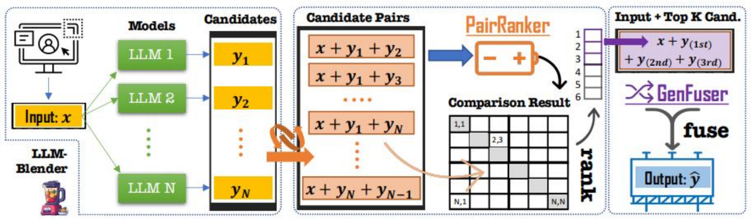
Weighted-Reward Preference Optimization for Implicit Model Fusion

Ziyi Yang* Fanqi Wan* Longguang Zhong Tianyuan Shi Xiaojun Quan †
School of Computer Science and Engineering, Sun Yat-sen University
yangzy39@mail2.sysu.edu.cn, quanxj3@mail.sysu.edu.cn

Backgrounds: Collective LLMs

LLM-BLENDER: Ensembling Large Language Models with Pairwise Ranking and Generative Fusion

Dongfu Jiang[♡] Xiang Ren[♣♠] Bill Yuchen Lin[♠]
dongfu@zju.edu.cn, xiangren@usc.edu, yuchenl@allenai.org
[♠]Allen Institute for Artificial Intelligence
[♣]University of Southern California [♡]Zhejiang University

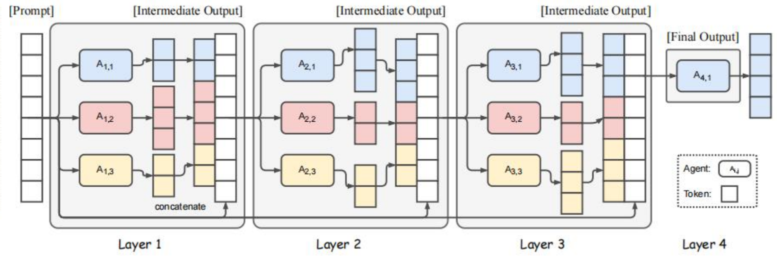


Mixture-of-Agents Enhances Large Language Model Capabilities

Junlin Wang
Duke University
Together AI
junlin.wang2@duke.edu

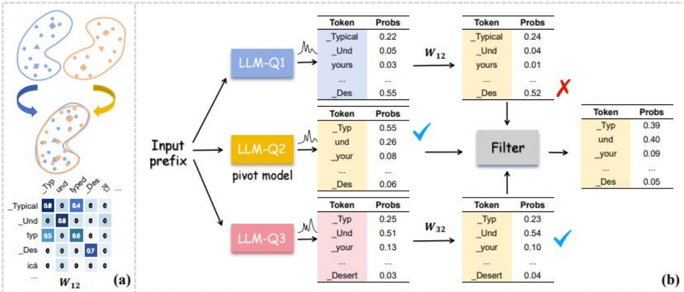
Jue Wang
Together AI
jue@together.ai

Ben Athiwaratkun
Together AI
ben@together.ai



Bridging the Gap between Different Vocabularies for LLM Ensemble

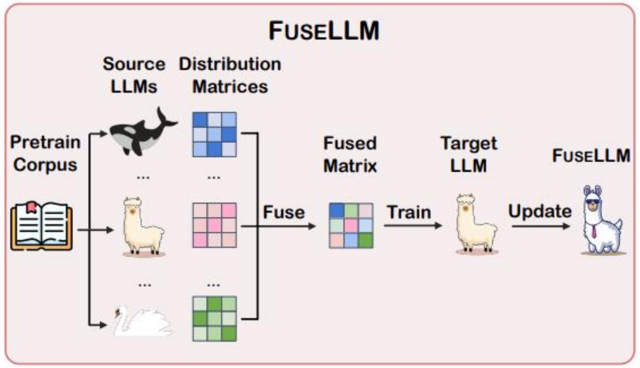
Yangyifan Xu[1,2]*, Jinliang Lu[1,2]*, Jiajun Zhang[1,2,3,4†]
[1]School of Artificial Intelligence, University of Chinese Academy of Sciences



Model Ensemble

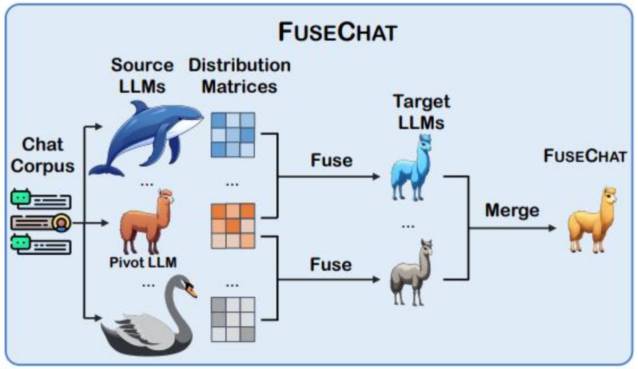
KNOWLEDGE FUSION OF LARGE LANGUAGE MODELS

Fanqi Wan[1]*, Xinting Huang[2]†, Deng Cai[2], Xiaojun Quan[1]†, Wei Bi[2], Shuming Shi[2]
[1]School of Computer Science and Engineering, Sun Yat-sen University, China
[2]Tencent AI Lab
wanfq@mail2.sysu.edu.cn, quanxj3@mail.sysu.edu.cn



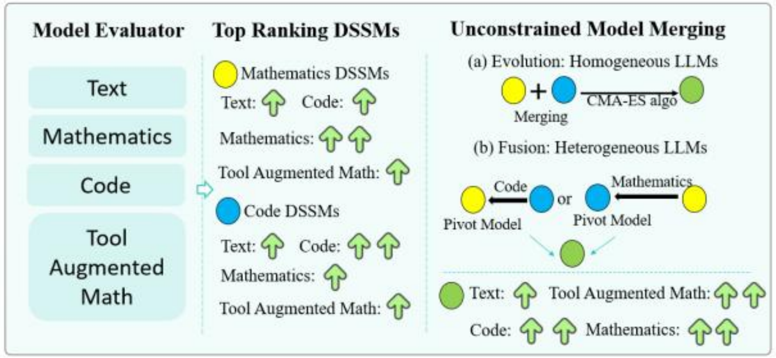
FUSECHAT: Knowledge Fusion of Chat Models

Fanqi Wan, Longguang Zhong, Ziyi Yang, Ruijun Chen, Xiaojun Quan*
School of Computer Science and Engineering, Sun Yat-sen University, China
wanfq@mail2.sysu.edu.cn, quanxj3@mail.sysu.edu.cn



Unconstrained Model Merging for Enhanced LLM Reasoning

Yiming Zhang[1], Baoyi He[2], Shengyu Zhang[2], Yuhao Fu[7], Qi Zhou[4], Zhijie Sang[3], Zijin Hong[1], Kejing Yang[3], Wenjun Wang[5], Jianbo Yuan[7] Guanghan Ning[7], Linyi Li[6], Chunlin Ji[7], Fei Wu[2], Hongxia Yang[1,3]*
[1] The Hong Kong Polytechnic University, [2] Zhejiang University, [3] Reallm Labs,
[4] Harbin Institute of Technology, Shenzhen [5] South China University of Technology,



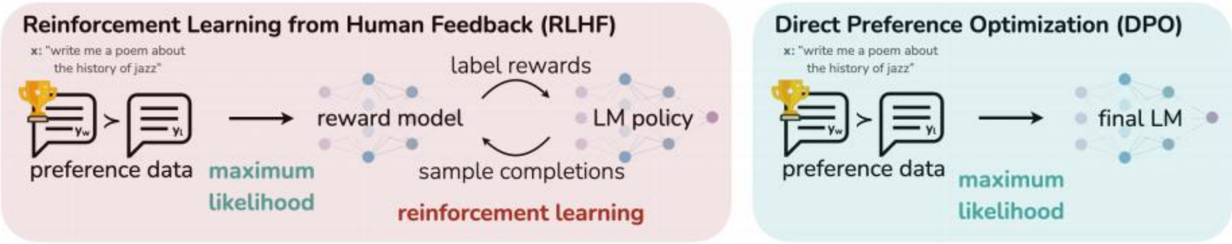
Model Fusion

Backgrounds: Direct Preference Optimization

Direct Preference Optimization: Your Language Model is Secretly a Reward Model

Rafael Rafailov*† Archit Sharma*† Eric Mitchell*†
Stefano Ermon†‡ Christopher D. Manning† Chelsea Finn†

†Stanford University ‡CZ Biohub
{rafailov,architsh,eric.mitchell}@cs.stanford.edu



A General Theoretical Paradigm to Understand Learning from Human Preferences

Mohammad Gheshlaghi Azar Mark Rowland Bilal Piot
Daniel Guo Daniele Calandriello Michal Valko Rémi Munos
Google DeepMind

Algorithm 1 Sampled IPO

Require: Dataset $\mathcal{D}$ of prompts, preferred and dis-preferred generations x, y_w and y_l, respectively. A reference policy π_{ref}

1: Define

$$h_\pi(y, y', x) = \log\left(\frac{\pi(y|x)\pi_{\text{ref}}(y'|x)}{\pi(y'|x)\pi_{\text{ref}}(y|x)}\right)$$

2: Starting from $\pi = \pi_{\text{ref}}$ minimize

$$\mathop{\mathbb{E}}_{(y_w, y_l, x)\sim D}\left(h_\pi(y_w, y_l, x) - \frac{\tau^{-1}}{2}\right)^2.$$

SimPO: Simple Preference Optimization with a Reference-Free Reward

Yu Meng[1]* Mengzhou Xia[2]* Danqi Chen[2]
[1]Computer Science Department, University of Virginia
[2]Princeton Language and Intelligence (PLI), Princeton University
yumeng5@virginia.edu
{mengzhou,danqic}@cs.princeton.edu

$$\mathcal{L}_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}}) = -\mathbb{E}\left[\log\sigma\left(\beta\log\frac{\pi_\theta(y_w \mid x)}{\pi_{\text{ref}}(y_w \mid x)} - \beta\log\frac{\pi_\theta(y_l \mid x)}{\pi_{\text{ref}}(y_l \mid x)}\right)\right]$$

$$\mathcal{L}_{\text{SimPO}}(\pi_\theta) = -\mathbb{E}\left[\log\sigma\left(\frac{\beta}{|y_w|}\log\pi_\theta(y_w \mid x) - \frac{\beta}{|y_l|}\log\pi_\theta(y_l \mid x) - \gamma\right)\right]$$

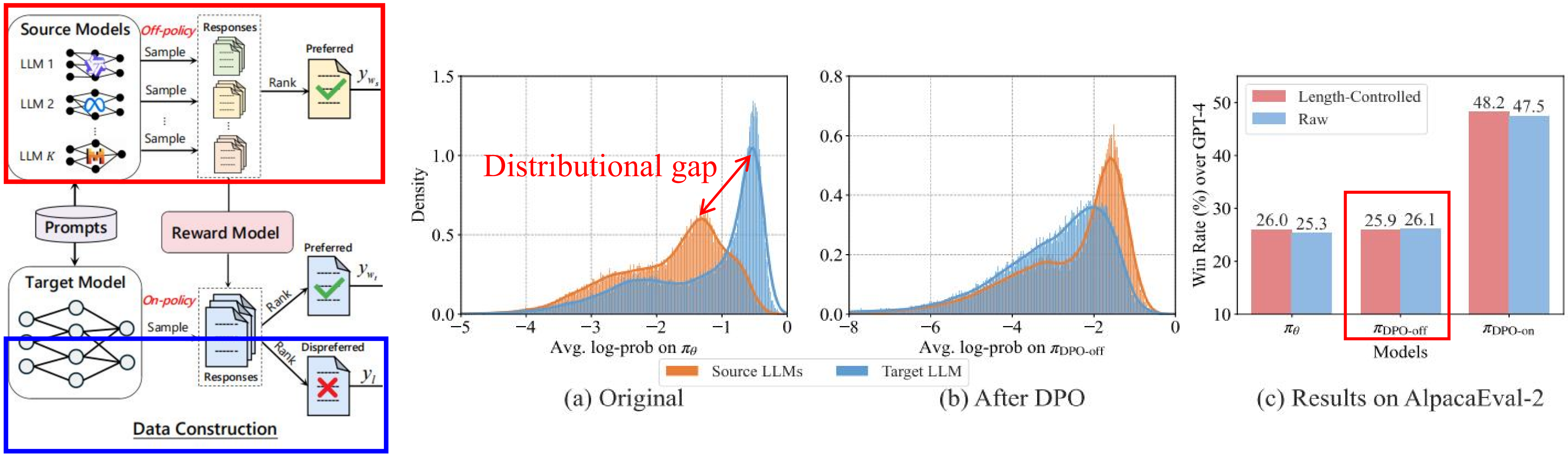
WRPO: Preliminary Experiment

Limitations of Existing Methods:

- Model ensemble: all models must **remain active** during inference
- Model fusion: **complicated** vocabulary and distribution matrices alignment process
- Direct Preference Optimization: **sensitive** to distribution shifts

Our Goal:

- Combining the strengths of multiple source LLMs into target LLM

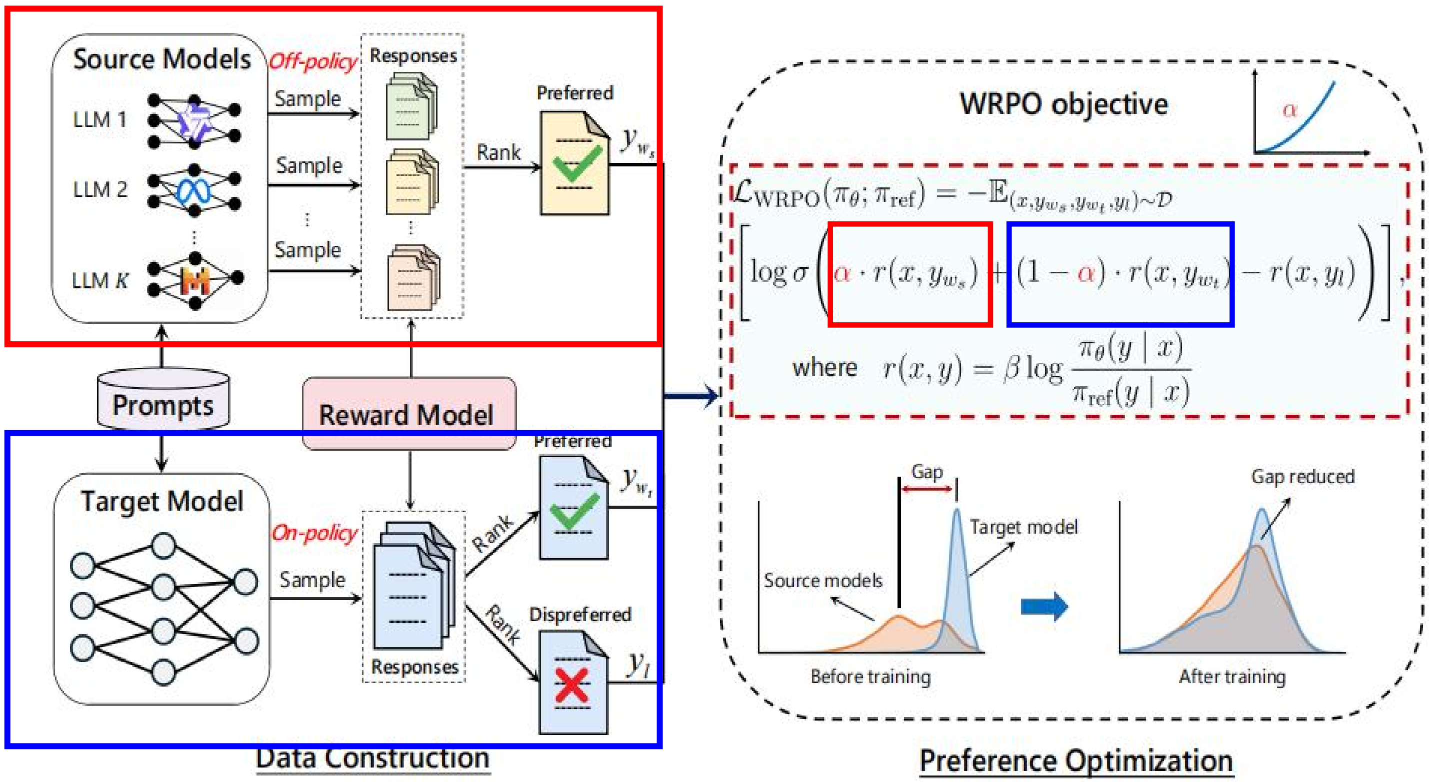


(a) Original (b) After DPO (c) Results on AlpacaEval-2

WRPO：Weighted-Reward Preference Optimization

WRPO Design

- **Implicit Fusion**： learning from discrepancies between y_{w_s} and y_{w_t}, y_l
- **Progressive adaptation**： shifting constitution of preferred response from y_{w_t} **to** y_{w_s}
- **Weigheted-reward mechanism**： increasing the weight for source LLMs and decreasing the weight of internal rewards for target LLM



WRPO: Experiment Setup

- **Target&Source LLMs:** Llama-3-8B-Instruct, Mistral-Large-Instruct-2407, Gemma-2-27B-it, Qwen-2-72B-Instruct, Llama-3-70B-Instruct, Gemma-2-9B-it, InternLM-2.5-20B-Chat, DeepSeek-V2-Chat, DeepSeek-Coder-V2-Instruct, Yi-1.5-34B-Chat, Phi-3-medium-4k-instruct
- **Training Dataset**
  - Prompt selection: UltraFeedback (60k, instruction following)
  - Responses: sampled from each source model (N=5, top-p = 0.95, temperature = 0.8)
  - Reward score: annotated by ArmoRM-Llama-3-8B-v0.1 reward model
- **Evaluation Benchmarks:** MT-Bench, AlpacaEval-2, and Arena-Hard
- **Baselines**
  - Target&Source LLMs
  - Collective LLMs: PackLLM, LLM-Blender, MOA, FuseLLM, FuseChat
  - Preference optimization methods: DPO, SimPO, IPO

WRPO: Main Results

Model	Size	AlpacaEval-2 (GPT-4-1106-Preview) LC(%)	AlpacaEval-2 (GPT-4-1106-Preview) WR(%)	Arena-Hard (GPT-4-1106-Preview) WR(%)	MT-Bench (GPT-4-0125-Preview) T1	MT-Bench (GPT-4-0125-Preview) T2	MT-Bench (GPT-4-0125-Preview) Overall
Source&Target LLMs							
Target	8B	26.0	25.3	20.6	7.41	7.04	7.23
Mistral-Large-Instruct-2407	123B	54.3	46.8	70.4	8.83	8.31	8.57
Gemma2-27B-IT	27B	55.5	41.0	57.5	8.34	8.03	8.19
Qwen2-72B-Instruct	72B	38.1	29.9	46.9	8.44	7.84	8.15
LLaMA3-70B-Instruct	70B	34.4	33.2	46.6	8.61	7.77	8.19
Gemma2-9B-IT	9B	51.1	38.1	40.8	8.27	7.44	7.86
Internlm2.5-20B-Chat	20B	37.4	45.3	31.2	8.03	7.23	7.64
DeepSeek-V2-Chat	236B	51.4	51.3	68.3	8.65	7.96	8.31
DeepSeek-Coder-V2-Instruct	236B	50.7	54.0	66.3	8.80	7.42	8.13
Yi-1.5-34B-Chat	34B	37.5	44.5	42.6	7.99	7.64	7.81
Phi-3-Medium-4K-Instruct	14B	29.8	24.2	33.4	8.63	7.46	8.04
Collective LLMs							
PackLLM-Top1-PPL	849B	49.1	48.0	64.8	8.29	8.20	8.25
LLM-Blender-Top1	849B	46.2	44.3	58.2	8.69	8.06	8.38
MoA	849B	61.3	77.2	83.1	9.04	8.03	8.54
Target-FuseLLM	8B	36.0	33.8	32.1	7.53	7.13	7.33
Target-FuseChat	8B	38.1	35.2	32.7	7.68	7.07	7.38
Preference Optimization Methods							
Target-DPO	8B	48.2	47.5	35.2	7.68	7.23	7.46
Target-SimPO	8B	53.7	47.5	36.5	7.73	7.00	7.38
Target-IPO	8B	46.8	42.4	36.6	7.89	7.19	7.54
Our Methods							
Target-SFT	8B	27.2	26.0	24.7	7.69	7.03	7.36
Target-SFT-DPO	8B	50.7	53.1	40.2	**7.98**	7.23	7.61
Target-SFT-WRPO-Medium	8B	53.5	53.8	41.6	7.80	7.03	7.42
Target-SFT-WRPO	8B	**55.9**	**57.6**	**46.2**	7.95	**7.31**	**7.63**

- Outperform all source LLMs on AlpacaEval-2
- Comparable to ensemble methods that are 106 times larger in scale (49.1->55.9)
- More superior to same size explicit model fusion technique (38.1->55.9)
- Consistently outperforms preference optimization baselines (48.2->55.9)

WRPO: Adaptability

- **Generalizability**: combining WRPO with SimPO (53.9 -> 55.8) and IPO (51.1 -> 53.3) consistently improves their performance

- **Scalability**: scaling up the number of source LLMs can enhance the overall performance of our method

Method	Objective
DPO (Rafailov et al., 2023)	$-\log\sigma\left(\beta\log\frac{\pi_\theta(y_w\|x)}{\pi_{\text{ref}}(y_w\|x)}-\beta\log\frac{\pi_\theta(y_l\|x)}{\pi_{\text{ref}}(y_l\|x)}\right)$
SimPO (Meng et al., 2024)	$-\log\sigma\left(\frac{\beta}{\|y_w\|}\log\pi_\theta(y_w\|x)-\frac{\beta}{\|y_l\|}\log\pi_\theta(y_l\|x)-\gamma\right)$
IPO (Azar et al., 2024)	$\left(\log\frac{\pi_\theta(y_w\|x)}{\pi_{\text{ref}}(y_w\|x)}-\log\frac{\pi_\theta(y_l\|x)}{\pi_{\text{ref}}(y_l\|x)}-\frac{1}{2\tau}\right)^2$
WRPO_{DPO}	$-\log\sigma\left(\alpha\cdot\beta\log\frac{\pi_\theta(y_{w_s}\|x)}{\pi_{\text{ref}}(y_{w_s}\|x)}+(1-\alpha)\cdot\beta\log\frac{\pi_\theta(y_{w_t}\|x)}{\pi_{\text{ref}}(y_{w_t}\|x)}-\beta\log\frac{\pi_\theta(y_l\|x)}{\pi_{\text{ref}}(y_l\|x)}\right)$
$\text{WRPO}_{\text{SimPO}}$	$-\log\sigma\left(\alpha\cdot\frac{\beta}{\|y_{w_s}\|}\log\pi_\theta(y_{w_s}\|x)+(1-\alpha)\cdot\frac{\beta}{\|y_{w_t}\|}\log\pi_\theta(y_{w_t}\|x)-\frac{\beta}{\|y_l\|}\log\pi_\theta(y_l\|x)-\gamma\right)$
WRPO_{IPO}	$\left(\alpha\cdot\log\frac{\pi_\theta(y_{w_s}\|x)}{\pi_{\text{ref}}(y_{w_s}\|x)}+(1-\alpha)\cdot\log\frac{\pi_\theta(y_{w_t}\|x)}{\pi_{\text{ref}}(y_{w_t}\|x)}-\log\frac{\pi_\theta(y_l\|x)}{\pi_{\text{ref}}(y_l\|x)}-\frac{1}{2\tau}\right)^2$

Table 3: Results of WRPO combined with different preference optimization objectives.

Method	AlpacaEval-2		MT-Bench
	LC(%)	WR(%)	Overall
SimPO	53.9	49.9	7.39
IPO	51.1	52.4	7.67
$\text{WRPO}_{\text{SimPO}}$	55.8	51.8	7.42
WRPO_{IPO}	53.3	57.7	7.72

Table 4: Results of our WRPO implemented with varying numbers of source LLMs on AlpacaEval-2 and MT-Bench.

Num	AlpacaEval-2		MT-Bench
	LC(%)	WR(%)	Overall
1	48.9	50.3	7.29
2	52.3	50.4	7.54
5	53.5	53.8	7.42
10	55.9	58.0	7.63

+

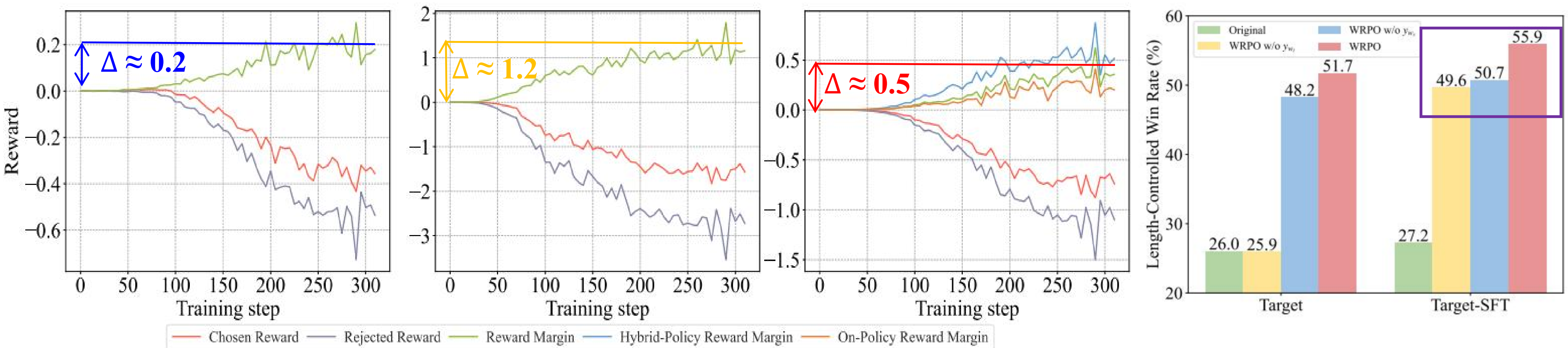
WRPO：Ablation Studies

$w/o\ y_{w_s}$

- Modest margin gain reveals a relatively **conservative** optimization process
- Exclusively relying on on-policy samples **limits model's exploration capability**

w/o y_{w_t}

- Faster margin gain reveals a more **aggressive** optimization behavior
- **Distribution shift** inherent in the hybrid setting may **compromise training stability**



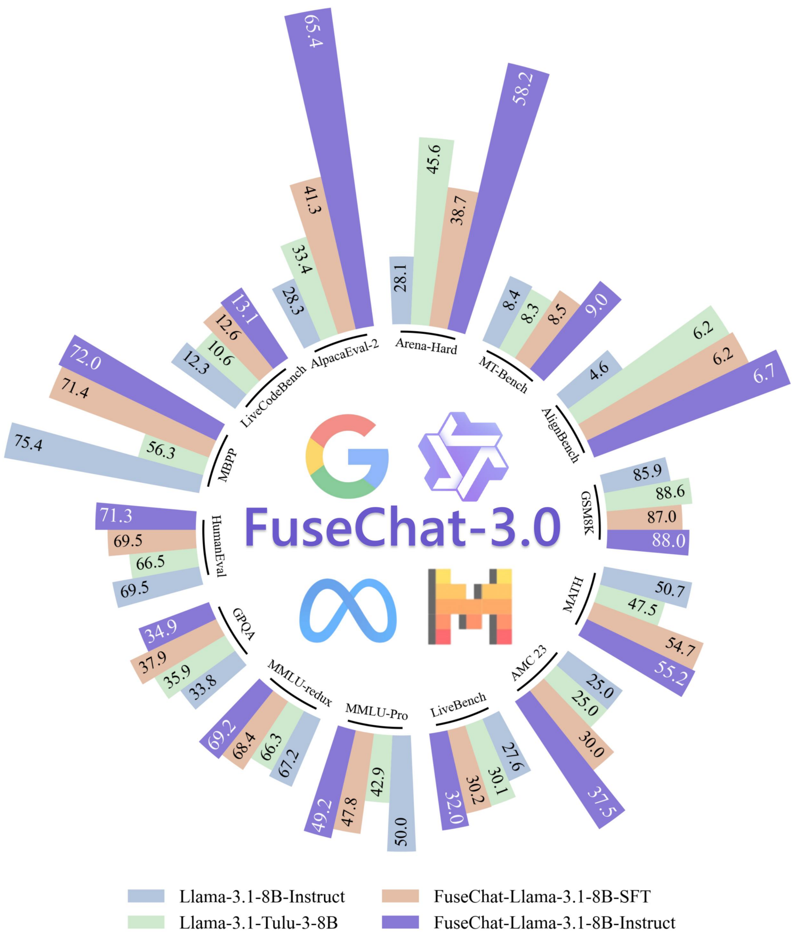
(a) DPO-on (b) DPO-hybrid (c) WRPO $\alpha = 0.5$

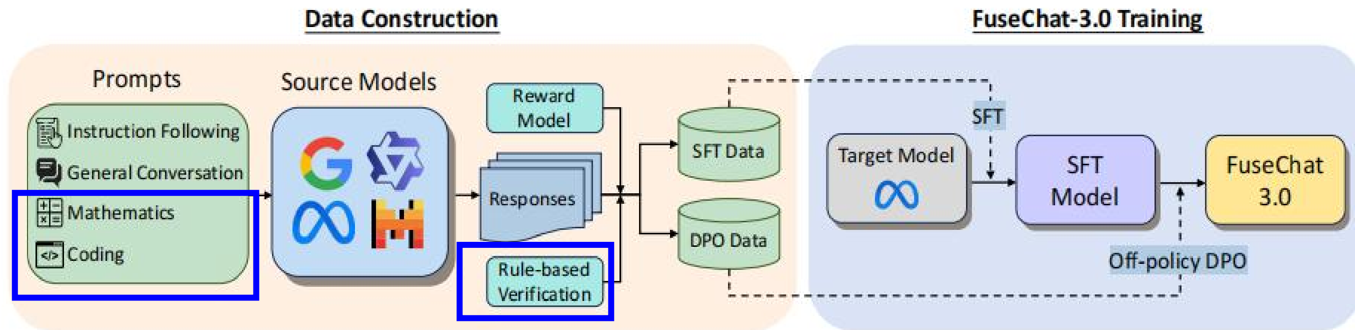
Conclusion

- We introduce Weighted-Reward Preference Optimization (WRPO) for the **implicit model fusion** of heterogeneous open-source LLMs, aiming to create a more **capable and robust target LLM**
- To address distributional deviations between source and target LLMs, we introduce a **progressive adaptation strategy** that gradually shifts reliance on preferred responses from the target LLM to the source LLMs
- Extensive experiments demonstrate that WRPO **consistently outperforms** existing knowledge fusion methods and various fine-tuning baselines

Future work

FuseChat-3.0: Preference Optimization Meets Implicit Model Fusion





Category	Benchmark	Llama-3.1-8B-Instruct			Llama-3.2-3B-Instruct			Llama-3.2-1B-Instruct		
		Base	SFT	FuseChat	Base	SFT	FuseChat	Base	SFT	FuseChat
Instruction Following	AlpacaEval-2 (LC %)	28.3	41.3	**65.4**	21.4	31.1	**54.0**	9.7	14.0	**25.3**
	Arena-Hard (WR %)	28.1	38.7	**58.2**	16.6	21.3	**30.2**	5.1	6.0	**8.6**
	MT-Bench	8.4	8.5	**9.0**	6.9	7.3	**7.7**	4.7	5.2	**5.7**
	AlignBench$_{v1.1}$	4.6	6.3	**6.7**	3.8	5.5	**5.9**	2.9	3.9	**4.3**
General	LiveBench$_{0831}$	27.6	30.2	**32.0**	23.4	24.5	**24.9**	14.0	13.9	**15.8**
	MMLU-Pro (0 shot, CoT)	**50.0**	47.8	49.2	39.3	**40.3**	40.3	**22.3**	21.5	21.3
	MMLU-redux (0 shot, CoT)	67.2	68.4	**69.2**	58.5	58.2	**59.0**	**43.7**	40.3	41.6
	GPQA-Diamond (0 shot, CoT)	33.8	**37.9**	34.9	29.8	33.3	**33.8**	21.2	**25.3**	24.2
Mathematics	GSM8K (0 shot, CoT)	85.9	87.0	**88.0**	82.0	**82.8**	82.0	46.3	**55.6**	54.5
	MATH (0 shot, CoT)	50.7	54.7	**55.2**	51.4	52.9	**53.1**	32.7	**34.7**	33.6
	AMC 23 (0 shot, CoT)	25.0	30.0	**37.5**	22.5	20.0	**35.0**	17.5	15.0	**20.0**
Coding	HumanEval (0 shot)	69.5	69.5	**71.3**	61.0	**62.8**	60.4	39.6	36.6	**40.2**
	MBPP (0 shot)	**75.4**	71.4	72.0	**68.5**	67.5	67.5	**49.5**	42.1	46.6
	LiveCodeBench$_{2408-2411}$	12.3	12.6	**13.1**	8.3	7.1	**9.0**	-	-	-
	Average	40.5	43.2	**47.3**	35.2	36.8	**40.2**	23.8	24.2	**26.3**

Thanks!