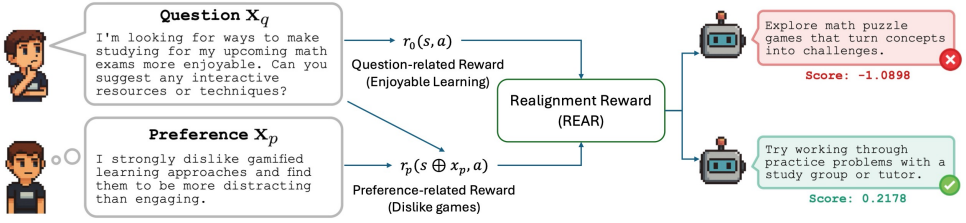
REAR: Test-time Preference Realignment through Reward Decomposition

Paper Code

Fuxiang Zhang[1], Pengcheng Wang[2], Chenran Li[2], Yi-Chen Li[3], Yuxin Chen[2], Lang Feng[1], Chenfeng Xu[2], Masayoshi Tomizuka[2], Bo An[1]

[1] Nanyang Technological University [2] UC Berkeley [3] Nanjing University

Controllable Realignment for General LLMs



- Human preferences are diverse and personal, yet a pretrained LLM bakes in a single, fixed preference balance from its training data.
- General LLMs are good at answering questions but bad at recognizing user preference.

How can we adjust the balance between question & preference of an LLM?

- Re-training / fine-tuning with specific preference level – costly
- Test-time alignment – no verifiable approach (diff. from math, code, …)

Key Insights

- The base model is already an implicit reward model.
- Under the following reward decomposition:
  - s: generation state wo. preference (question only)
  - $s \oplus x_p$: generation state with preference
  - The base model maximizes the question-related reward wo. preference and maximizes a combination of question-related and preference-related reward with some α

Prop. 3.1 — latent reward structure

$$r(s \oplus x_p, a) = \underbrace{r_0(s,a)}_{\text{question}} + \alpha \underbrace{r_p(s \oplus x_p, a)}_{\text{preference}}$$

REAR: Realignment Reward

- How to rebalance two reward terms? Adjusting α to $\hat{\alpha}$ results in a new reward

Token-level Reward

Lemma 3.2. *The realignment reward $r_{\text{REAR}}(s \oplus x_p, a)$ is equal to*

$$r_{\text{REAR}}(s \oplus x_p, a) = \frac{(\alpha - \hat{\alpha})\,\beta}{\alpha} \log \pi(a \mid s) + \frac{\hat{\alpha}\beta}{\alpha} \log \pi(a \mid s \oplus x_p) + Z(s) - \gamma Z(s'), \quad (8)$$

where the state-dependent term $Z(s) = \left(1 - \frac{\hat{\alpha}}{\alpha}\right) V^{\pi}(s) + \frac{\hat{\alpha}}{\alpha} V^{\pi}(s \oplus x_p)$.

Cumulative Rewards over Trajectory

$$\begin{aligned} R_{\text{REAR}}(\tau) &= \sum_{s_t, a_t} \gamma^t r_{\text{REAR}}(s_t \oplus x_p, a_t) \\ &= \sum_{s_t, a_t}^{\infty} \gamma^t \Big((1-\lambda)\beta \log \pi(a_t|s_t) \\ &\quad + \lambda\beta \log \pi(a_t|s_t \oplus x_p) \Big) + Z(s_0), \end{aligned}$$

REAR Score (Drop state-only terms)

$$S_{\text{REAR}}(\tau) = \sum_{t=0}^{T} \gamma^t \Big((1-\lambda) \log \pi(a_t|s_t) + \lambda \log \pi(a_t|s_t \oplus x_p) \Big).$$

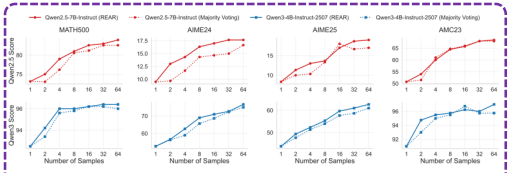
Now we have two test-time methods using S_{REAR} :

(1) Best-of-N (2) Tree search (DVTS) (calculated from sub-trajectories)

Experiment Results

Benchmark	DVTS w/ REAR (Ours)	BoN w/ REAR (Ours)	BoN w/ GenRM	Amulet	LA	Greedy
PrefEval Scores						
Explicit Preference	**77.7**$^{\uparrow 10.7}$	74.1$^{\uparrow 7.1}$	69.0$^{\uparrow 2.0}$	68.5$^{\uparrow 1.5}$	64.2$^{\downarrow 2.8}$	67.0
Implicit Choice	**78.6**$^{\uparrow 7.1}$	78.2$^{\uparrow 6.7}$	74.7$^{\uparrow 3.2}$	70.4$^{\downarrow 1.1}$	78.0$^{\uparrow 6.5}$	71.5
Implicit Preference	**19.1**$^{\uparrow 7.1}$	16.2$^{\uparrow 4.2}$	12.9$^{\uparrow 0.9}$	13.1$^{\uparrow 1.1}$	12.8$^{\uparrow 0.8}$	12.0
Multifaceted Bench	**76.8**$^{\uparrow 1.5}$	76.3$^{\uparrow 1.0}$	76.1$^{\uparrow 0.8}$	75.4$^{\uparrow 0.1}$	75.6$^{\uparrow 0.3}$	75.3
Ping-Pong Bench	3.03$^{\uparrow 0.06}$	**3.07**$^{\uparrow 0.10}$	3.01$^{\uparrow 0.04}$	2.87$^{\downarrow 0.10}$	3.01$^{\uparrow 0.04}$	2.97

Preference: General, Role-Play, …



Preference: Do good math

Preference: Read the image without hallucinations

Method	MMHal Score (↑)	Hallucination (↓)
BoN w/ REAR	**84.20**	**21.87**
BoN w/ GenRM	80.21	23.96
Greedy	78.99	28.12