Exploring *Model Kinship* for *Merging* Large Language Models

Yedi Hu[1], Yunzhi Yao[1], Ningyu Zhang[1], Shumin Deng[2], Huajun Chen[1]

[1]Zhejiang University [2]National University of Singapore

Project Paper

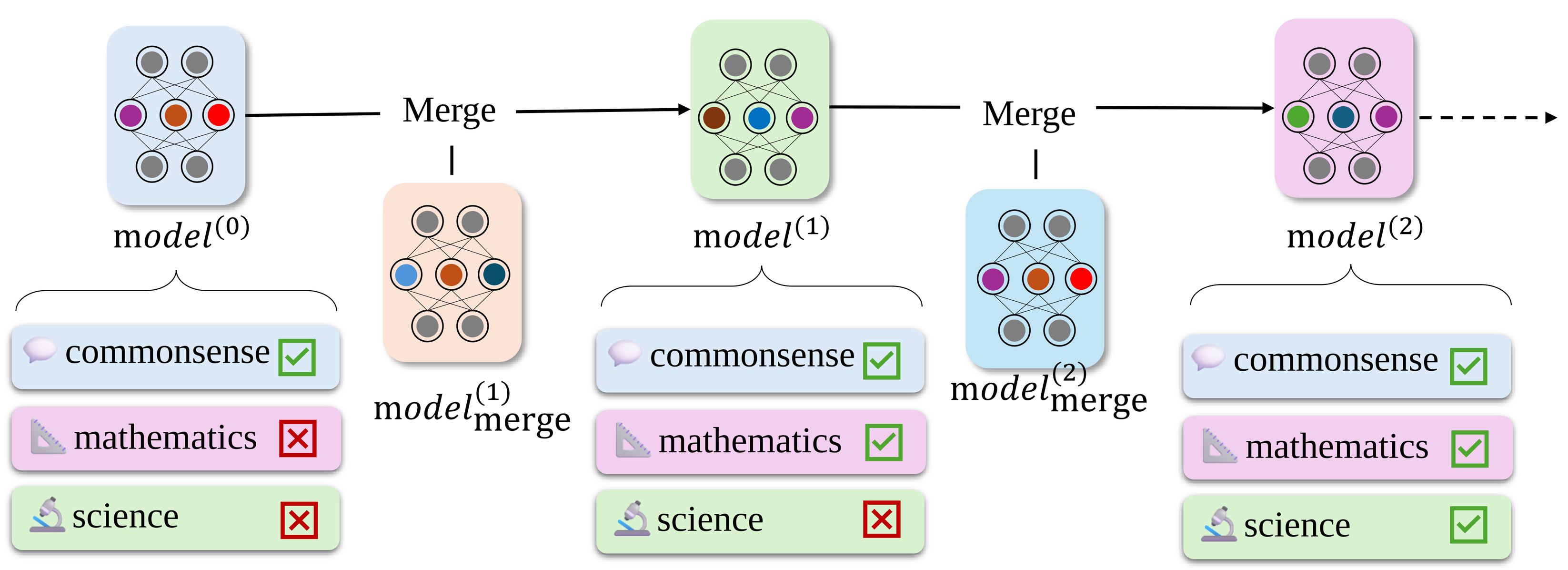
Conceptual Motivation

What we know:

- **LLMs** need require **frequent and continuous updates** to keep up with growing user **demands across diverse tasks**.
- **Model Merging** has emerged as a **cost-effective** approach to combine the strengths of different models **without altering their underlying architecture**.

Model Evolution:

- LLMs can evolve to integrate new capabilities by **iteratively merging with one another**.



$$\mathrm{E}\left(\theta^{(t)} \mid \theta^{(t-1)}, \theta^{(t-2)}, \dots, \theta^{(0)}\right) = \mathrm{E}\left(\theta^{(t)} \mid \theta^{(t-1)}, \theta^{(t)}_{merge}\right)$$

Question: How can we systematically select models for merging to achieve greater performance improvements?

Key Contributions

- **Model Kinship:** Introduce a kinship metric to assess LLM similarity during merging, guiding auto-merging strategies.
- **Empirical Analysis**: Evaluate iterative merging dynamics, highlighting multitask performance gains and stagnation, with insights from model kinship.
- **Practical Strategies:** Propose Top-k Greedy Merging with model kinship to efficiently tackle merging optimization challenges.

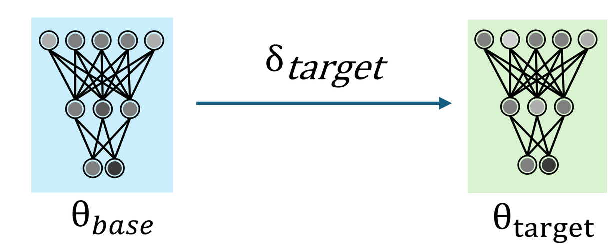
Model Kinship

We need **an efficient metric** to evaluate the **overall capability difference** between models, without relying on a dataset for evaluation, **to estimate the outcome of future merges (high or low)**.

- Extract **delta parameter δ_{target}** of the target models.
- Calculate the **kinship** between **delta parameters**.

$$\delta_{target} = \theta_{target} - \theta_{base}$$

$$k = sim(\delta_1, \delta_2)$$

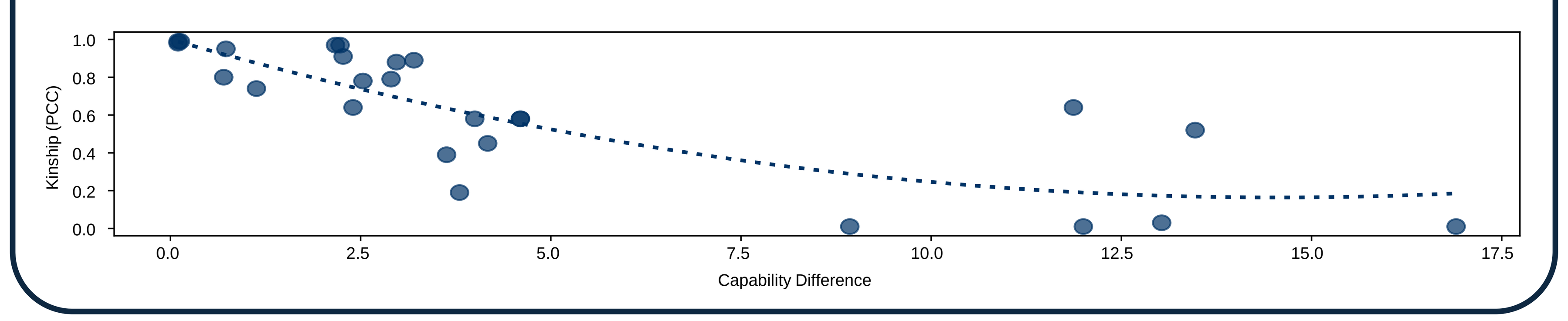


Parameter changes introduced through fine-tuning/iterative merging processes.

e.g. adopt ***Pearson Correlation Coefficient*** as similarity metric:

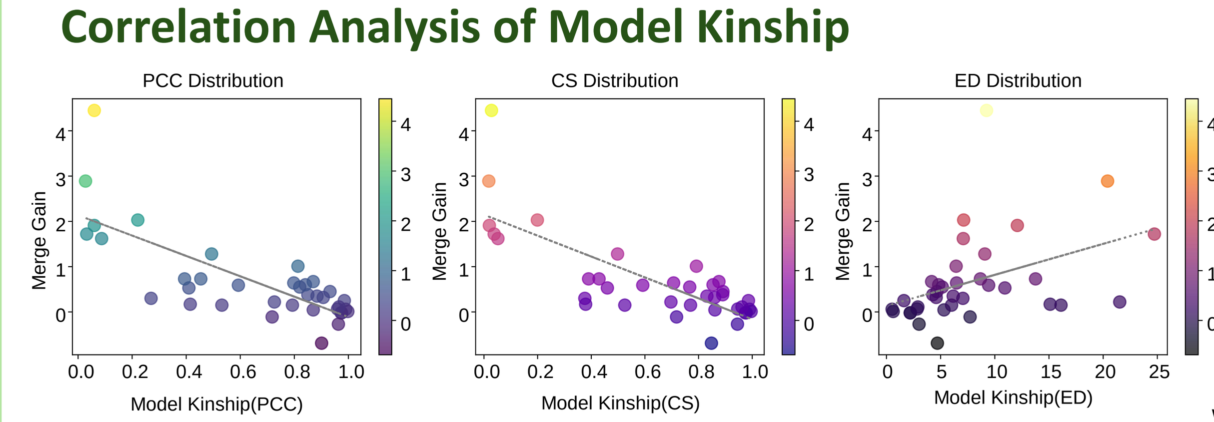
$$sim(\delta_i, \delta_j) = \frac{cov(\delta_i, \delta_j)}{\sigma_{\delta_i}\sigma_{\delta_j}} = \frac{\sum(\delta_i - \bar{\delta}_i)(\delta_j - \bar{\delta}_j)}{\sqrt{\sum(\delta_i - \bar{\delta}_i)^2}\sqrt{\sum(\delta_j - \bar{\delta}_j)^2}}$$

- **Correlation** between task capability difference and kinship

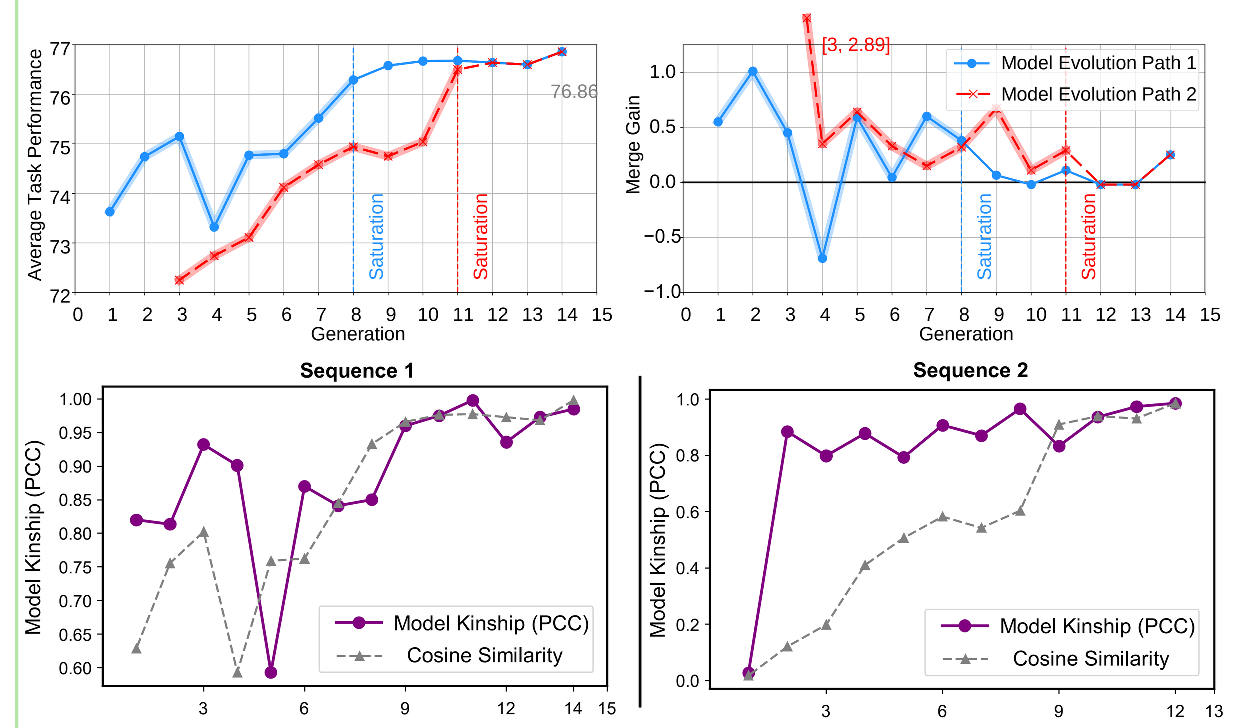


Initial Analysis

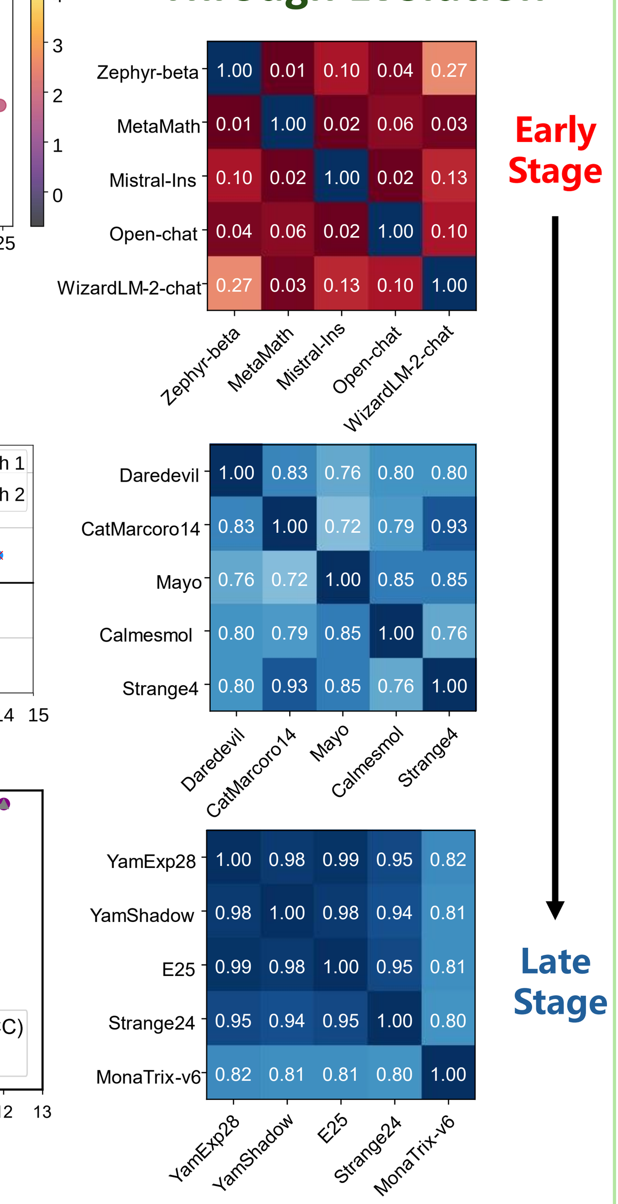
Correlation Analysis of Model Kinship



Kinship Change Through Model Evolution Paths



Mutual Kinship of Candidate Models Through Evolution



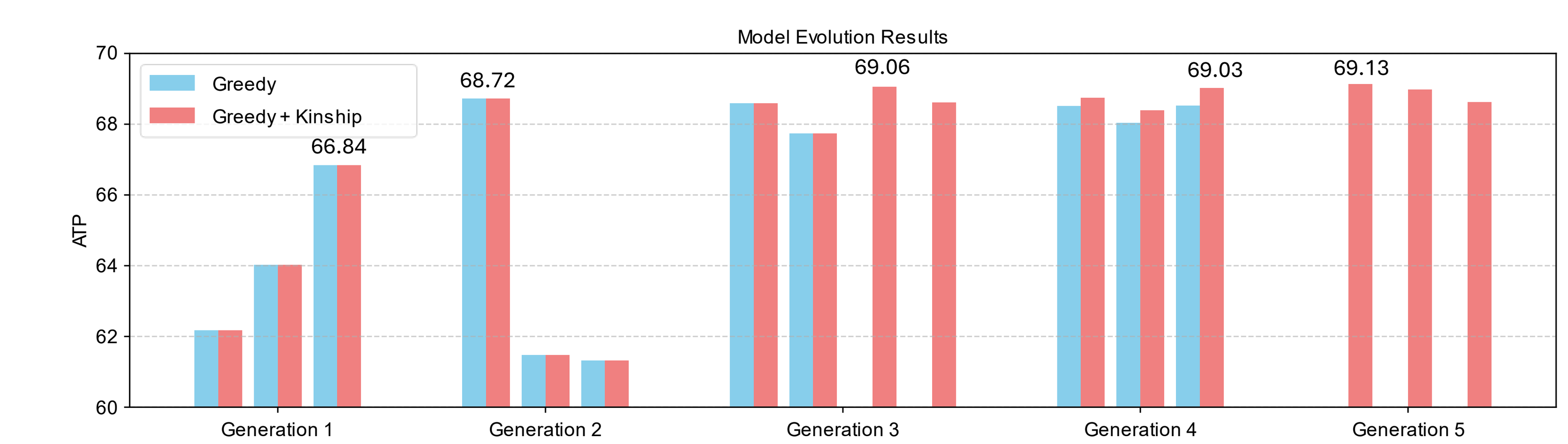
Main Results

- Enhance the greedy merging strategy by **adding additional merges based on low model kinship.**

Table 2: Results of merging experiments comparing the vanilla greedy strategy and our proposed approach. The first three models serve as the foundation models in both experiments.**Note**: The model kinship experiment was terminated at generation 5, as it has already outperformed the greedy strategy by that point.

Greedy Strategy				+ Model Kinship			
Model	**Avg.**	**Gain**	**Kinship**	**Model**	**Avg.**	**Gain**	**Kinship**
MetaMath	63.72	/	/	**MetaMath**	63.72	/	/
Instruct	61.82	/	/	**Instruct**	61.82	/	/
Open-chat	66.28	/	/	**Open-chat**	66.28	/	/
model-1-1	62.17	-0.6	0.01	model-1-1	62.17	-0.6	0.01
model-1-2	64.02	-0.03	-0.02	model-1-2	64.02	-0.03	-0.02
model-1-3	66.84	+1.84	0.05	model-1-3	66.84	+1.84	0.05
model-2-1	**68.72**	**+2.16**	**0.93**	model-2-1	68.72	+2.16	0.93
model-2-2	61.47	-3.96	0.57	model-2-2	61.47	-3.96	0.57
model-2-3	61.32	-3.83	0.58	model-2-3	61.32	-3.83	0.58
model-3-1	68.59	+1.09	0.95	model-3-2	67.74	+1.09	0.93
model-3-2	67.74	-0.04	0.93	model-3-3	69.06	+0.74	0.24
	-	-	-	model-3-4	68.61	+1.13	0.32
model-4-1	68.51	-0.14	0.98	model-4-4	68.75	-0.14	0.54
model-4-2	68.04	-0.19	0.98	model-4-5	68.39	-0.27	0.66
model-4-3	68.53	+0.37	0.94	model-4-6	69.03	+0.15	0.52
	-	-	-	**model-5-1**	**69.13**	**+0.04**	**0.65**
	-	-	-	model-5-2	68.98	+0.07	0.65
	-	-	-	model-5-3	68.63	-0.37	0.98

Stop improving



Findings:

- **Merging low-kinship models introduces novel weight variations, enabling continued iterative merging.**
- **Early stopping at high kinship reduces resource consumption while maintaining performance.**

Future works

- Advance **interpretability** by designing more **informative metrics** to explain model merging behavior.
- Explore **foundational theory** for **automated, self-evolving model merging pipelines.**