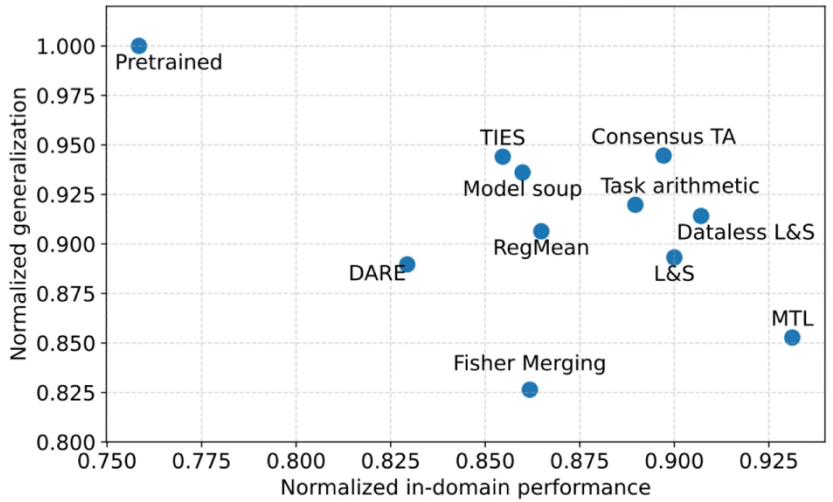


Performance on 8B Instruction-Tuned Models

- Model merging is more effective on stronger base models.
  - Larger models merge better: More capacity, less conflicts
  - IT models merge better: Better performance, longer shared trajectories
- The two Localize-and-Stitch variants consistently show superior performance.

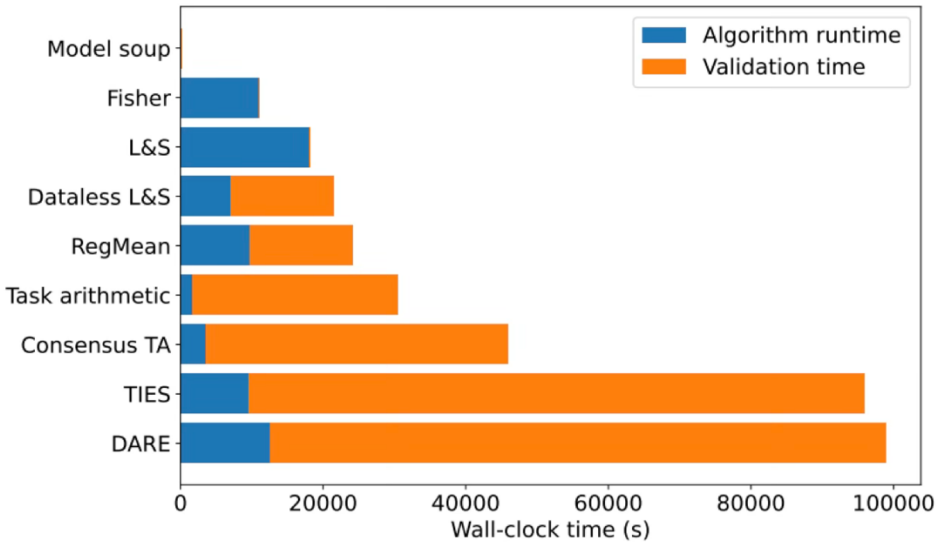
Forgetting

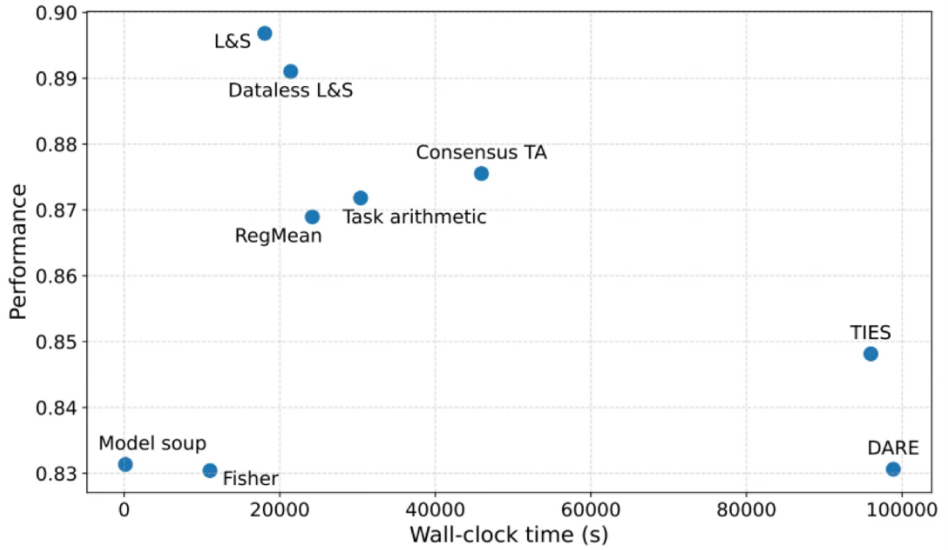




- Multi-task learning (MTL) models perform well on in-domain tasks but often sacrifice generalization to unseen domains.
- Merged models better retain base model knowledge.

Runtime





- Validation cost often dominate algorithm runtime
- Localize-and-Stitch, RegMean and Task arithmetic shows favorable trade-off between performance and efficiency

Future directions

- Opportunities for improving merging efficiency
  - Can we gain more insights on the hyperparameters?
- Mix data or merge models?
  - May be particularly useful for imbalanced or low-resource settings
- Positioning model merging in LLM pipelines
  - Llama-3 uses model soup to average models with different hyperparameters
  - Command A combines separately trained specialized models
  - UI-TARS-2 combines separately trained models on different platforms

Thank you!

Please join us at Exhibit Hall C,D,E on Dec 5 (Fri) 11 AM -2 PM!