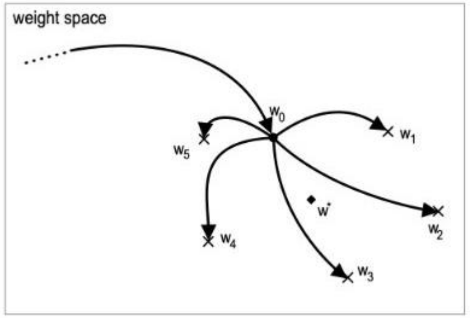
Explore Model Kinship For Merging LLMs

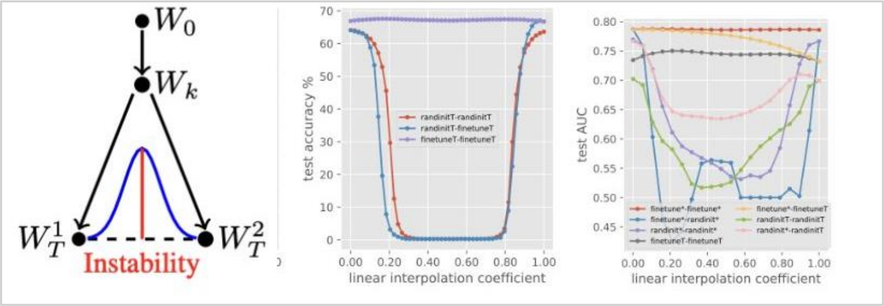
Yedi Hu
Zhejiang University

Overview

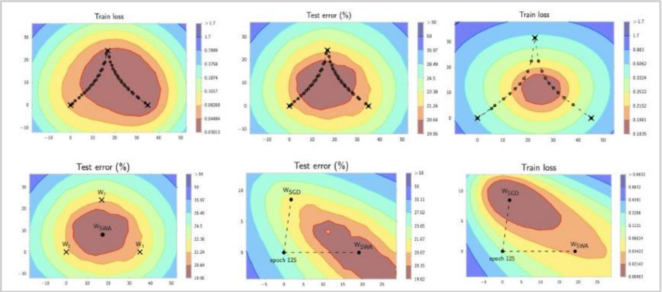
Quick view: Model Merging Research Timeline



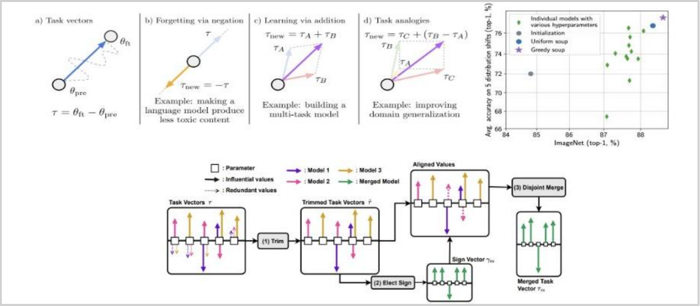
Initial Model Merging



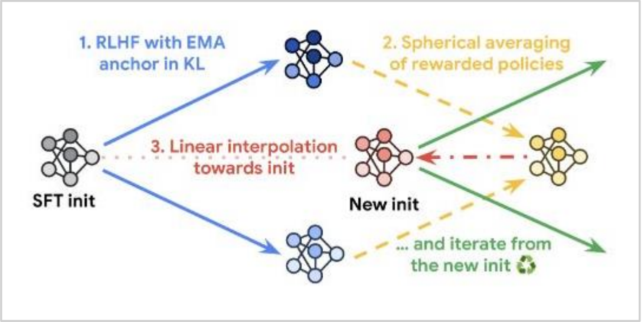
Averaging along the **training trajectory**



Linear mode connectivity



Various merging **approaches**



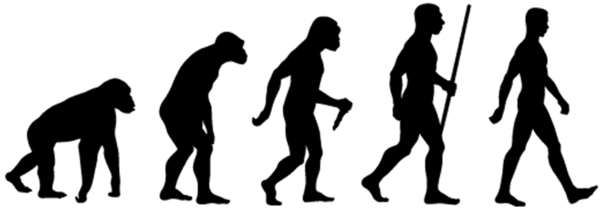
Model Merging in **LLMs**

Modified from https://x.com/cwolferesearch/status/1828567528710513141

Community Experiments

Hugging face Open Leaderboard

T	Model	Average	ARC	HellaSwag	MMLU	TruthfulQA	Winogrande	GSM8K
	automerger/YamshadowExperiment28-7B	76.86	73.29	89.25	64.38	78.53	85.24	70.51
	liminerity/M7-7b	76.82	72.87	89.15	64.5	77.93	84.77	71.72
	allknowingroger/MultiverseEx26-7B-slerp	76.8	72.95	89.17	64.36	78.12	85.16	71.04
	Kukedlc/NeuralSynthesis-7B-v0.1	76.8	73.04	89.18	64.37	78.15	85.24	70.81
	AurelPx/Percival_01-7b-slerp	76.79	73.21	89.16	64.42	77.97	85.08	70.89
	shyamieee/J4RVIZ-v6.0	76.78	73.29	89.15	64.41	77.87	85	70.96
	automerger/Ognoexperiment27Multi_verse_model-7B	76.77	72.95	89.29	64.39	78.04	84.85	71.11
	shyamieee/B3E3-SLM-7b-v3.0	76.76	73.04	89.14	64.48	78.2	85	70.74
	Kukedlc/NeuralSynthesis-7b-v0.4-slerp	76.76	73.21	89.14	64.28	78.07	84.85	71.04
	BarraHome/Mistroll-7B-v2.2	76.76	72.78	89.16	64.35	78.1	85	71.19
	nlpguy/T3QM7	76.75	73.12	89.14	64.48	77.96	85.08	70.74
…								
	Walmart-the-bag/openchat-3.5-Infinity	67.95	62.63	84.05	64.65	51.99	80.11	64.29
	Isaak-Carter/J.O.S.I.E.3-Beta10-7B-slerp	67.95	63.48	83.79	62.88	56.88	79.64	61.03
	jondurbin/bagel-dpo-7b-v0.1	67.95	66.72	84.16	64.24	64.05	80.9	47.61
	P0x0/IceMerge-7b-32k	67.94	65.53	85.65	64.66	53.09	80.51	58.23
	vicgalle/SystemHermes-2-7B	67.92	65.02	84.05	63.16	56.42	77.35	61.56
	allknowingroger/DolphinChat-7B-slerp	67.92	64.59	84.21	64.23	50.86	81.37	62.24
	Liangmingxin/ThetaWave-7B-sft	67.92	63.14	84.42	63.78	59.74	79.64	56.79
	ichigoberry/pandafish-7b	67.88	65.19	85.28	64.99	52.69	80.82	58.3
	Weyaxi/Einstein-openchat-7B	67.87	65.1	83.57	64.01	54.51	79.16	60.88
	openagi-project/OpenAGI-7B-v0.1	67.87	68.26	85.06	61.6	59.4	79.79	53.07
	indischepartij/MiaLatte-Indo-Mistral-7b	67.86	66.55	85.23	63.93	56.04	80.35	55.04
	nlpguy/Hermes-low-tune-2	67.85	65.27	84.41	63.63	53.12	78.22	62.47



Model merging toolkits such as **Mergekit** enable users with limited expertise to easily conduct merging experiments, **leading to an evolution of LLMs for the community.**

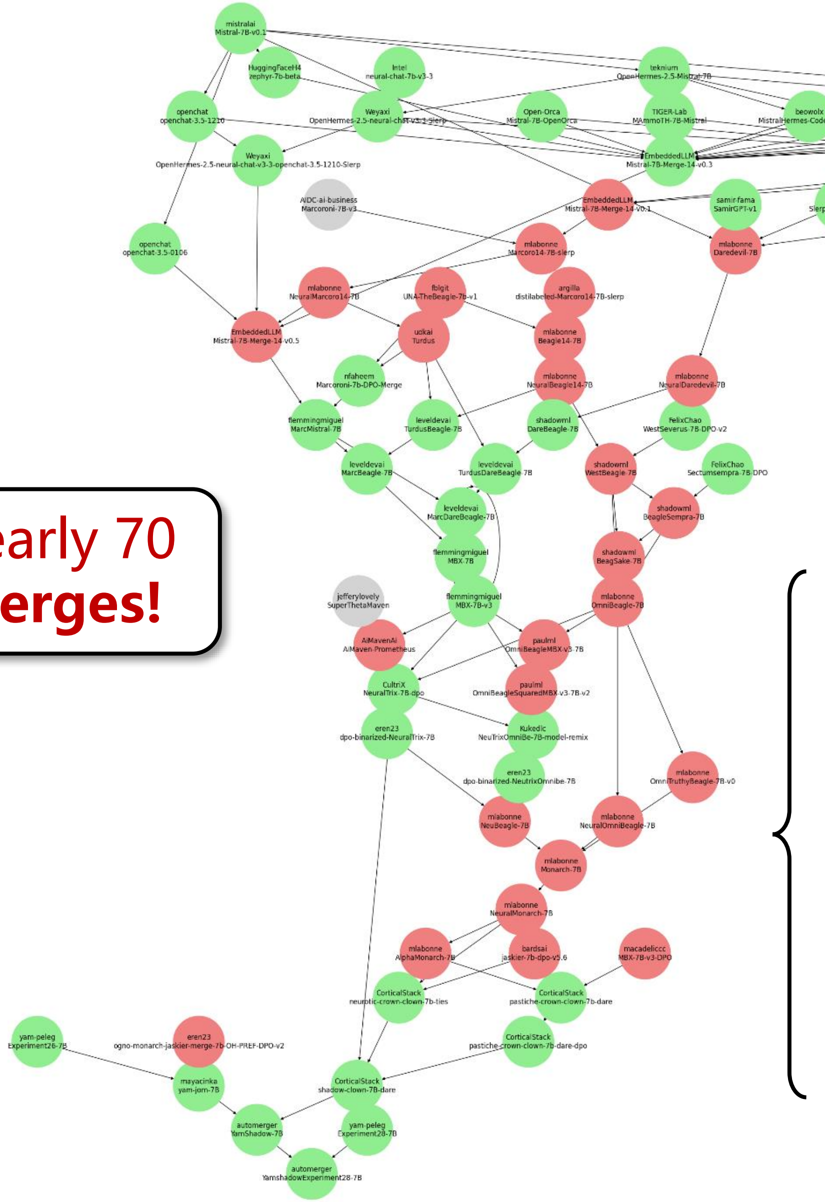
Just for **Mistral-7B**

More than 500 merged models

https://huggingface.co/spaces/open-llm-leaderboard-old/open_llm_leaderboard

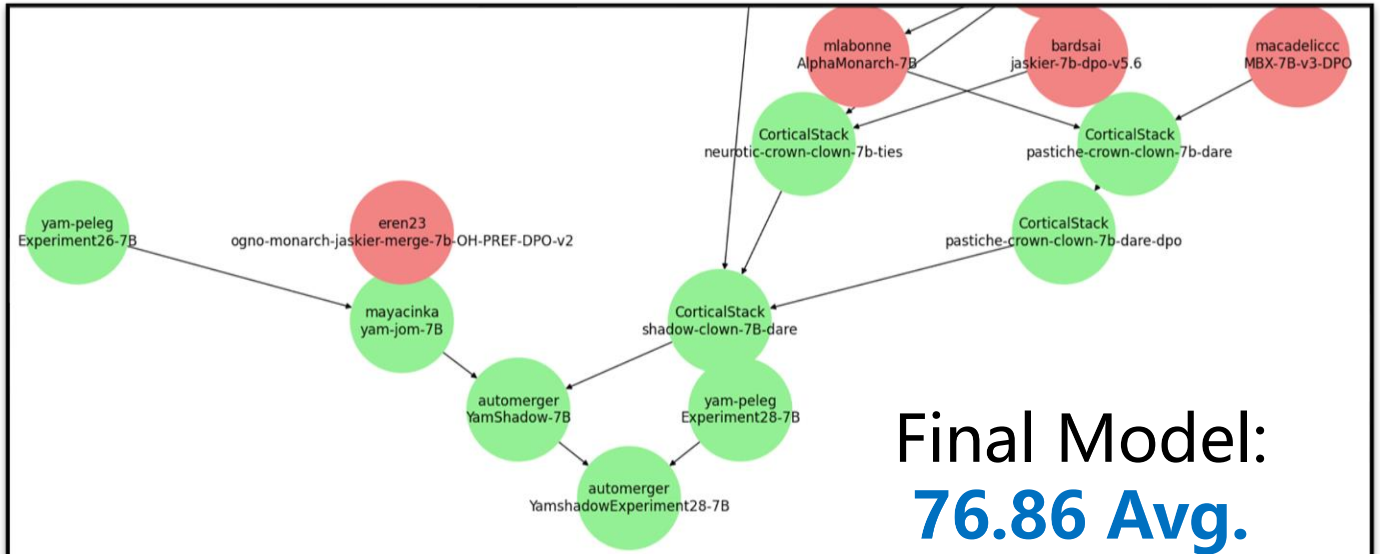
Model Evolution

Foundation Models:
~ 67 Avg.



Model Family Tree of
Model Evolution

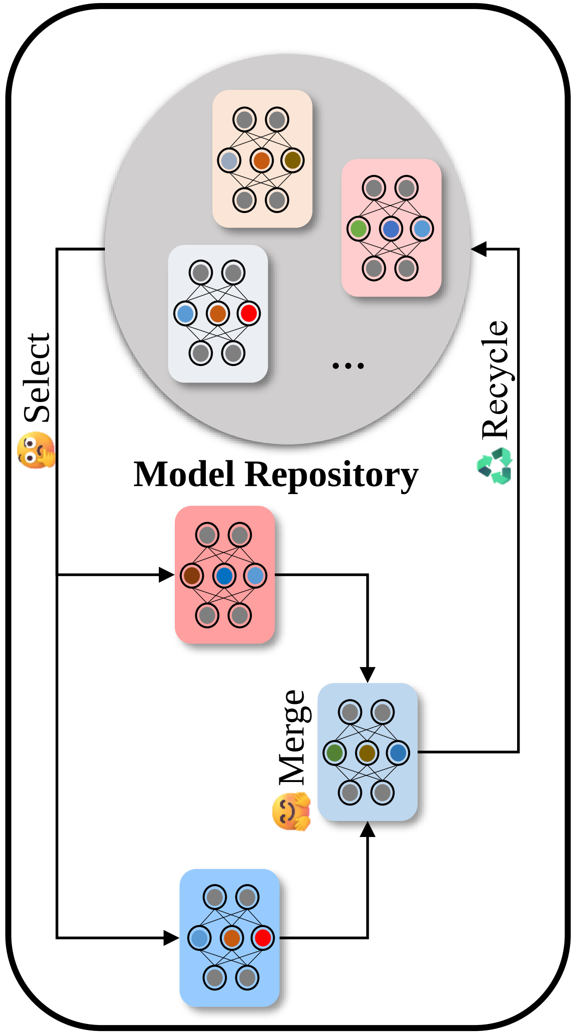
Nearly 70
merges!



Final Model:
76.86 Avg.

How is **Model Evolution** achieved ?

Iterative Merging Improving LLMs via iterative process



Select

Select potential models for the next merge using a specific strategy (e.g. **performance-prior**)

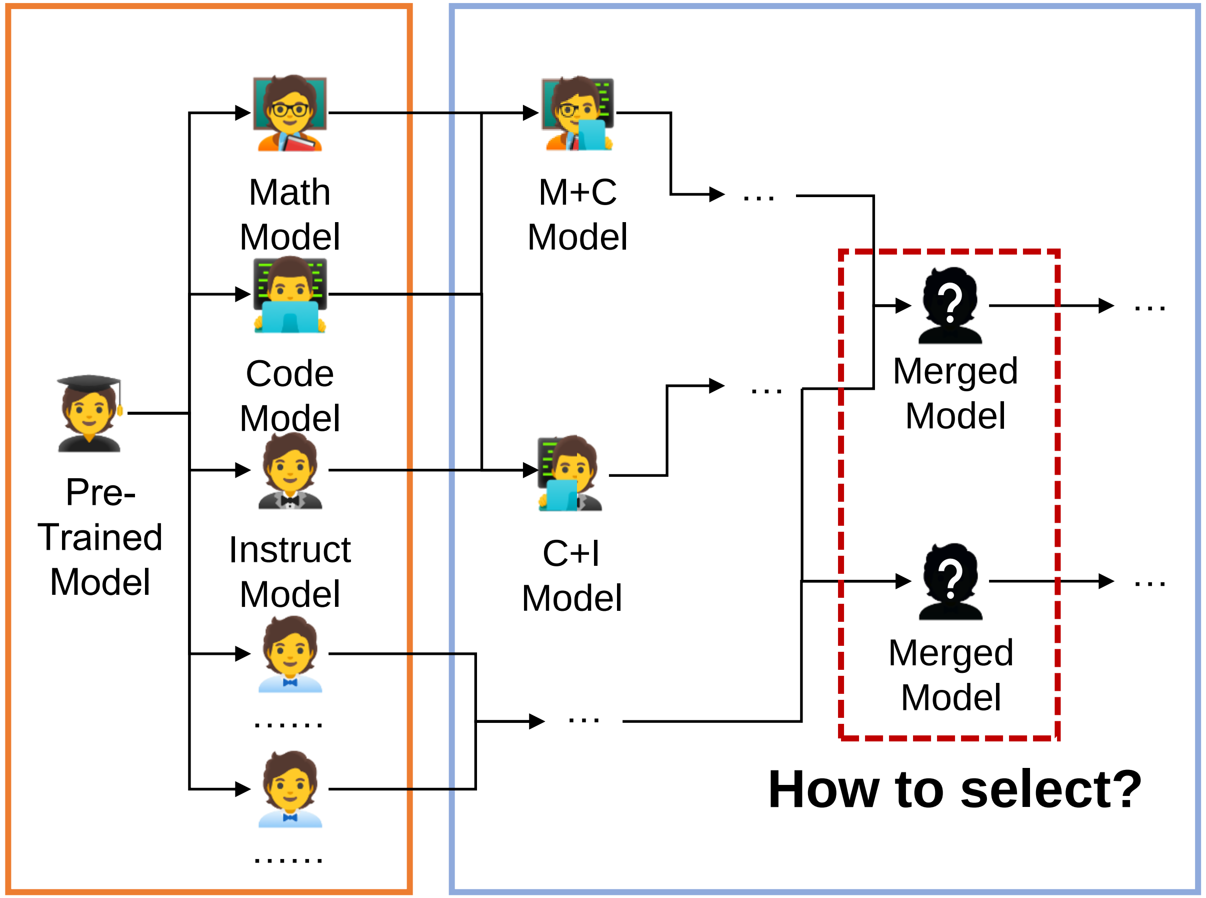
Merge

Merge selected models via model merging approach (e.g. **Linear Averaging**)

Recycle

Recycle merged model to **Model Repository** for future merging.

Merging Challenges



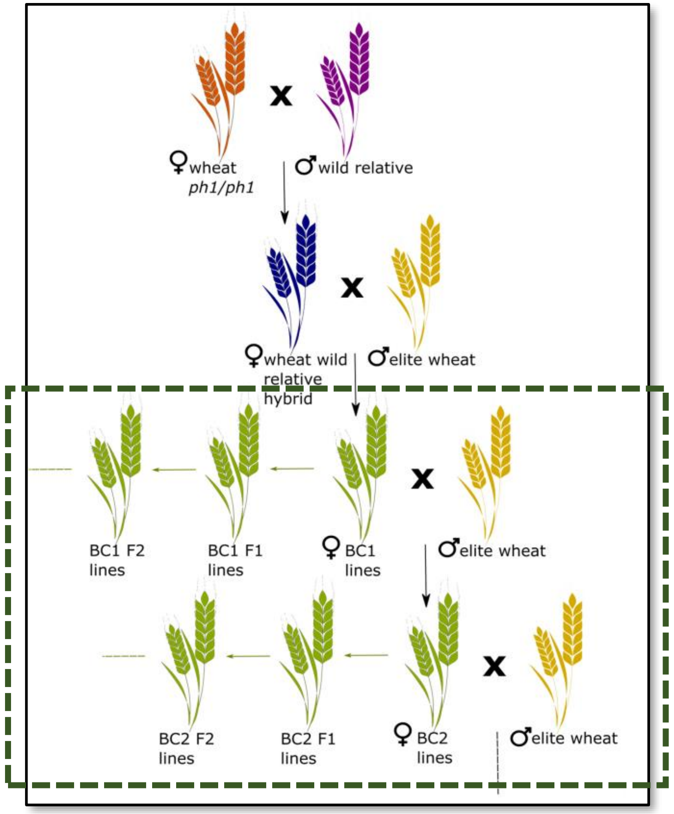
How can we select models after **multiple merges?**

- **Track foundation models?**
  - **Possible for early merged models.**
  - **Difficult for merged models after multiple merges.**
- **Compare each task performance?**
  - Possible for comparison between 2 or 3 tasks.
  - **Difficult for merging multiple tasks.**
- **Highest average task performance?**
  - Easy to identify.
  - **Potential problems?**

What else can we do?

Artificial Evolution in Biology

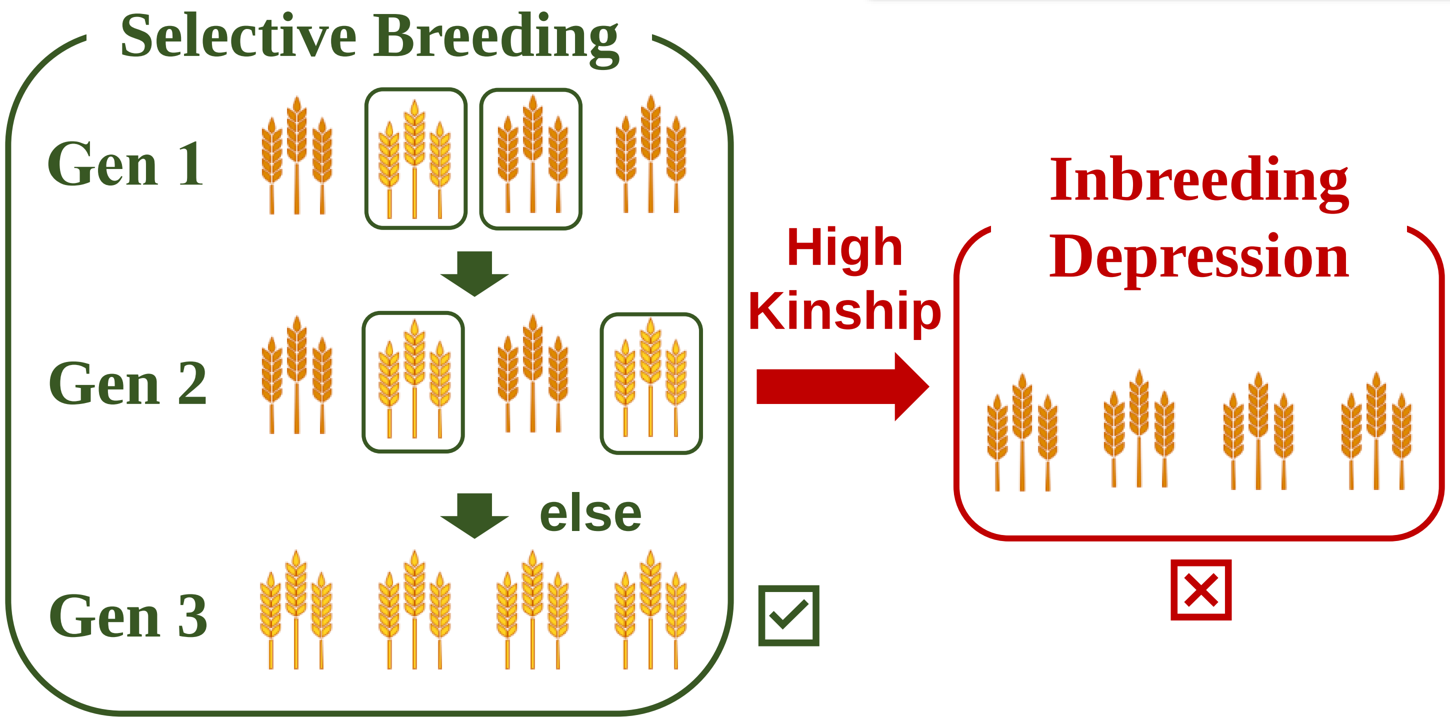
Hybridization

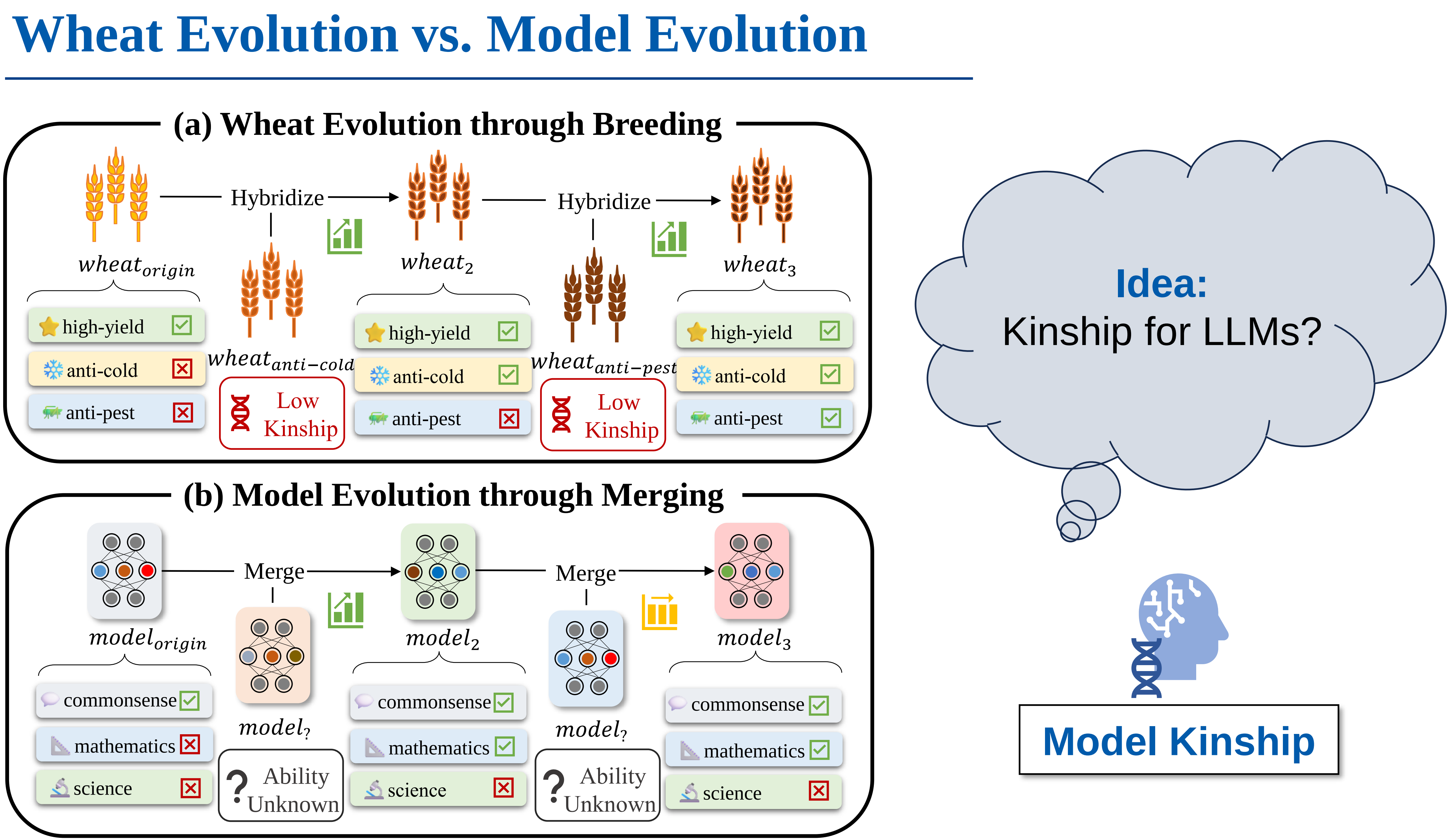


Laugerotte et. al'22

Selective breeding involves choosing parents with particular characteristics to breed together and **produce offspring with more desirable characteristics.**

Inbred populations are more likely to suffer from <u>genetic conditions</u> caused by recessive gene variants, resulting in **Inbreeding Depression**

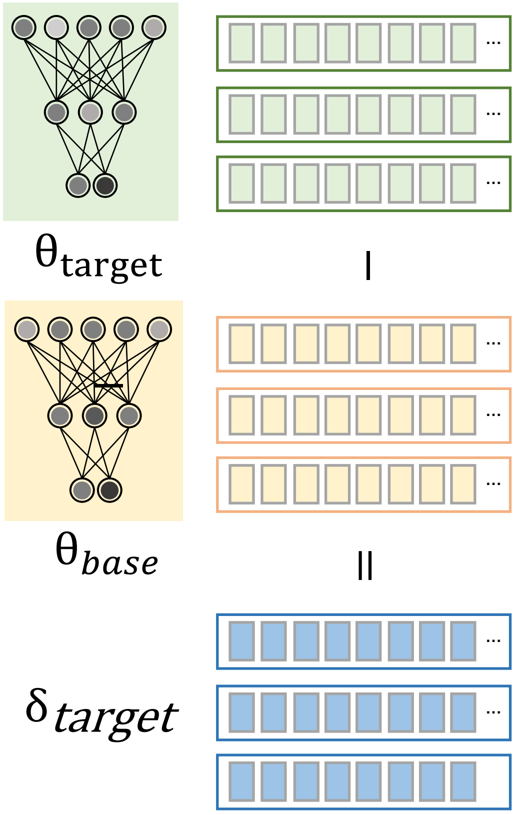



Wheat Evolution vs. Model Evolution
(a) Wheat Evolution through Breeding
Hybridize
Hybridize
wheat_origin
wheat_anti-cold
wheat_2
wheat_anti-pest
wheat_3
high-yield
anti-cold
anti-pest
Low Kinship
Low Kinship
(b) Model Evolution through Merging
Merge
Merge
model_origin
model_?
model_2
model_?
model_3
commonsense
mathematics
science
Ability Unknown
Ability Unknown
Idea:
Kinship for LLMs?
Model Kinship

Model Kinship

Delta Parameters

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



$\theta \in R^d$ is the weight of LLMs

Merging two models

$$model\ kinship = \boxed{sim(\delta_1, \delta_2)}$$

similarity metric

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

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

Merging multiple models

$$model\ kinship = \frac{2}{n(n-1)} \sum_{1 \le i < j \le n} sim(\delta_i, \delta_j)$$

Additional Metrics

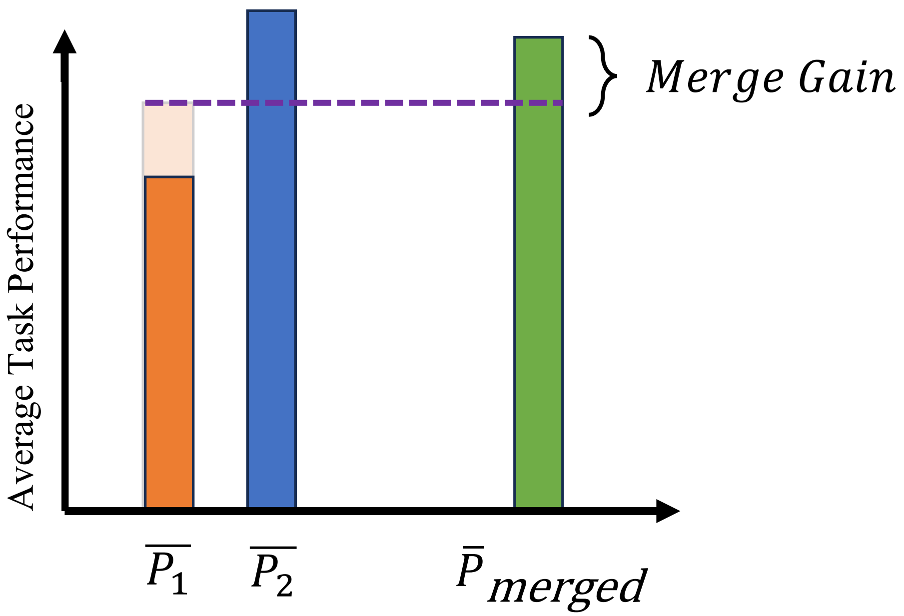
Average Task Performance

$$\bar{P} = \frac{1}{n}\sum_{i=1}^{n} \bar{P_i}$$

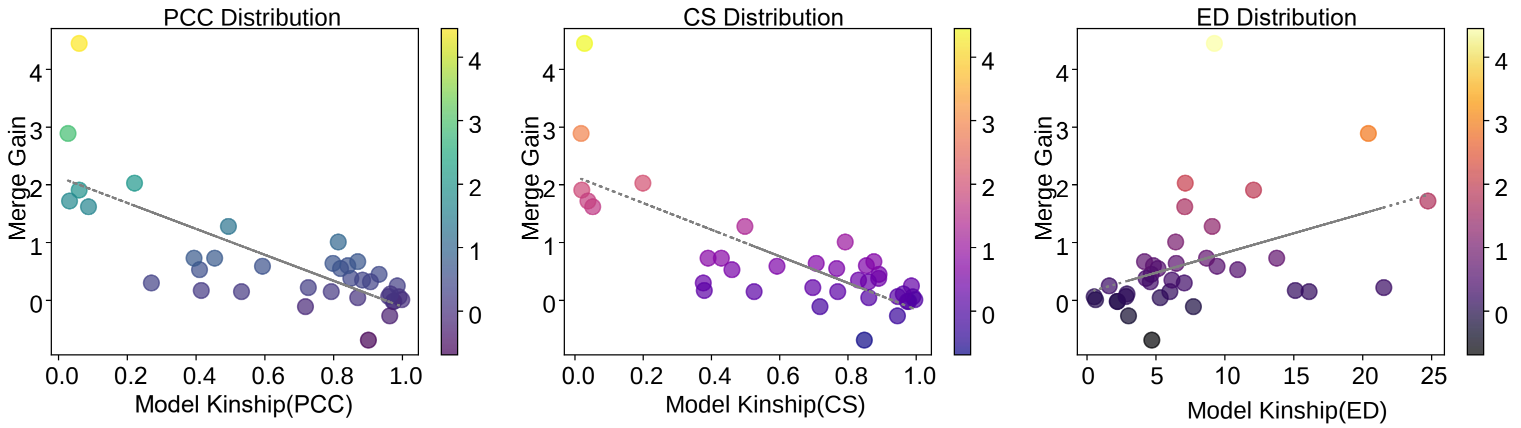
Average	ARC	HellaSwag	MMLU	TruthfulQA	Winogrande	GSM8K
81.22	79.78	91.15	77.95	74.5	87.85	76.12

Merge Gain

$$Gain = \bar{P}_{merged} - \frac{1}{k}\sum_{i=1}^{k} \bar{P}_{pre\_i}$$



Correlation: Model Kinship and Merge Gain



Bias from users?

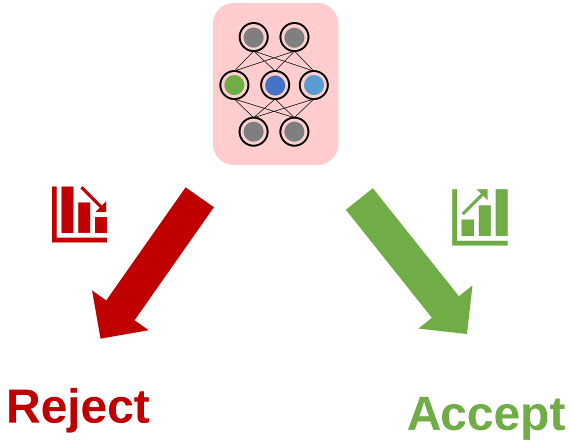


Table 1: **Correlation** of Model Kinship based on different correlation function $sim(\cdot, \cdot)$ with Merge Gain, along with their corresponding p-values.

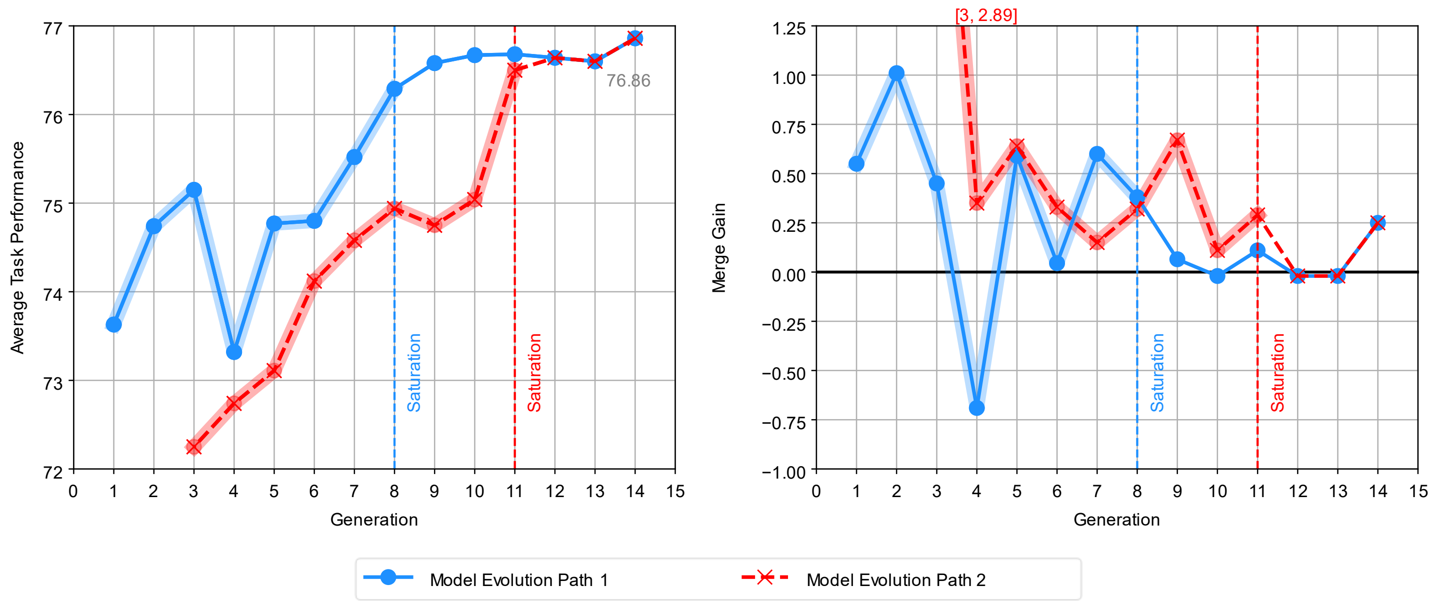
Metric	**Correlation** (Normal Value)	**Correlation (Absolute Value)**
PCC	-0.50	**-0.59**
P-value	0.063	**0.023**
CS	-0.45	**-0.66**
P-value	0.098	**0.008**
ED	0.46	**0.67**
P-value	0.091	**0.007**

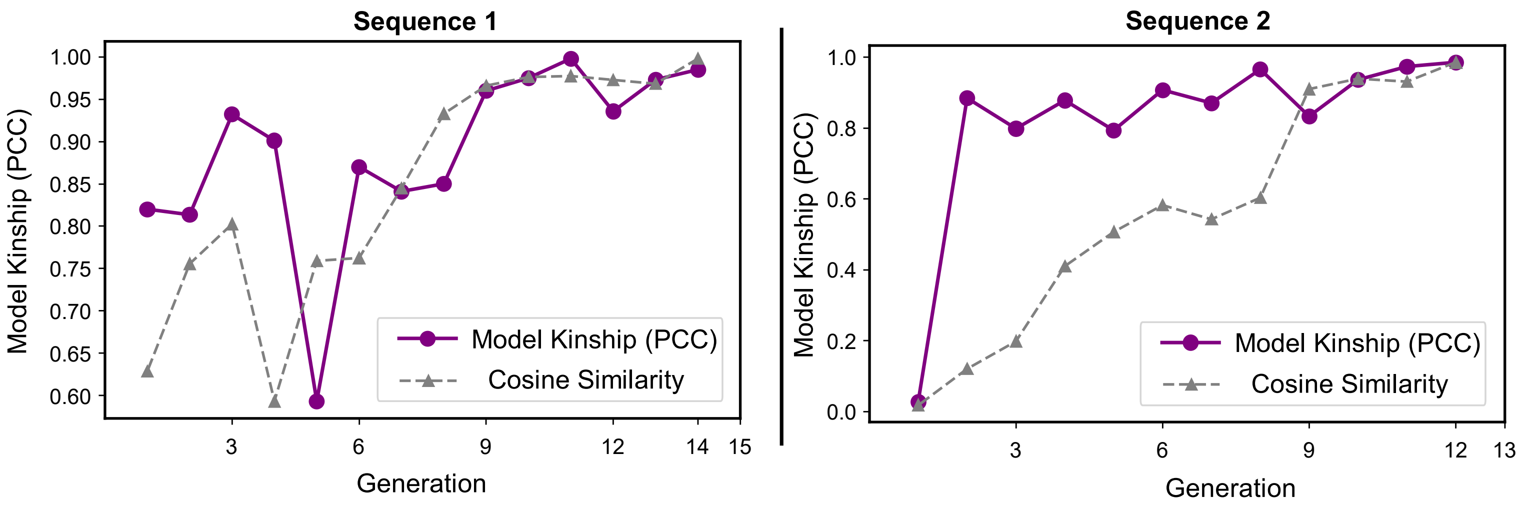
Stronger Correlation
More Confidence

Initial conclusion:
Farthest model kinship means higher merge gain

Conclusion:
Farthest model kinship means higher variation in merge performance

Analysis of Model Evolution Paths





◻ Learning Stage

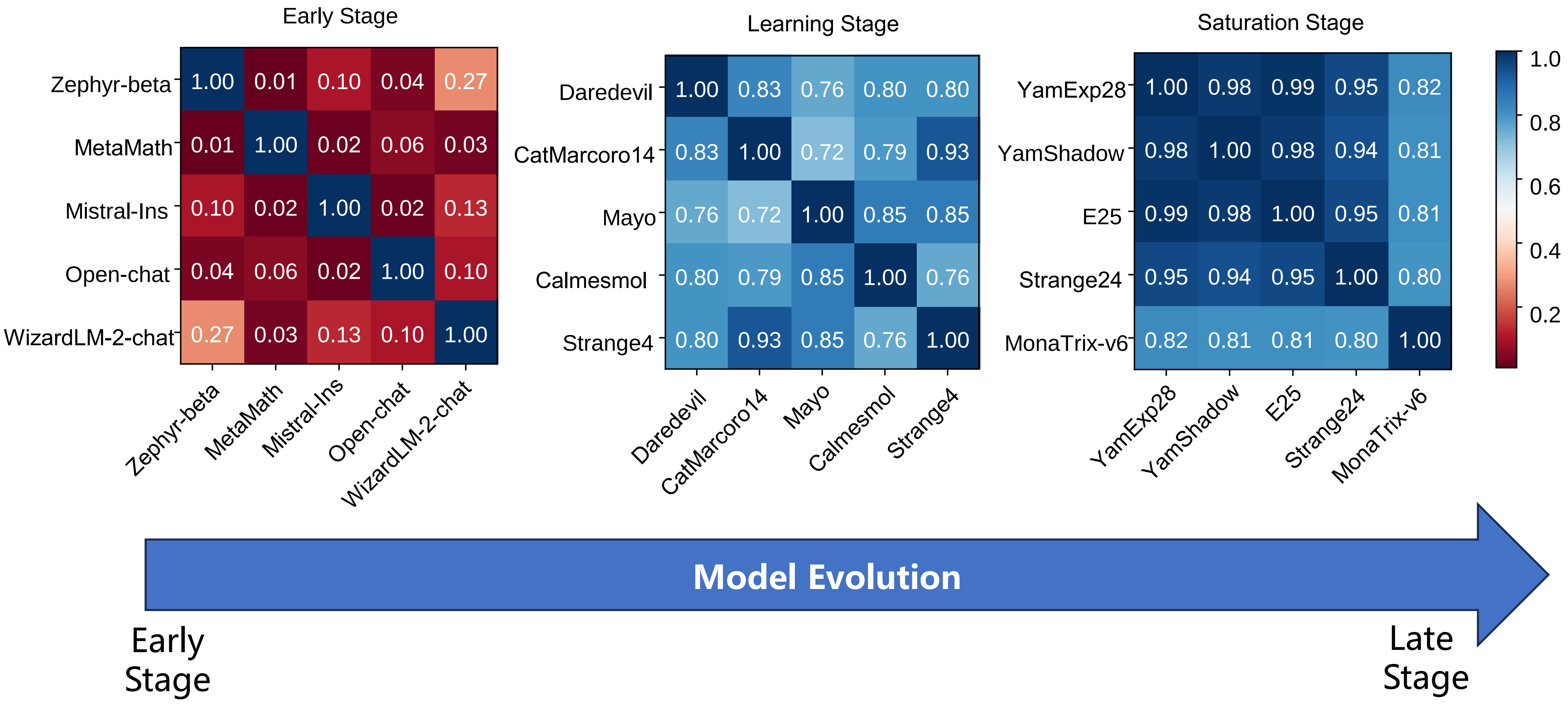
the average task performance generally experiences a rapid increase.

◻ Saturation Stage

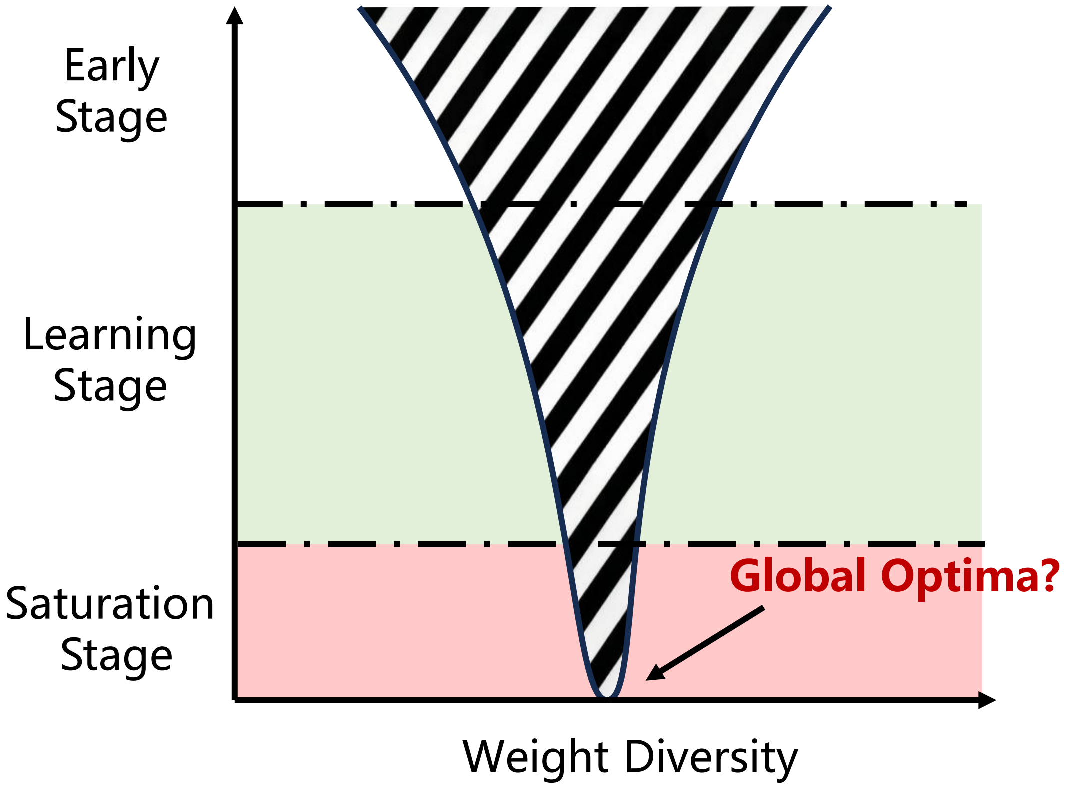
the model can no longer benefit from the merging process and has ceased to improve

Why does evolution stop?

Model Kinship within Different Stage



Convergence of Weight Space



The **weight diversity** of the top merged models **decreases** throughout the evolution.

The evolution progresses **toward an optimal set of weights**, which could either be a global optimum that fully integrates all information or **a local optimum.**

Methodology

Algorithm 1 Top k Greedy Merging with Model Kinship.

Require: A set M of n foundation models $\{M_1, M_2, \ldots, M_n\}$, Evaluation function f, Similarity metric function $sim(\cdot,\cdot)$ for model kinship.

Generate the first generation of merged models by merging each pair in set M $\{M_1, M_2, \ldots, M_n\}$.
Evaluate each merged model using f and select the top k models. Denote this set as $S = \{M_1, M_2, \ldots, M_n\}$.
Initialize a variable $S_{\text{prev}} = \emptyset$ to store the top m models from the previous iteration.

Initialization

while $S \neq S_{\text{prev}}$ **do**
  Set $M_{\text{prev}} = M$.
  Set $S_{\text{prev}} = S$.
  Select k pairs of models from S with the highest performance according to f.
  Identify the current best model $M_{best} \in S$.
  Identify the model $M_f \in S$ with the highest model kinship to M_{best} from the M_{prev} according to the similarity metric $sim(\cdot,\cdot)$.
  Merge M_f with M_{best} to generate a new model M_{exp} and add M_{exp} into set M.
  Merge each selected pair to M_{merged} (named as **Model-gen-id**) for merged models and add merged models into set M.
  Evaluate each new model using f and update S to be the new top k models.
end while

Iterative Merging

Note: The blue-highlighted steps are only executed in modified experiments incorporating model kinship-based exploration.

Main Experiment Results

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. Merged models are labeled using the structure: **Model-{generation number}-{model identification number}**.

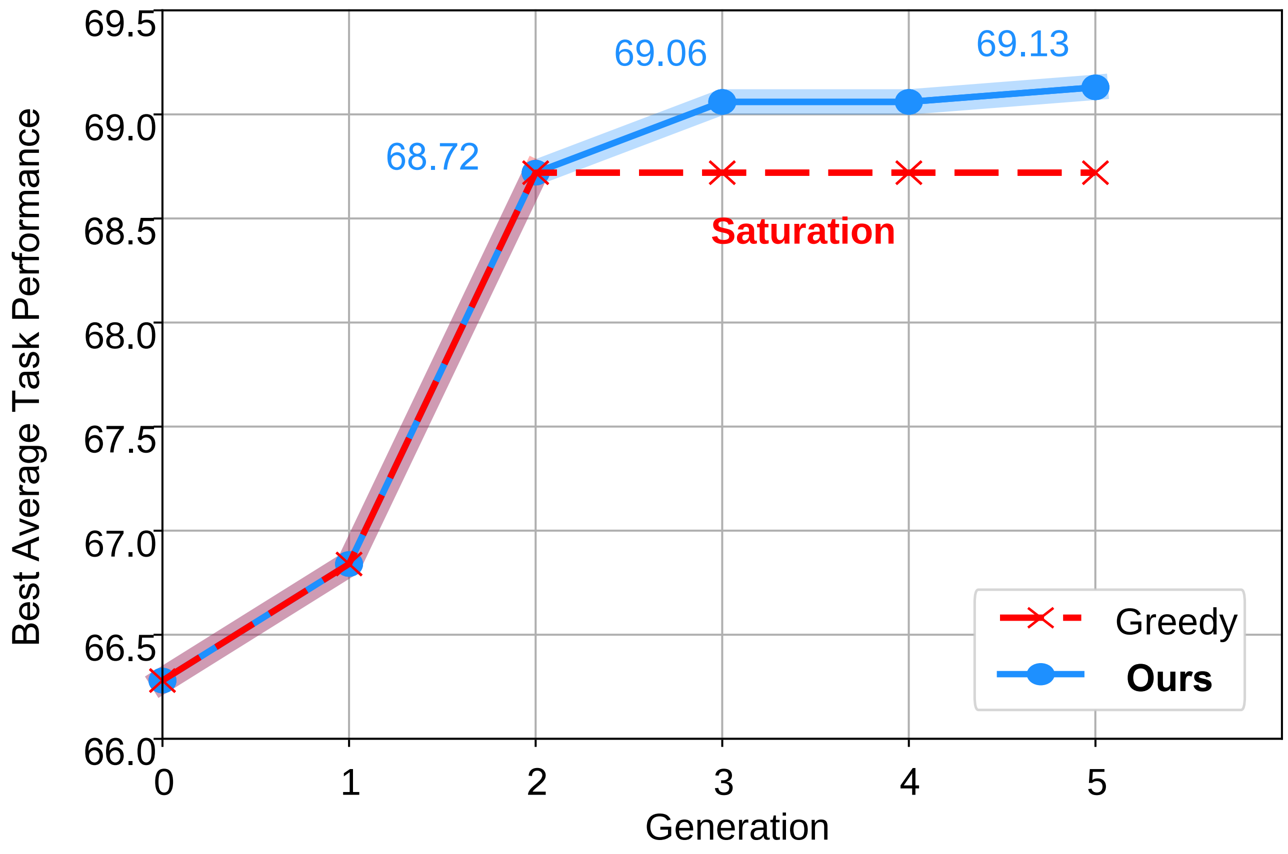
Greedy Strategy				+ Model Kinship			
Model	**Avg.**	**Gain**	**Kinship**	**Model**	**Avg.**	**Gain**	**Kinship**
MetaMath	63.72	/	/	**MetaMath**	63.72	/	/
Instruct	61.82	/	/	**Instruct**	61.82	/	/
Open-chat	66.28	/	/	**Open-chat**	66.28	/	/
model-1-1	62.17	-0.6	0.01	model-1-1	62.17	-0.6	0.01
model-1-2	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

Same selection

Models in each merge are **highly related.**

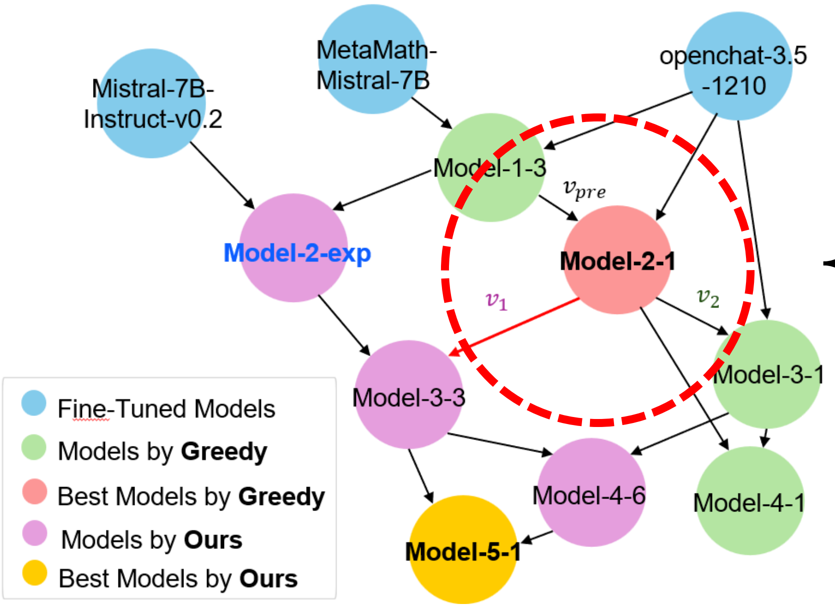
Exploration Strategy finds a better evolution path.

Main Experiment Results

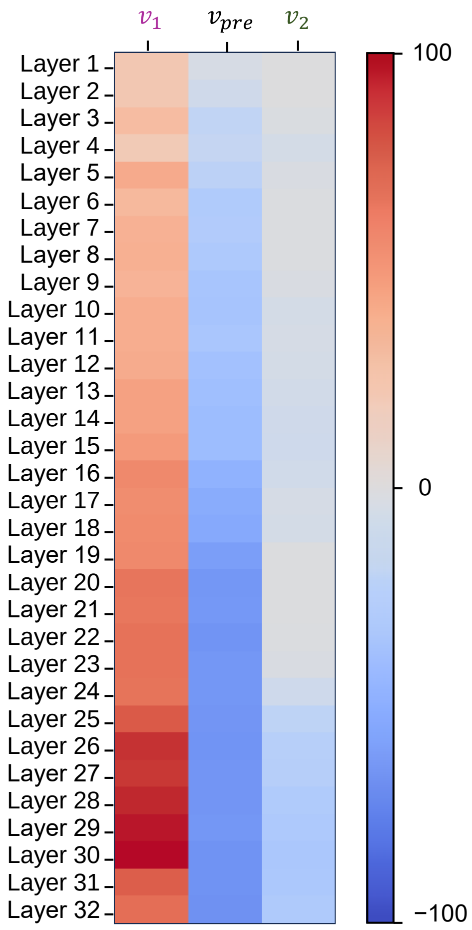


Exploration Strategy enable evolution to continue.

Weight Change in Bifurcation point



Model Tree **(Bifurcation Point)**



Weight change
in different path

Opposite direction

&

Stronger change

Effect of Model Kinship

- **Merging Models with Low Kinship can improve performance**
  - Expand searching range.
  - Escape the local optima traps.

- **Early Stopping at High Kinship can improve Efficiency**
  - Fair trade-off: small performance gains versus large time consumption

Conclusion & Future Work

- **Model Kinship: An Effective Guiding Metric for Model Evolution**
  - **Predict the Merging Outcomes.**
  - **Control the Evolution Directions.**

- **More works need to be done**
  - **Better similarity metric?**
  - **Theoretical framework of Model Evolution?**
  - **Support sustainable evolution (with the help of extrapolation[1]) ?**

[1] Weak-to-Strong Extrapolation Expedites Alignment (CoRR 2024)

Model Kinship

https://github.com/zjunlp/ModelKinship