Amulet: ReAlignment During Test Time for Personalized Preference Adaptation of LLMs

Zhaowei Zhang

One-Sentence Summary

We introduce Amulet, a training-free framework that enables real-time optimization to satisfy user's personalized preferences for LLMs at test time.

OUTLINE

01 Motivations

Motivations

- AI alignment is a heated topic lately.
- Current alignment method can not process human preference changing with culture, value, or time.
- The alignment target set by the developers is usually different from the users.

Therefore, efficient and dynamic alignment at inference time is important!

02 Background

Background

Alignment at Training Time

- RLHF
- DPO
- KTO

...

Assisted Inference Methods

- EFT, DeRa
- BoN
- Aligner, BP

...

Need fine-tuning at each time

Tuning-free Methods (self-distillation)

- URIAL
- RAIN

...

Tuning-free Methods (Efficient)

- Beam Search
- Linear Alignment

...

High cost for the former,
Less effective for the latter

03 Methodology

Methodology

- Optimization Objective:

$$\pi^*(a) = \arg\max_{\pi \in \Pi} \mathbb{E}_{a \sim \pi(\cdot|s,s_0)} r(a|s_0, s),$$

- Modelling through the FTRL framework:

$$\pi_{t+1} = \arg\max_{\pi \in \Pi} \left[\sum_{i=1}^{t} \mathcal{U}_i(\pi) - \frac{1}{\eta} D_{\mathrm{KL}}(\pi \| \pi_t) \right].$$

We can freely choose the utility, here we use:

$$\mathcal{U}_t(\pi) := u_t(\pi) - \lambda D_{\mathrm{KL}}(\pi \| \pi_1).$$

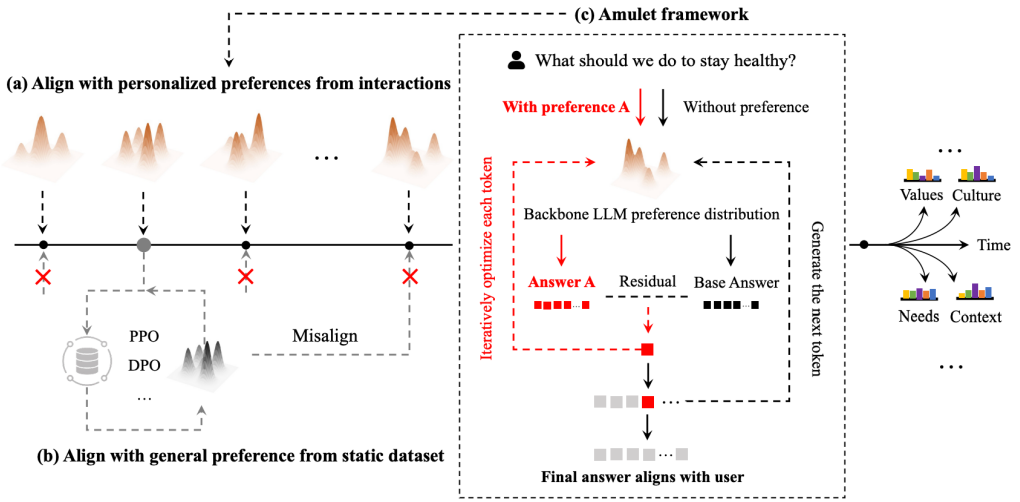
- Get the closed-form solution (Main contribution) :

$$\pi_{t+1}(a) \propto \exp\left(\frac{1}{t\lambda\eta + 1} \left(\eta \sum_{i=1}^{t} u_i(a) + \lambda\eta t \log \pi_1(a) + \log \pi_t(a) \right) \right).$$

Algorithm 1 Decoding Process with Amulet

Require: LLM for generating policy; basic prompt p_{base}; preference prompt p_{pref}; current generated sequence s, iteration number T; maximum new token number M; parameters α, λ, and η; blank string s

repeat
  generate $\pi_1(a) = P_{\text{LLM}}(a|p_{\text{base}}, p_{\text{pref}}, s)$, $\pi_{\text{base}}(a) = P_{\text{LLM}}(a|p_{\text{base}}, s)$ with the given LLM
  for $t = 1, 2, \ldots, T-1$ **do**
    calculate $u_t(\pi_t(a)) := \alpha(\log \pi_t(a) - \log \pi_{\text{base}}(a))$
    update the policy with the iteration given by Equation 6
  end for
  get the optimized policy $\pi^*(a) \leftarrow \pi_T(a)$
  sample the generated token a with $\pi^*(a)$
  update the current sequence $s \leftarrow s + a$
until the length of s reaches M **or** generation is ended
return the full generation sequence s



04 Experiment

Experiment Settings

Dataset

- Truthful QA
- Personal Preference Eval
- UltraChat
- HelpSteer

Eval-Metrics

- ArmoRM-8B reward model score
- GPT-4o win rate

Tested LLM

- Qwen2-7B-Instruct
- Llama-2-7B-chat
- Llama-3.1-8B-Instruct
- Mistral-7B-Instruct-v0.2

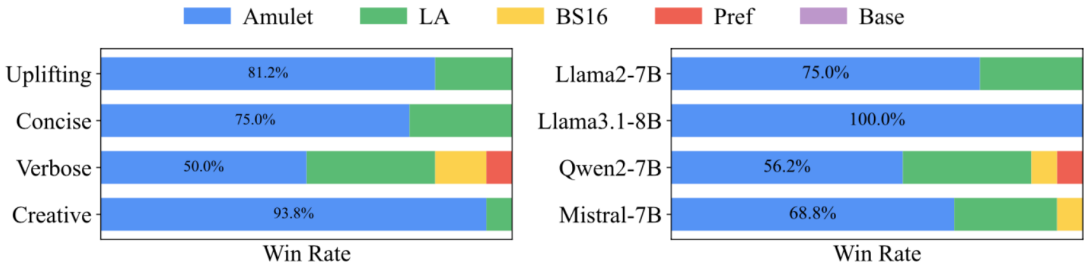
Baselines

- Base
- Pref
- Beam Search 16
- Linear Alignment

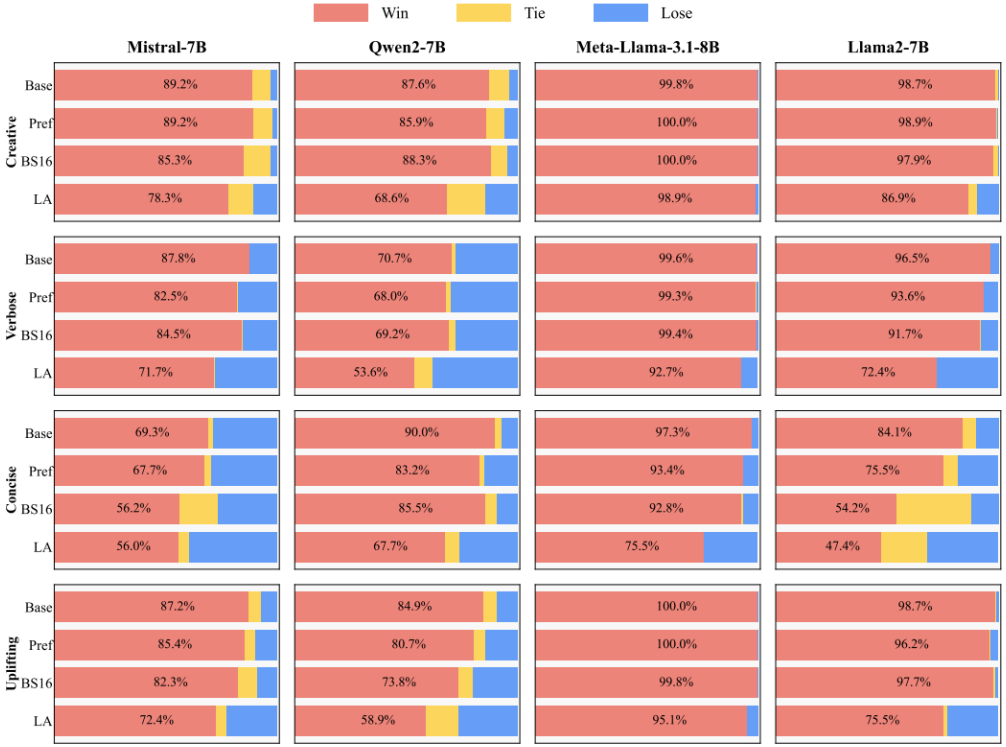
Experiment Results

ArmoRM-8B Score

Model	Dataset	Creative					Verbose					Concise					Uplifting				
		Base	Pref	BS16	LA	Amulet	Base	Pref	BS16	LA	Amulet	Base	Pref	BS16	LA	Amulet	Base	Pref	BS16	LA	Amulet
Mistral-7B	HelpSteer	0.30	0.30	0.34	0.36	**0.39**	0.27	0.27	**0.31**	0.31	0.30	0.41	0.42	0.50	0.52	**0.52**	0.33	0.33	0.39	0.40	**0.41**
	Personal	0.34	0.34	0.35	0.38	**0.42**	0.30	0.30	0.30	0.30	**0.30**	0.47	0.49	0.50	**0.54**	0.53	0.41	0.42	0.42	0.45	**0.46**
	Truthful QA	0.32	0.33	0.34	0.38	**0.41**	0.30	0.31	0.31	**0.33**	0.32	0.41	0.44	0.47	**0.51**	0.49	0.36	0.38	0.39	0.47	**0.47**
	Ultra Chat	0.34	0.35	0.35	0.36	**0.38**	0.31	0.31	0.31	**0.32**	0.31	0.45	0.46	0.47	0.49	**0.51**	0.38	0.39	0.39	0.41	**0.42**
	Average	0.32	0.33	0.34	0.37	0.40	0.30	0.30	0.31	0.32	0.31	0.43	0.45	0.48	0.52	0.51	0.37	0.38	0.40	0.43	0.44
Qwen2-7B	HelpSteer	0.34	0.34	0.35	0.35	**0.36**	0.31	0.32	**0.33**	0.33	0.30	0.43	0.48	0.50	0.57	**0.59**	0.38	0.38	0.39	0.39	**0.41**
	Personal	0.33	0.34	0.34	0.37	**0.41**	0.31	**0.31**	0.31	0.30	0.28	0.41	0.48	0.49	0.53	**0.54**	0.40	0.42	0.42	**0.43**	0.42
	Truthful QA	0.32	0.33	0.33	0.34	**0.36**	0.30	0.31	0.32	**0.33**	0.32	0.41	0.46	0.50	**0.54**	0.51	0.36	0.38	0.39	0.44	**0.45**
	Ultra Chat	0.34	0.34	0.34	0.35	**0.36**	0.31	0.32	0.32	**0.32**	0.31	0.40	0.45	0.46	0.54	**0.57**	0.38	0.39	0.39	**0.40**	0.39
	Average	0.33	0.34	0.34	0.35	0.37	0.31	0.32	0.32	0.32	0.30	0.41	0.47	0.49	0.55	0.55	0.38	0.39	0.40	0.42	0.42
Llama-3.1-8B	HelpSteer	0.33	0.34	0.36	0.44	**0.50**	0.30	0.31	0.33	0.36	**0.41**	0.40	0.43	0.45	0.53	**0.57**	0.36	0.37	0.39	0.45	**0.50**
	Personal	0.35	0.36	0.36	0.46	**0.62**	0.31	0.31	0.31	0.35	**0.49**	0.39	0.44	0.45	0.53	**0.67**	0.42	0.44	0.43	0.49	**0.61**
	Truthful QA	0.31	0.33	0.33	0.41	**0.56**	0.29	0.29	0.31	0.34	**0.44**	0.37	0.40	0.42	0.49	**0.52**	0.34	0.36	0.37	0.43	**0.49**
	Ultra Chat	0.33	0.34	0.34	0.42	**0.57**	0.31	0.32	0.32	0.36	**0.41**	0.38	0.41	0.41	0.48	**0.53**	0.37	0.38	0.38	0.44	**0.48**
	Average	0.33	0.34	0.35	0.43	0.56	0.30	0.31	0.32	0.35	0.44	0.38	0.42	0.43	0.51	0.57	0.37	0.39	0.39	0.45	0.52
Llama-2-7B	HelpSteer	0.32	0.33	0.35	**0.37**	0.36	0.28	0.29	0.31	**0.31**	0.30	0.39	0.42	0.44	**0.48**	0.47	0.36	0.37	0.39	**0.40**	0.38
	Personal	0.32	0.33	0.32	0.39	**0.45**	0.26	0.27	0.27	0.29	**0.32**	0.38	0.41	0.43	0.49	**0.53**	0.40	0.41	0.41	0.45	**0.49**
	Truthful QA	0.30	0.32	0.31	0.36	**0.41**	0.27	0.28	0.28	0.30	**0.32**	0.30	0.35	0.37	0.44	**0.49**	0.34	0.36	0.36	0.40	**0.44**
	Ultra Chat	0.32	0.33	0.34	0.37	**0.41**	0.29	0.30	0.30	0.32	**0.34**	0.39	0.43	0.43	0.47	**0.50**	0.37	0.38	0.39	0.40	**0.43**
	Average	0.32	0.33	0.33	0.37	0.41	0.28	0.29	0.29	0.30	0.32	0.36	0.40	0.42	0.47	0.50	0.37	0.38	0.39	0.41	0.44



GPT-4o Win Rate



More Models

Model	Creative					Verbose					Concise					Uplifting				
	Base	Pref	BS16	LA	Amulet	Base	Pref	BS16	LA	Amulet	Base	Pref	BS16	LA	Amulet	Base	Pref	BS16	LA	Amulet
Qwen2-0.5B	0.27	**0.30**	0.28	0.27	0.29	0.23	**0.25**	0.25	0.21	0.23	0.31	0.34	0.19	0.33	**0.37**	0.33	0.35	0.27	0.36	**0.41**
Llama-3.2-1B	0.28	0.27	0.34	0.34	**0.35**	0.23	0.22	0.30	0.33	**0.36**	0.31	0.29	**0.40**	0.38	0.39	0.34	0.32	**0.41**	0.39	0.38
Llama-2-13B	0.28	0.31	0.32	0.42	**0.46**	0.24	0.25	0.27	0.30	**0.33**	0.34	0.38	0.42	0.49	**0.54**	0.35	0.37	0.40	0.45	**0.48**
Llama-2-70B	0.33	0.33	0.33	0.39	**0.43**	0.28	0.28	0.28	0.32	**0.33**	0.44	0.51	0.50	0.57	**0.62**	0.40	0.41	0.41	0.44	**0.47**

THANKS