Paper Review

6 February 2025

It's Morphing Time: Unleashing the Potential of Multiple LLMs via Multi-objective Optimization

Bingdong Li[1], Zixiang Di[1], Yanting Yang[1], Hong Qian[1], Peng Yang[2],
Hao Hao[3], Ke Tang[2], Aimin Zhou[1]
[1]East China Normal University
[2]Southern University of Science and Technology
[3]Shanghai Jiao Tong University
bdli@cs.ecnu.edu.cn, {51265901113, 51255901098}@stu.ecnu.edu.cn,
hqian@cs.ecnu.edu.cn, yangp@sustech.edu.cn,
haohao@sjtu.edu.cn, tangk3@sustech.edu.cn, amzhou@cs.ecnu.edu.cn

Why and What This Paper is About?

- Background & Current Gap
  - Finetuning is expensive, model merging is more promising
  - But, model merging requires profound knowledge and intuition on how to balance the weights
  - Or you can conduct grid search (expensive, defeat the purpose)

- Motivation
  - Automatically finding the ‘perfect’ weight for model merging, without sacrificing one or the other
  - This is done using multi-objective optimization

Why and What This Paper is About?

- Main Contributions
    1. Formalizing model merging as a multi-objective optimization problem
    2. Automated and enhanced acquisition strategy
        - Basically, how to search the best configuration faster
    3. Additional optimization objective
        - To make sure that the model generalize on different tasks
        - Reduce overfitting

Approach & Results

- Methods

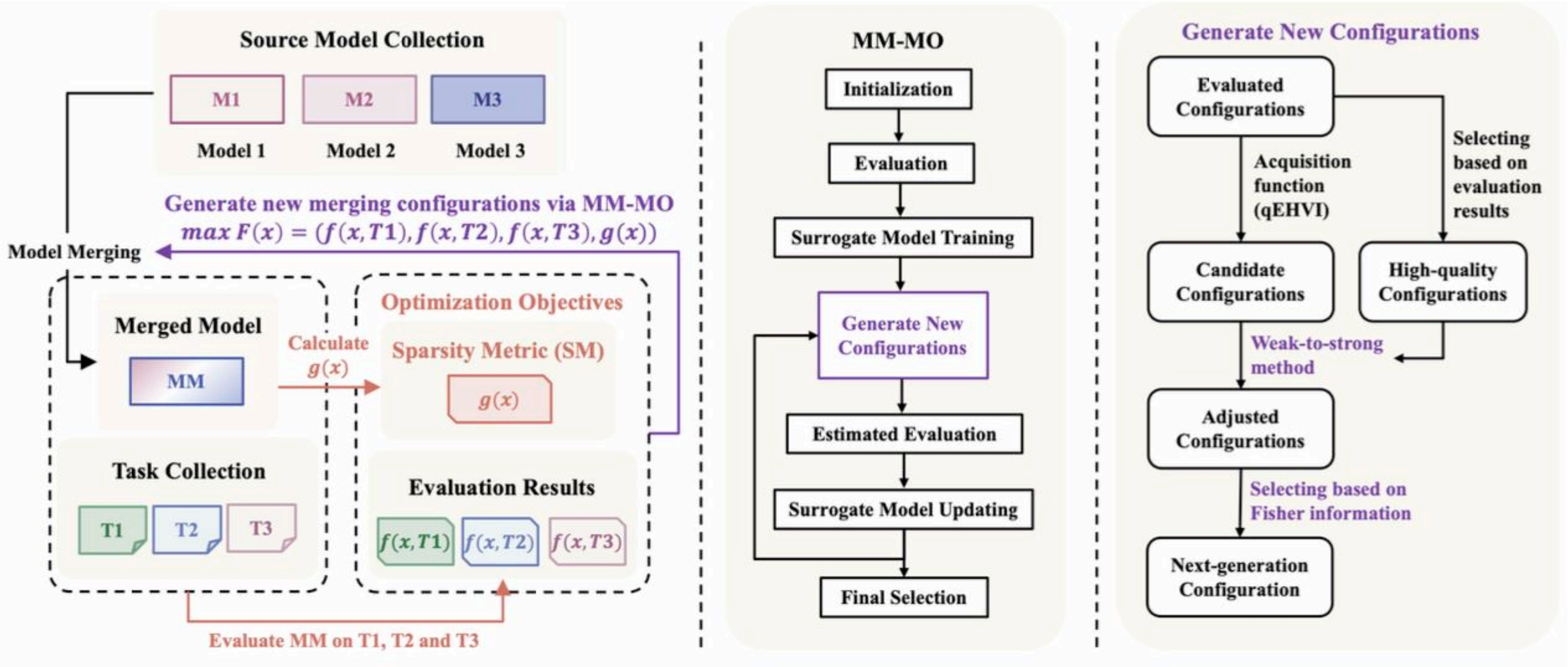
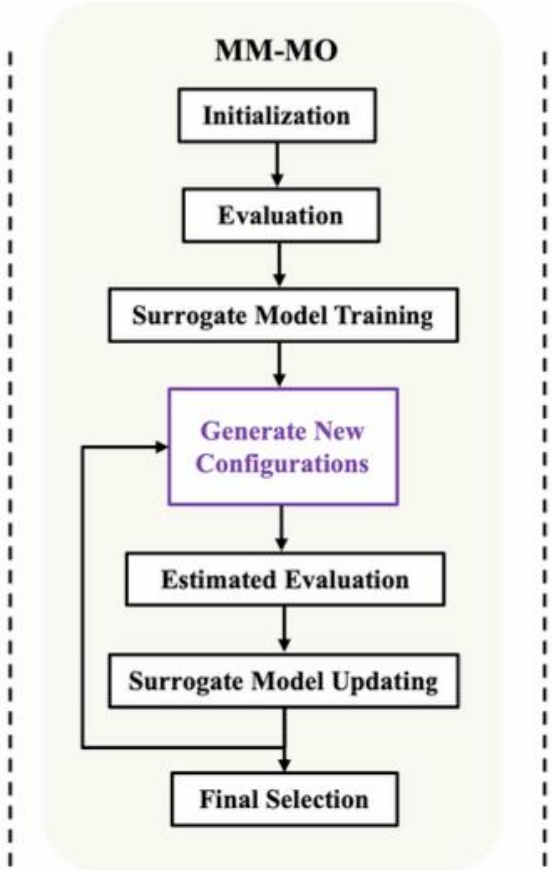


Fig. 3. An illustration of automated model merging with multi-objective optimization (MM-MO).



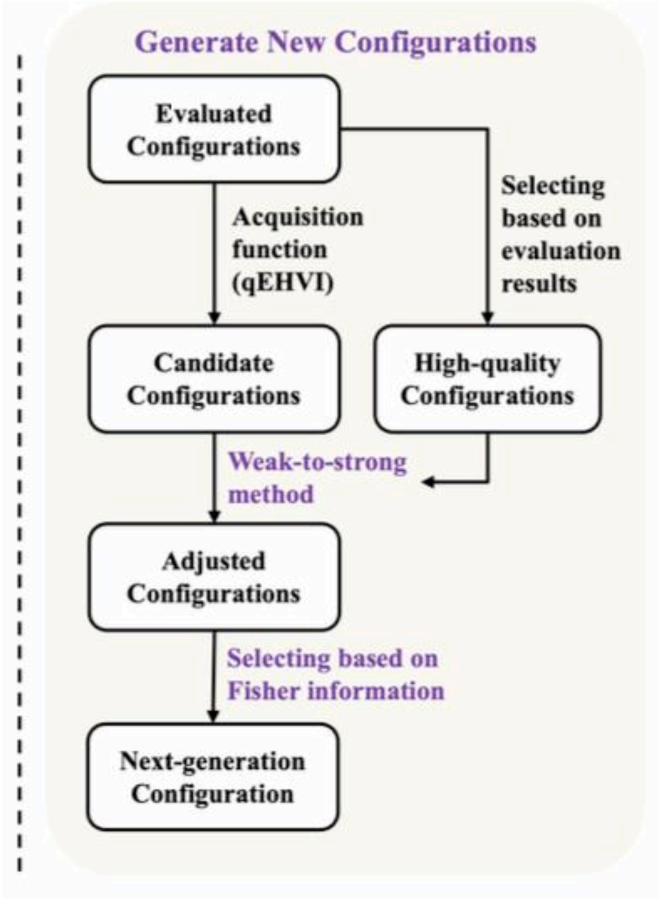
- We start with a merged model using TIES and DARE
  - TIES: Trim redundant parameters, resolve conflicting signs, take average for parameters with the same sign
  - DARE: Randomly drop parameters according to a drop rate and rescale the remaining parameters
  - In the paper, they used qwen1.5-chat, liberated-qwen (specialized in coding), firefly-qwen (specialized in Chinese)

- Next, we initialize various random configurations

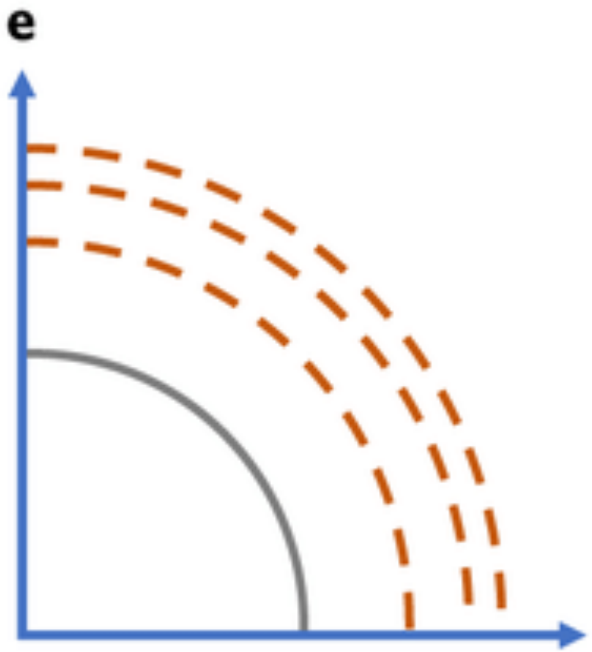
- For each of the model, we evaluate with the 2 tasks you want to optimize for
  - In the paper, they use GSM8K (math tasks) and C-EVAL (Chinese tasks)
  - Plus sparsity metric

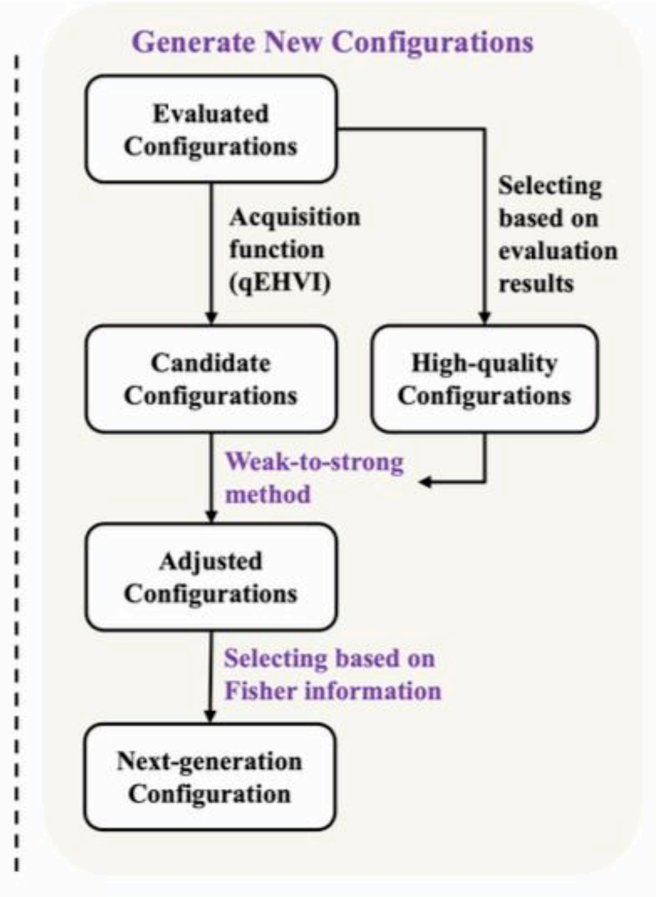
- Use the results to train a surrogate model
  - This surrogate model will aim to predict the evaluation results from different configurations

- Then, use this model to predict which configuration is best



- Using the evaluated configurations, we use qEHVI (q-Expected Hypervolume Improvement) to determine the next best configuration
  - Haven't really look deep into the math, but it quantifies how much a configuration can expand the existing Pareto frontier
  - Need the surrogate model to determine the value
  - End result: a bunch of candidate configurations

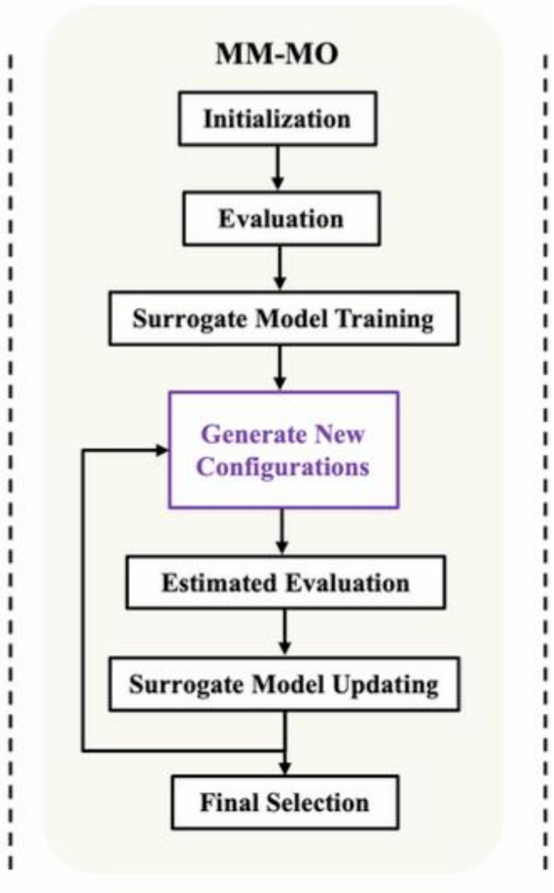




- Now, we have no narrow down the candidate to find the best ones

- They used weak-to-strong method
  - From existing evaluated configurations, pick 5 of the best ones
  - Apply differential evolution to the above 5 (why, I have no idea)
  - Then apply stochastic perturbation on the candidate configurations. This replaces some of the candidate configurations parameter with those obtained during DE (why, I also have no idea)
  - Not particularly sure why are we doing this…
  - But, apparently it can help improving the configurations

- With the improved configuration, we need to prioritize the ones with least Fisher information
  - Fisher info quantifies the uncertainty of certain configurations
  - Less certain spots are prioritized as it provides most learning about the search space
  - End result: bestest next configuration to be evaluated



- After evaluating the bestest configurations using our two tasks (GSM8K + C-EVAL), we update the surrogate model
- Iteratively find the next bestest configurations for 5 iterations
- The best performing configuration will be selected last

Approach & Results

- Key Findings
  - MM-MO method significantly outperforms existing model merging approaches

TABLE III
PERFORMANCE COMPARISON OF DIFFERENT MERGING METHODS AND SINGLE MODELS. (MODEL SIZE: 7B PARAMS)

Merging Method	Models	Average Score	C-EVAL	GSM8K	HellaSwag	HumanEval	MBPP	MMLU	WinoGrande
Single Model 1	Qwen1.5-7B-Chat	56.46	68.7	54.59	68.13	46.95	34.20	60.06	62.59
Single Model 2	Liberated-Qwen1.5-7B	57.29	69.7	53.30	71.02	48.78	38.80	58.84	60.62
Single Model 3	firefly-qwen1.5-en-7b	51.32	70.0	49.81	65.69	33.54	28.40	51.66	60.14
Linear (Model Soup)	Single Model 1 + 2 + 3	58.88	71.1	54.89	72.72	50.61	39.80	60.80	62.27
Task Arithmetic	Single Model 1 + 2 + 3	56.67	70.1	55.50	69.54	46.34	37.80	55.04	62.35
Dare + Task Arithmetic	Single Model 1 + 2 + 3	58.38	69.8	55.27	69.97	51.22	39.40	60.23	62.75
TIES	Single Model 1 + 2 + 3	53.32	65.7	53.15	69.58	34.76	29.00	58.05	62.98
DARE + TIES	Single Model 1 + 2 + 3	57.78	69.5	55.72	69.23	49.39	37.80	60.06	62.75
Model Breadcrumbs	Single Model 1 + 2 + 3	58.56	70.3	56.10	70.91	49.39	40.60	60.47	62.12
Model Breadcrumbs + TIES	Single Model 1 + 2 + 3	58.41	70.2	55.72	70.59	50.00	39.80	60.35	62.19
DARE + TIES w/ MM-MO (Ours)	Single Model 1 + 2 + 3	**60.97**	**71.9**	**57.77**	**74.44**	**55.49**	**42.20**	**60.81**	**64.17**

Approach & Results

- Key Findings
  - Adding sparsity metric improves generalization
    - MM-MO performs better than other model merging method in non-trained fields like common sense and reasoning
    - MM-MO enhances the overall potential of the model, instead of optimizing in specific given tasks

TABLE III
PERFORMANCE COMPARISON OF DIFFERENT MERGING METHODS AND SINGLE MODELS. (MODEL SIZE: 7B PARAMS)

Merging Method	Models	Average Score	C-EVAL	GSM8K	HellaSwag	HumanEval	MBPP	MMLU	WinoGrande
Single Model 1	Qwen1.5-7B-Chat	56.46	68.7	54.59	68.13	46.95	34.20	60.06	62.59
Single Model 2	Liberated-Qwen1.5-7B	57.29	69.7	53.30	71.02	48.78	38.80	58.84	60.62
Single Model 3	firefly-qwen1.5-en-7b	51.32	70.0	49.81	65.69	33.54	28.40	51.66	60.14
Linear (Model Soup)	Single Model 1 + 2 + 3	58.88	71.1	54.89	72.72	50.61	39.80	60.80	62.27
Task Arithmetic	Single Model 1 + 2 + 3	56.67	70.1	55.50	69.54	46.34	37.80	55.04	62.35
Dare + Task Arithmetic	Single Model 1 + 2 + 3	58.38	69.8	55.27	69.97	51.22	39.40	60.23	62.75
TIES	Single Model 1 + 2 + 3	53.32	65.7	53.15	69.58	34.76	29.00	58.05	62.98
DARE + TIES	Single Model 1 + 2 + 3	57.78	69.5	55.72	69.23	49.39	37.80	60.06	62.75
Model Breadcrumbs	Single Model 1 + 2 + 3	58.56	70.3	56.10	70.91	49.39	40.60	60.47	62.12
Model Breadcrumbs + TIES	Single Model 1 + 2 + 3	58.41	70.2	55.72	70.59	50.00	39.80	60.35	62.19
DARE + TIES w/ MM-MO (Ours)	Single Model 1 + 2 + 3	**60.97**	**71.9**	**57.77**	**74.44**	**55.49**	**42.20**	**60.81**	**64.17**

Approach & Results

- Key Findings
  - MM-MO outperforms evolutionary model merge

TABLE VI
PERFORMANCE COMPARISON OF MM-MO AND EMM. (MODEL SIZE: 7B & 13B PARAMS)

Merging Method	Models	C-EVAL	MMLU	GSM8K	Human Eval	MBPP
Single Model 1	Qwen1.5-7B-Chat	68.7	60.06	54.59	46.95	34.20
Single Model 2	Liberated-Qwen1.5-7B	69.7	58.84	53.30	48.78	38.80
Single Model 3	firefly-qwen1.5-en-7b	70.0	51.66	49.81	33.54	28.40
Single Model 4	WizardLM-13B	34.8	51.47	55.50	35.98	30.60
Single Model 5	WizardMath-13B	30.2	51.27	60.50	14.02	25.20
Single Model 6	llama-2-13b-code-alpaca	33.2	52.99	29.72	21.95	30.00
DARE + TIES	Single Model 1 + 2 + 3	69.5	60.06	55.72	49.39	37.80
DARE + TIES w/ EMM	Single Model 1 + 2 + 3	68.0	58.99	**62.02**	34.76	29.60
DARE + TIES w/ MM-MO (Ours)	Single Model 1 + 2 + 3	**71.9**	**60.81**	57.77	**55.49**	**42.20**
DARE + TIES	Single Model 4 + 5 + 6	37.7	55.57	60.73	33.54	33.00
DARE + TIES w/ EMM	Single Model 4 + 5 + 6	32.5	52.16	60.05	34.15	25.20
DARE + TIES w/ MM-MO (Ours)	Single Model 4 + 5 + 6	**38.0**	**56.36**	**62.85**	**36.59**	**36.80**

Approach & Results

- Key Findings
  - qEHVI is the most effective acquisition function for multi-objective optimization

TABLE VII
PERFORMANCE COMPARISON OF DIFFERENT ACQUISITION FUNCTIONS

Method	Average Score	C-EVAL	MMLU	GSM8K	Human Eval	MBPP
DARE + TIES	54.49	69.5	60.06	55.72	49.39	37.80
MM-MO / qNEHVI	56.09	71.6	**61.07**	57.16	50.61	40.00
MM-MO / qNParEGO	55.09	70.3	60.45	**59.74**	46.95	38.00
MM-MO / qEHVI (Ours)	**57.63**	**71.9**	60.81	57.77	**55.49**	**42.20**

Analysis

- Strengths
  - Interesting approach to model merging, very technical
  - Relevant, find the ‘sweet spot’ in the Pareto front
  - Strong experimental validation, even with small models
  - I like how it improves general ability of the LLM as well

- Weaknesses
  - Requires the source models to be homologous to ensure compatibility
  - Very complex. Not sure how to implement.
  - No code given

Questions

1. Do you think it will be hard to implement? Not feasible?
2. Not sure about how qEHVI works
   - AFAIK, we want to maximize the task evaluations and minimize the sparsity metric
   - Where and how is this objective relayed to qEHVI?