InfiFPO: Implicit Model Fusion via Preference Optimization in Large Language Models

[1] The Hong Kong Polytechnic University [c] InfiX.ai [3] Zhejiang University

Yanggan Gu[1 2], Yuanyi Wang[1], Zhaoyi Yan[2], Yiming Zhang[1], Qi Zhou[1], Fei Wu[3], Hongxia Yang[1 2] *

* Corresponding author.

Project Page: https://github.com/InfiXAI/InfiFPO

What is *Model Fusion*?

Model fusion is typically performed during the Supervised Fine-Tuning (SFT) stage. Given N source models $\{\mathcal{M}_i^s\}_{i=1}^N$ and a pivot model (also called target model) $\mathcal{M}^p$, the model fusion can be cast as a optimization problem:

$$\arg\max_{\mathcal{M}^p} \underbrace{\mathbb{E}_{\boldsymbol{x},\boldsymbol{y}\sim\mathcal{D}}\left[\log \mathcal{M}^p(\boldsymbol{y}|\boldsymbol{x})\right]}_{\text{(a) SFT loss}} + \beta \underbrace{\sum_{i=1}^{N} -\mathbb{D}_{\text{TKL}}\left[\mathcal{M}^p(\boldsymbol{y}|\boldsymbol{x}) \| \mathcal{M}_i^s(\boldsymbol{y}|\boldsymbol{x})\right]}_{\text{(b) Imitation of source models behavior}},$$

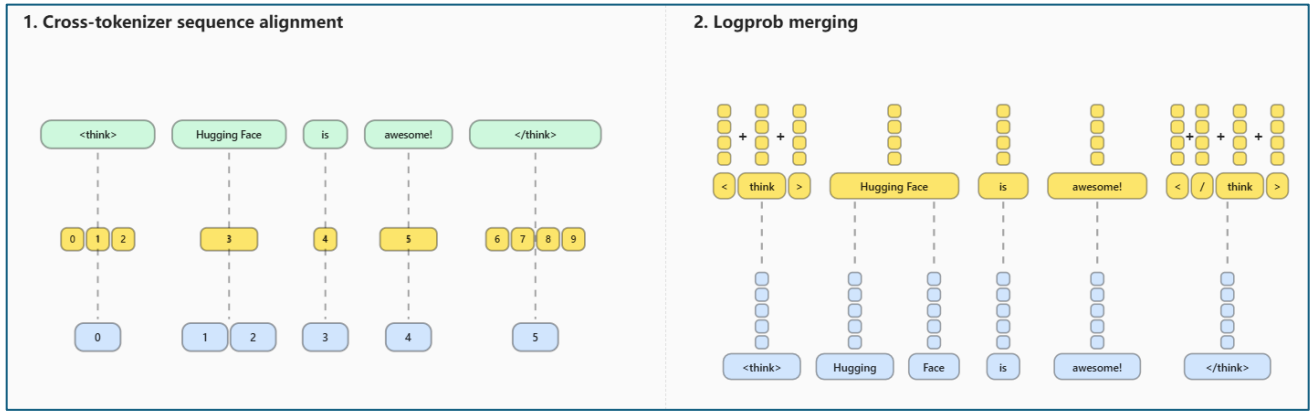
where each sample in the dataset $\mathcal{D}$ contains a prompt sequence $\boldsymbol{x}$ and its corresponding response sequence $\boldsymbol{y}$. $\mathbb{D}_{\text{TKL}}$ indicates token-level KL divergence.

What is the *challenges* for Model Fusion?

Current Model Fusion relies on Token-level KLD, which requires the two models share the same vocabularies. However, heterogeneous models (e.g., Llama and Qwen) usually employ incompatible tokenizers, leading to a fundamental **vocabulary conflict**. It contains two parts:

- **Token Misalignment**: Different tokenizers produce misaligned probability distributions over their vocabularies, making Token-level KLD inapplicable.
- **Sequence Misalignment**: After tokenization, the token sequence lengths for the same text may be different.

For these two misalignment, current work often introduce complex alignment processes, significantly increase the fusion cost.



https://huggingface.co/spaces/HuggingFaceH4/on-policy-distillation

Can we bypass *vocabulary conflict*?

Quick Review:

The objective of Reinforcement Learning from Human Feedback (RLHF) can be formalized as

$$\arg\max_{\mathcal{M}^\theta} \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D},\boldsymbol{y}\sim\mathcal{M}^\theta(\boldsymbol{y}|\boldsymbol{x})}[\underbrace{\boldsymbol{r}(\boldsymbol{x},\boldsymbol{y})}_{\text{(a) Rewards}}] + \beta \underbrace{[-\mathbb{D}_{\text{SKL}}[\mathcal{M}^\theta(\boldsymbol{y}|\boldsymbol{x})\|\mathcal{M}^{\text{ref}}(\boldsymbol{y}|\boldsymbol{x})]]}_{\text{(b) Preventing excessive deviation from } \mathcal{M}^{\text{ref}}}.$$

where $\boldsymbol{r}$ is a reward model that evaluates how good the response $\boldsymbol{y}$ generated by policy model $\mathcal{M}^\theta$ is. $\mathbb{D}_{\text{SKL}}$ indicates sequence-level KL divergence. Usually, the base reference model $\mathcal{M}^{\text{ref}}$ is the initial $\mathcal{M}^\theta$, and β is a parameter controlling the deviation from $\mathcal{M}^{\text{ref}}$.

Objective of Model Fusion:

$$\arg\max_{\mathcal{M}^p} \underbrace{\mathbb{E}_{\boldsymbol{x},\boldsymbol{y}\sim\mathcal{D}}[\log \mathcal{M}^p(\boldsymbol{y}|\boldsymbol{x})]}_{\text{(a) SFT loss}} + \beta \underbrace{\sum_{i=1}^{N} -\mathbb{D}_{\text{TKL}}[\mathcal{M}^p(\boldsymbol{y}|\boldsymbol{x})\|\mathcal{M}_i^s(\boldsymbol{y}|\boldsymbol{x})]}_{\text{(b) Imitation of source models behavior}},$$

Similar Formulation

If we combine model fusion and RLHF,
maybe we can bypass vocabulary conflict with Seq-level KLD.

How to *combine* RLHF and Model Fusion?

FuseRLHF:

$$\arg\max_{\mathcal{M}^p} \mathbb{E}_{\boldsymbol{x},\boldsymbol{y}\sim\mathcal{D}}\left[\boldsymbol{r}(\boldsymbol{x},\boldsymbol{y})\right]$$
$$\text{s.t. } \mathbb{E}_{\boldsymbol{x},\boldsymbol{y}\sim\mathcal{D}}\left[\mathbb{D}_{\text{SKL}}\left[\mathcal{M}_i^s(\boldsymbol{y}|\boldsymbol{x})||\mathcal{M}^p(\boldsymbol{y}|\boldsymbol{x})\right]\right] \leq \varepsilon, \quad \forall i \in \{1,...,N\},$$

where the initial pivot model is included in the source models. Each constraint keeps the pivot policy within an ε-ball (in KL divergence) of every teacher, thereby fusing their behaviour while still allowing preference alignment through the reward $\boldsymbol{r}$.

Can we transform RL into *efficient offline learning*?

Similar to DPO, after derivation we can get FusePO:

$$\mathcal{L}_{\text{FPO}}(\mathcal{M}^p; \{\mathcal{M}_i^s\}_{i=1}^N) = -\mathbb{E}_{(\boldsymbol{x}, \boldsymbol{y}_w, \boldsymbol{y}_l) \sim \mathcal{D}^p} \left[\log \sigma \left(\beta \log \frac{\mathcal{M}^p(\boldsymbol{y}_w|\boldsymbol{x})}{\mathcal{M}_{\text{fu}}^s(\boldsymbol{y}_w|\boldsymbol{x})} - \beta \log \frac{\mathcal{M}^p(\boldsymbol{y}_l|\boldsymbol{x})}{\mathcal{M}_{\text{fu}}^s(\boldsymbol{y}_l|\boldsymbol{x})} \right) \right]$$

where a fused source model $\mathcal{M}_{\text{fu}}^s(\boldsymbol{y}|\boldsymbol{x}) = \prod_{i=1}^N (\mathcal{M}_i^s(\boldsymbol{y}|\boldsymbol{x}))$

Further Improvement: *InfiFPO*

Naïve FPO. Directly training with FuseRLHF requires significant time and resources. To reduce that, we convert it to offline FPO:

$$\mathcal{L}_{\text{FPO}}(\mathcal{M}^p;\{\mathcal{M}_i^s\}_{i=1}^N) = -\mathbb{E}_{(\boldsymbol{x},\boldsymbol{y}_w,\boldsymbol{y}_l)\sim\mathcal{D}^p}\left[\log\sigma\left(\beta\log\frac{\mathcal{M}^p(\boldsymbol{y}_w|\boldsymbol{x})}{\mathcal{M}_{\text{fu}}^s(\boldsymbol{y}_w|\boldsymbol{x})} - \beta\log\frac{\mathcal{M}^p(\boldsymbol{y}_l|\boldsymbol{x})}{\mathcal{M}_{\text{fu}}^s(\boldsymbol{y}_l|\boldsymbol{x})}\right)\right]$$

Strategy I: Length Normalization. To address the ***length bias*** issue arising from varying token sequence lengths across models, we introduce:

$$\overline{\log}\,\mathcal{M}^{(\cdot)}(\boldsymbol{y}|\boldsymbol{x}) = \frac{1}{|\boldsymbol{y}|}\log\mathcal{M}^{(\cdot)}(\boldsymbol{y}|\boldsymbol{x})$$

Strategy II: Probability Clipping. To address the ***source model degradation*** issue caused by noisy outputs from poor-performing models, we propose:

$$\text{Clip}(\mathcal{M}_i^s(\boldsymbol{y}|\boldsymbol{x})) = \begin{cases}\max(\mathcal{M}_i^s(\boldsymbol{y}|\boldsymbol{x}),\mathcal{M}_{\text{init}}^p(\boldsymbol{y}|\boldsymbol{x})), & \text{if } \boldsymbol{y} \text{ is } \boldsymbol{y}_w,\\ \min(\mathcal{M}_i^s(\boldsymbol{y}|\boldsymbol{x}),\mathcal{M}_{\text{init}}^p(\boldsymbol{y}|\boldsymbol{x})), & \text{else.}\end{cases}$$

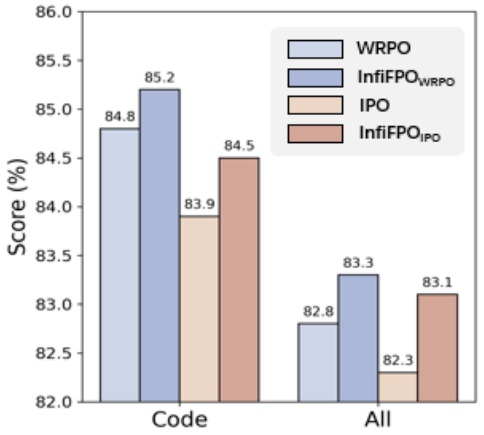
Final Objective: InfiFPO. Combining the strategies listed above, we have:

$$\mathcal{L}_{\text{InfiFPO}}(\mathcal{M}^p;\{\mathcal{M}_i^s\}_{i=1}^N) = -\mathbb{E}_{(\boldsymbol{x},\boldsymbol{y}_w,\boldsymbol{y}_l)\sim\mathcal{D}^p}\left[\log\sigma\left(\beta\overline{\log}\frac{\mathcal{M}^p(\boldsymbol{y}_w|\boldsymbol{x})}{\mathcal{M}_{\text{fu}}^{\text{sclip}}(\boldsymbol{y}_w|\boldsymbol{x})} - \beta\overline{\log}\frac{\mathcal{M}^p(\boldsymbol{y}_l|\boldsymbol{x})}{\mathcal{M}_{\text{fu}}^{\text{sclip}}(\boldsymbol{y}_l|\boldsymbol{x})}\right)\right]$$

Main Experimental Results

Models	Math			Code		General Reasoning			InstFol	Text Reasoning		Avg	Model Size	GPU Hours
	GSM8K	MATH	ThmQA	MBPP	HEval	BBH	ARC	MMLU	IFEval	DROP	HS			
Pivot Model														
Phi-4	87.41	80.04	51.12	75.40	83.54	68.84	93.90	85.62	77.34	88.67	87.62	79.95	14B	∼1.0M
Source Models														
Qwen2.5-Instruct	91.13	78.16	47.25	81.70	83.54	77.59	92.20	80.22	85.01	85.56	88.28	80.97	14B	∼1.8M
Mistral-Small	92.42	69.84	48.50	68.80	84.15	81.59	91.86	81.69	82.25	86.52	**91.84**	79.95	24B	∼1.6M
Qwen2.5-Coder	89.16	74.18	38.88	**85.40**	**90.90**	75.40	89.49	75.08	74.70	84.34	79.83	77.94	14B	∼1.8M
Gemma-3-Instruct	**93.71**	**82.90**	49.62	72.60	82.32	**85.70**	71.19	77.61	**90.77**	86.43	83.34	79.65	12B	-
Qwen2.5-Math	92.27	81.70	20.25	1.40	46.34	33.12	65.76	40.20	35.49	81.96	25.57	47.64	7B	∼0.5M
Model Fusion Methods														
FuseLLM*	90.24	80.25	53.52	79.28	84.00	77.62	92.08	83.92	78.56	88.74	87.81	81.46	14B	225
FuseChat*	91.21	77.52	51.88	81.80	84.15	83.37	93.56	84.23	78.90	89.23	87.42	82.12	14B	650
InfiFusion*	90.07	80.94	55.62	81.80	83.54	80.94	94.24	**85.81**	76.02	**89.27**	87.91	82.38	14B	160
Preference Optimization Methods														
SFT	88.70	79.58	55.12	78.20	86.59	74.66	93.56	84.36	80.06	88.72	87.75	81.57	14B	15
SFT-DPO	89.76	80.02	**57.88**	82.50	84.76	77.86	**94.58**	84.27	81.89	88.56	87.31	82.67	14B	50
SFT-IPO	90.45	80.18	55.25	82.50	85.37	77.13	94.24	84.08	80.94	88.67	87.36	82.38	14B	50
SFT-WRPO	89.92	80.02	**57.88**	83.10	86.59	78.18	94.24	83.98	81.18	88.41	87.30	82.80	14B	57
InfiFPO														
InfiFPO*	89.92	79.88	57.00	82.00	85.98	81.26	94.24	83.33	80.46	88.68	87.36	82.74	14B	55
InfiFPO	90.07	80.10	57.25	82.50	87.80	82.02	94.24	84.27	82.25	88.83	87.29	**83.33**	14B	58

+ 3.4% on Avg.
+ 6.1% In ThmQA
+7.1% In MBPP
+4.9% In IFEval



① **InfiFPO consistently outperforms existing model fusion and preference optimization methods.**

② **InfiFPO's versatility enables effective integration with various PO objectives, enhancing performance.**

Thanks!