Multi-Personality Generation of LLMs at Decoding-time

WeChat Paper

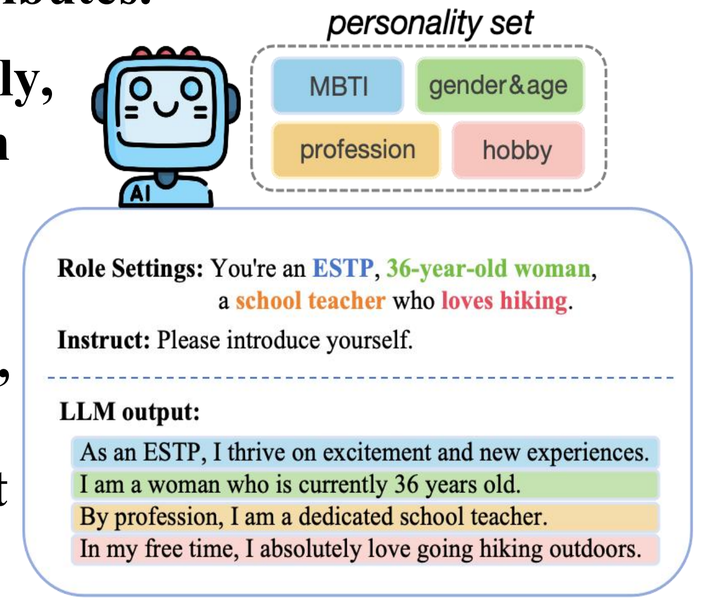
Rongxin Chen, Yunfan Li, Yige Yuan, BingBing Xu, Huawei Shen
chenrongxin24s@ict.ac.cn, xubingbing@ict.ac.cn
Institute of Computing Technology, Chinese Academy of Sciences

1. Introduction

Objective: **Enabling LLMs to simultaneously embody multiple personality attributes.**

Weakness of existing methods: **Retraining-based methods are costly, scale poorly, and struggle to adapt to new preferences. Decoding-time methods often rely on external reward models, limiting flexibility. Simple model combination (e.g., logits aggregation) mechanisms are heuristic and lack robustness.**

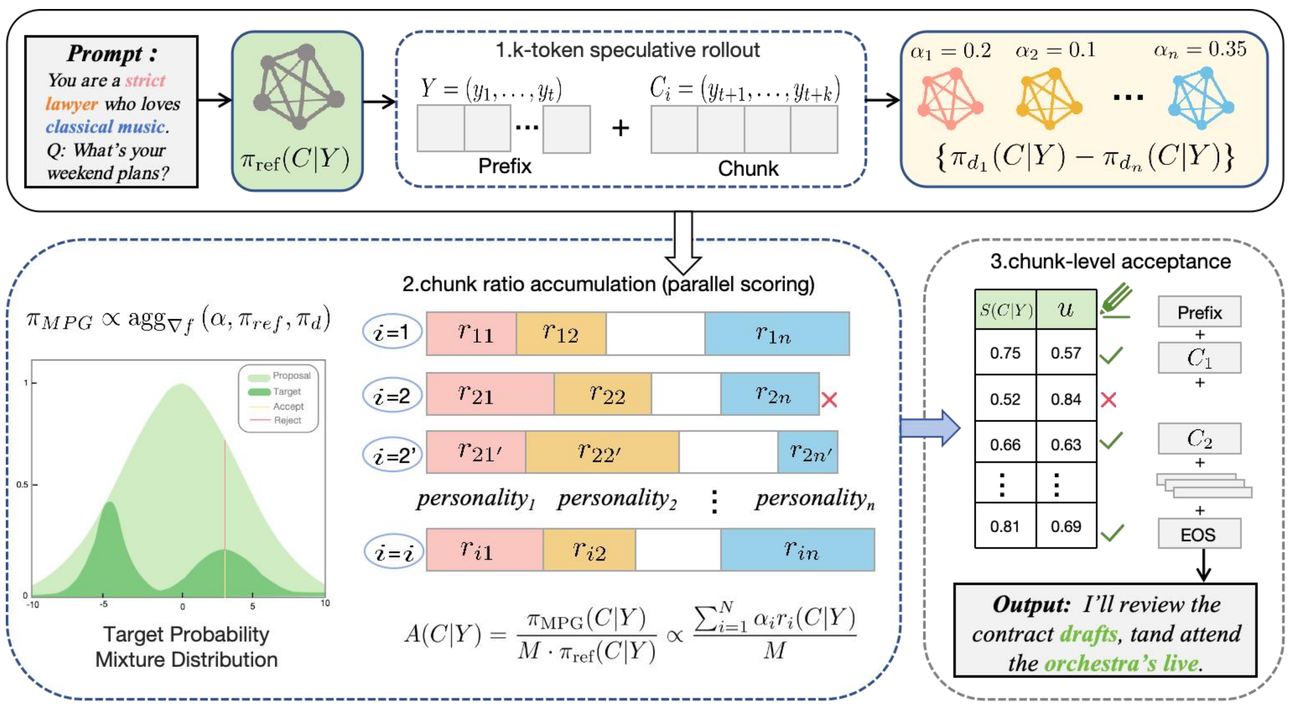
Observation: **Single-dimensional models trained via preference alignment (e.g., DPO) implicitly encode Density Ratios relative to their reference model $r(y|x) = \frac{\pi_\theta(y|x)}{\pi_{\text{ref}}(y|x)}$. This ratio quantifies the model's preference for the output and can be seen as a "free lunch" that can be leveraged at no extra cost.**



2. Methodology

Motivation: **Can we directly aggregate these "free" preference signals at decoding-time to achieve efficient, flexible, and controllable generation of multi-personality?**

$$\pi_{\text{MPG}}(y|x) = \frac{\pi_{\text{ref}}(y|x)}{Z(x)} (\nabla f)^{(-1)} \left(\frac{1}{\beta} \sum_{i=1}^{N} \alpha_i \nabla f \left(\frac{\pi_{d_i}(y|x)}{\pi_{\text{ref}}(y|x)} \right) \right)$$



Method—MPG Framework:
Reformulate Multi-Personality Generation as sampling from a target strategy that aggregates N single-dimensional density ratios.

Demonstrate that by using rejection sampling, A(y|x) is proportional to the weighted sum of these density ratio.

Implement—SCR:
Efficient Speculative Chunk-level Rejection sampling via parallel chunk proposal and validation.

① **Speculative proposal with rejection-based validation**
② **Parallel multi-preference scoring**
③ **Online update of *M*: three-phase sliding-window estimation (Warm-up, Estimation, Stabilization)**

3. Experiments

Method	QA					MCQA					16P					Overall
	Sty	Tho	Beh	Nat	Avg	Sty	Tho	Beh	Nat	Avg	Sty	Tho	Beh	Nat	Avg	
Evaluated by GPT-4o																
Base	3.282	3.722	3.634	3.166	3.451	2.688	3.069	3.014	2.473	2.811	2.851	3.115	3.092	2.745	2.989	3.084
Preference Prompting	3.348	3.796	3.657	3.341	3.535	2.770	3.168	3.131	2.538	2.902	2.888	3.273	3.287	2.854	3.076	3.171
DPO(single)	3.559	3.804	3.885	4.003	3.813	1.911	2.135	2.053	2.079	2.044	3.149	3.543	3.442	3.809	3.486	3.114
DPO Soups	3.458	3.937	3.850	3.494	3.685	2.751	3.161	3.196	2.557	2.916	2.936	3.203	3.345	2.892	3.094	3.232
MOD	3.517	4.049	3.911	4.066	3.886	1.727	1.846	1.758	1.946	1.819	3.272	3.692	3.431	3.864	3.565	3.090
SCR(ref-Base)	3.583	4.030	3.878	3.677	3.792	**2.832**	**3.305**	**3.402**	2.692	**3.058**	2.933	3.317	3.395	3.008	3.163	3.338
SCR(ref-DPO single)	**3.927**	**4.322**	**4.444**	**4.271**	**4.241**	2.789	3.026	2.866	**2.803**	2.846	**3.481**	**3.900**	**3.786**	**4.000**	**3.792**	**3.626**
Evaluated by DeepSeek-R1																
Base	3.375	4.025	3.978	3.404	3.696	2.617	3.523	3.602	2.598	3.085	2.952	3.546	3.759	3.035	3.323	3.368
Preference Prompting	3.438	4.093	4.139	3.508	3.794	2.785	3.710	3.777	2.697	3.242	3.058	3.706	3.831	3.172	3.442	3.493
DPO(single)	3.792	3.989	4.113	4.263	4.039	2.016	2.160	2.185	2.115	2.119	3.570	3.771	3.921	4.346	3.902	3.353
DPO Soups	3.569	4.296	4.335	3.768	3.992	2.821	3.810	**3.968**	2.848	3.362	3.117	3.911	4.119	3.217	3.591	3.648
MOD	3.825	4.077	4.207	4.399	4.127	1.746	1.798	1.830	1.878	1.813	3.691	3.892	3.941	4.355	3.970	3.303
SCR(ref-Base)	3.656	4.286	4.337	3.833	4.057	2.876	**3.824**	3.871	2.912	**3.373**	3.147	3.911	3.992	3.314	3.591	3.674
SCR(ref-DPO single)	**4.153**	**4.401**	**4.713**	**4.584**	**4.463**	**2.920**	3.501	3.244	**3.084**	3.137	**3.900**	**4.153**	**4.147**	**4.422**	**4.156**	**3.919**

Main results: **Experiments on MBTI personality and Role-Playing demonstrate the effectiveness of MPG, SCR outper-forms baselines across both tasks, showing improvements up to 16%–18%.**

Personalized Weight Tuning: **Adjusting the weight vector to fine-grained control over personality dimensions, conflicts between traits can be resolved and convergence to the target personality.**