Advancing LLM Reasoning Generalists with Preference Trees

Lifan Yuan*[1,2], Ganqu Cui*[1], Hanbin Wang*[3], Ning Ding[1], Xingyao Wang[2], Boji Shan, Zeyuan Liu[1], Jia Deng, Huimin Chen[1], Ruobing Xie, Yankai Lin, Zhenghao Liu, Bowen Zhou[1], Hao Peng[2], Zhiyuan Liu[1], Maosong Sun[1]

[1]Tsinghua University [2]University of Illinois Urbana-Champaign [3]Peking University

lifan4@illinois.edu cgq22@mails.tsinghua.edu.cn wanghanbin95@stu.pku.edu.cn

Eurus: Reasoning Generalists Covering Multiple Tasks.

Eurus variants achieve the **best** overall performance among open-source models of similar sizes *at the time of its release*, outperforming specialized models in corresponding domains in many cases. Particularly, Eurus-70B **beats GPT-3.5 Turbo in reasoning** through a benchmarking across 12 tests (**mostly OOD**) covering five tasks.

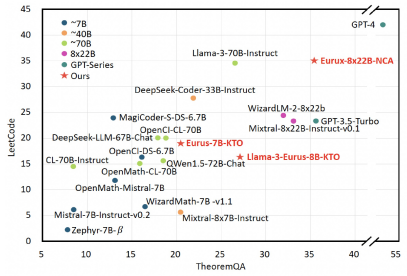

Figure 1: Evaluation results on LeetCode and TheoremQA, two challenging OOD coding and math benchmarks with only test sets. Our EURUS-7B is comparable with baselines that are 10x larger and EURUX-8X22B is the only one on par with GPT-3.5 Turbo.

Table 3: Overall performance. All test sets except MATH are out-of-distribution to our models and most baselines. MAmmoTH, OpenChat, and Starling-LM have been trained on TheoremQA test sets. We ~~strikethrough~~ the contaminated numbers.

Model	Coding			Math					Reasoning	Ins-Following	Multi-Turn		Avg.
	HumanEval	MBPP	LeetCode	GSM-Plus	MATH	TheoQA	SVAMP	ASDiv	BBH	IFEval	Code	Math	
~7B													
Mistral-7B-Instruct-v0.2	39.0	30.8	6.1	15.7	9.5	8.5	42.9	49.5	62.4	44.4	7.4	26.2	28.5
Zephyr-7B-β	29.3	35.8	2.2	23.3	5.0	7.8	19.1	28.0	61.8	39.7	5.2	16.9	22.8
OpenChat-3.5-1210	64.0	61.7	11.7	46.7	28.1	10.1	75.4	77.0	67.0	50.3	21.3	32.4	46.2
Starling-LM-7B-α	46.3	51.1	8.9	23.7	21.5	12.0	26.3	39.8	67.1	26.1	18.4	28.9	30.8
Magicoder-S-DS-6.7B	75.6	**70.4**	**23.9**	16.4	19.9	13.1	61.6	62.8	57.0	21.1	**27.9**	8.0	38.1
WizardCoder-33B-V1.1	**76.8**	66.2	16.1	41.5	31.6	16.1	74.5	79.8	53.9	22.6	5.9	1.3	40.5
MAmmoTH-7B-Mistral	24.4	42.4	7.2	40.1	36.0	26.3	60.7	72.3	57.7	34.9	3.7	6.7	34.4
WizardMath-7B-v1.1	50.0	53.9	6.7	54.6	30.0	16.5	57.8	73.5	64.4	22.6	16.2	8.9	37.9
OpenMath-Mistral-7B	33.5	46.6	11.7	**59.4**	**39.1**	13.1	83.4	79.8	58.6	15.0	2.9	5.3	37.4
EURUS-7B-SFT	55.5	59.1	20.0	52.1	32.6	20.0	82.2	84.1	64.6	44.0	15.4	28.4	46.5
+ DPO	50.6	52.1	8.3	51.0	28.3	20.9	78.7	83.8	55.0	42.5	20.6	32.4	44.5
+ KTO	56.1	58.6	18.9	55.0	33.2	20.6	84.4	85.0	67.6	43.1	19.1	**43.6**	48.8
+ NCA	55.5	60.2	14.4	54.9	34.2	20.9	**84.6**	85.4	64.3	42.7	21.3	38.7	48.1
LLAMA-3-EURUS-8B-SFT	51.2	57.9	17.2	50.7	32.0	21.3	82.2	83.7	**72.4**	47.1	18.4	24.5	46.6
+ DPO	43.9	50.1	11.7	45.3	26.8	21.4	54.1	67.5	71.3	**56.7**	21.3	39.2	42.4
+ KTO	51.8	58.1	15.6	54.8	34.2	24.9	80.1	86.7	71.7	50.6	26.5	37.4	49.4
+ NCA	50.6	60.4	15.6	55.2	34.8	**25.4**	79.9	**87.5**	71.7	56.2	21.3	36.3	**49.6**
~40B													
Mixtral-8x7B-Instruct	50.6	50.1	5.6	**49.6**	**25.9**	20.4	66.4	68.8	**73.5**	**48.8**	12.5	**37.3**	42.5
DeepSeek-Coder-33B-Ins	**82.3**	**73.9**	**27.8**	29.5	20.2	**21.9**	**75.2**	**85.0**	61.5	26.1	**35.3**	21.8	**46.7**
~70B													
CodeLLaMA-70B-Instruct	56.7	58.6	14.4	34.9	12.0	8.4	63.5	70.1	74.5	24.0	3.7	14.2	36.3
DeepSeek-LM-67B-Chat	70.7	65.7	20.0	65.0	41.0	17.9	74.0	78.9	82.7	52.7	30.9	41.8	53.5
QWen1.5-72B-Chat	71.3	56.9	15.6	65.4	43.4	18.5	79.5	79.1	78.0	53.4	27.2	38.2	52.2
OpenCI-CL-70B	**77.4**	**71.7**	20.0	46.1	29.2	18.8	76.1	79.4	66.7	26.8	30.9	12.0	46.3
OpenMath-CL-70B	39.0	52.6	15.0	62.2	45.9	15.9	86.6	82.8	59.9	15.7	14.0	0.4	40.8
WizardLM-2-8x22B	72.0	64.2	24.4	57.0	**50.9**	32.0	81.2	82.2	85.3	**68.9**	13.2	43.2	56.2
Mixtral-8x22B-Instruct-v0.1	76.2	68.7	23.3	51.2	49.6	33.1	88.3	**92.4**	**86.0**	67.1	**39.0**	59.7	61.2
EURUS-8X22B-KTO	71.3	68.9	29.4	**68.3**	48.1	35.3	91.5	90.6	83.6	67.1	38.2	57.5	62.5
EURUS-8X22B-NCA	75.0	69.7	**35.0**	68.1	49.0	**38.5**	**91.6**	92.1	83.5	67.1	33.1	**63.0**	**63.6**
Proprietary Models													
GPT-3.5 Turbo	76.8	82.5	23.3	61.2	37.8	35.6	83.0	90.6	70.1	56.6	29.4	36.9	57.0
GPT-4	85.4	**83.5**	**41.8**	**85.6**	**69.7**	**52.4**	**94.8**	**92.6**	**86.7**	**79.7**	**59.6**	**65.8**	**74.8**

Eurus-RM: (was) State-of-the-Art 7B Reward Model.

We also train Eurus-RM-7B with a **new RM objective**, which is to **directly increase the reward of the chosen actions and vice versa**. In many cases, our RM achieves better correlation with humans than baselines, and can improve LLMs' reasoning performance by a large margin through reranking.

$$\mathcal{L}_{\text{ULTRAINTERACT}} = \underbrace{-\log\left(\sigma\left(r_\theta\left(x, y_c\right) - r_\theta\left(x, y_r\right)\right)\right)}_{\mathcal{L}_{BT}\text{ optimize relative rewards}} \underbrace{-\log\left(\sigma\left(r_\theta\left(x, y_c\right)\right)\right) - \log\left(\sigma\left(-r_\theta\left(x, y_r\right)\right)\right)}_{\mathcal{L}_{\mathcal{R}}:\text{ increase } r_\theta(x,y_c) \text{ and decrease } r_\theta(x,y_r)}$$

Table 4. Results on reward modeling benchmarks. UF: UltraFeedback; US: UltraSafety. The best performance in each benchmark is in bold and the second best one is underlined. Most baseline results are from (Jiang et al., 2023b) and (Lambert et al., 2024).

Model	Reward Bench					AutoJ				MT-Bench
	Chat	Chat-Hard	Safety	Reasoning	Avg.	Code	Math	Others	Overall	
PairRM	90.2	53.0	31.5	60.0	58.7	58.3	52.8	58.9	59.1	59.0
Starling-RM-7B	98.0	43.4	88.6	74.6	76.2	59.2	47.2	61.4	60.8	56.8
Starling-RM-34B	96.9	59.0	89.9	90.3	**84.0**	65.8	54.2	62.3	62.6	60.4
UltraRM-13B	96.1	55.3	45.8	82.0	69.8	55.0	43.1	59.6	59.9	56.0
GPT-3.5 Turbo	-	-	-	-	-	36.6	40.3	41.2	42.7	57.1
GPT -4	-	-	-	-	-	69.2	51.4	61.4	61.9	63.9
EURUS-RM-7B	96.5	65.3	80.7	87.0	82.4	87.5	82.5	78.0	80.7	79.4
w/o $\mathcal{L}_{\text{DR}}$	96.4	59.9	79.5	77.5	78.3	83.8	82.5	78.9	80.7	79.3
w/o $\mathcal{L}_{\text{BT}}$	96.8	58.5	83.8	84.2	80.8	88.8	92.5	79.4	**81.9**	**79.6**
w/o US	96.5	66.2	67.7	81.7	73.3	87.5	90.0	79.2	81.8	79.2
w/o UF + US	95.1	61.1	63.7	73.4	78.0	73.8	80.0	71.7	72.8	73.0

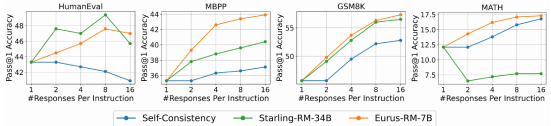

Figure 4. Results on reranking Mistral-7B-Instruct-v0.2's responses. Full results in Table 9.

The Secret Behind Eurus?
UltraInteract: Preference Trees For Reasoning.

UltraInteract collects a preference tree for each instruction, with **the instruction being the root** and **each action a node**, two nodes at each turn. All nodes of correct actions can be used for SFT. Paired **correct** and **incorrect** trajectories can be used for preference learning and reward modeling.

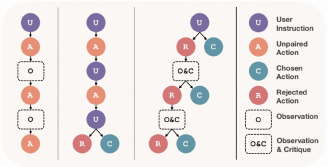

Figure 2. Left: CodeActInstruct (Wang et al., 2024) and Code-Feedback (Zheng et al., 2024); Middle: HH-RLHF (Bai et al., 2022); Right: ULTRAINTERACT. Each instruction in ULTRAINTERACT is constructed as a preference tree.

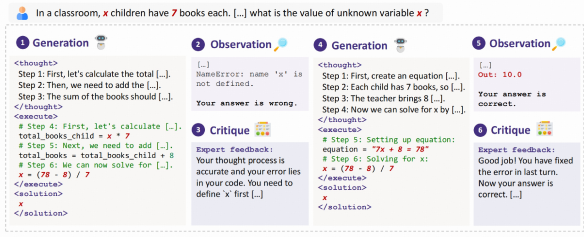

Figure 3. An illustrative example of an ULTRAINTERACT trajectory over two turns. In each turn, the actor model generates step-by-step reasoning chains, and the environment and the critique model provide observations and textual critique respectively.

Explicit Reward as A Proxy?
Hypothesis for Preference Learning in Reasoning.

We find that **KTO and NCA can improve** model performance on top of SFT. Inspecting the rewards, **they optimize not only reward margins but also absolute values.** We assume this behavior is necessary in preference learning for reasoning, where **correct answers are more important**.

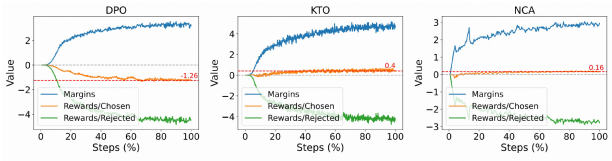

Figure 5. Reward patterns of EURUS-7B preference learning with DPO, KTO, and NCA. For all algorithms, the rewards of rejected data keep decreasing and the margins between chosen and rejected data keep increasing. However, the rewards of chosen data decrease below zero in DPO while keeping increasing and staying positive in KTO and NCA. The absolute values of the reward in the last step (in red) of the three algorithms positively correlate with their performance in Table 3.

Ablation Study: Tree Structure Benefits Multi-turn Interaction Ability

We **decomposed a multi-turn tree into multiple single-turn pairs** and trained Llama-3-Eurus-8B on single-turn pairwise data. Compared to training on single-turn pairs, **training on multi-turn trees enjoys huge benefits on multi-turn interaction ability** and slightly improves the overall performance.

Table 15: Model performance when trained on multi-turn and single-turn pairs from ULTRAINTERACT.

Model	Coding			Math					Reasoning	Ins-Following	Multi-Turn		Avg.
	HumanEval	MBPP	LeetCode	GSMPLUS	MATH	TheoremQA	SVAMP	ASDiv	BBH (CoT)	IFEval	Code	Math	
Llama-3-Eurus-8B-SFT	51.2	57.9	17.2	50.7	32.0	21.3	82.2	83.7	72.4	47.1	18.4	24.5	46.6
+ DPO	43.9	50.1	11.7	45.3	26.8	21.4	54.1	67.5	71.3	56.7	21.3	39.2	42.4
+ DPO (Single Turn)	49.4	51.9	9.4	54.5	27.1	22.8	76.9	85.1	72.2	57.1	22.0	37.0	47.1
+ KTO	51.8	58.1	15.6	54.8	34.2	24.9	80.1	86.7	71.7	50.6	26.5	37.4	49.4
+ KTO (Single Turn)	53.7	59.1	14.4	54.8	30.7	23.1	77.8	86.2	72.1	49.9	22.8	33.0	48.1
+ NCA	50.6	60.4	15.6	55.2	34.8	25.4	79.9	87.5	71.7	56.2	21.3	36.3	49.6
+ NCA (Single Turn)	53.7	55.9	16.1	55.4	30.5	25.4	79.3	87.5	72.2	54.2	17.7	35.5	48.6