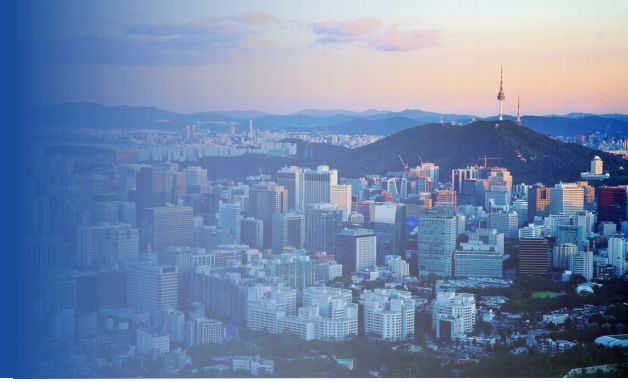


Multi-Personality Generation of LLMs at Decoding-time

Rongxin Chen[1,2], Yunfan Li[1,2], Yige Yuan[1,2], Bingbing Xu[1], Huawei Shen[1,2]

[1]Institute of Computing Technology, Chinese Academy of Sciences

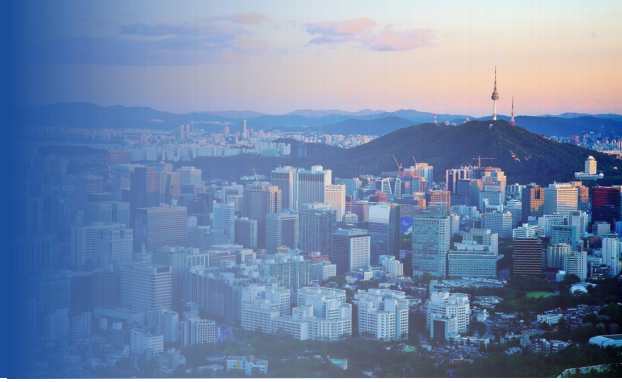
[2]University of Chinese Academy of Sciences

Paper

WeChat

Presenter: Rongxin Chen

Email: chenrongxin24s@ict.ac.cn

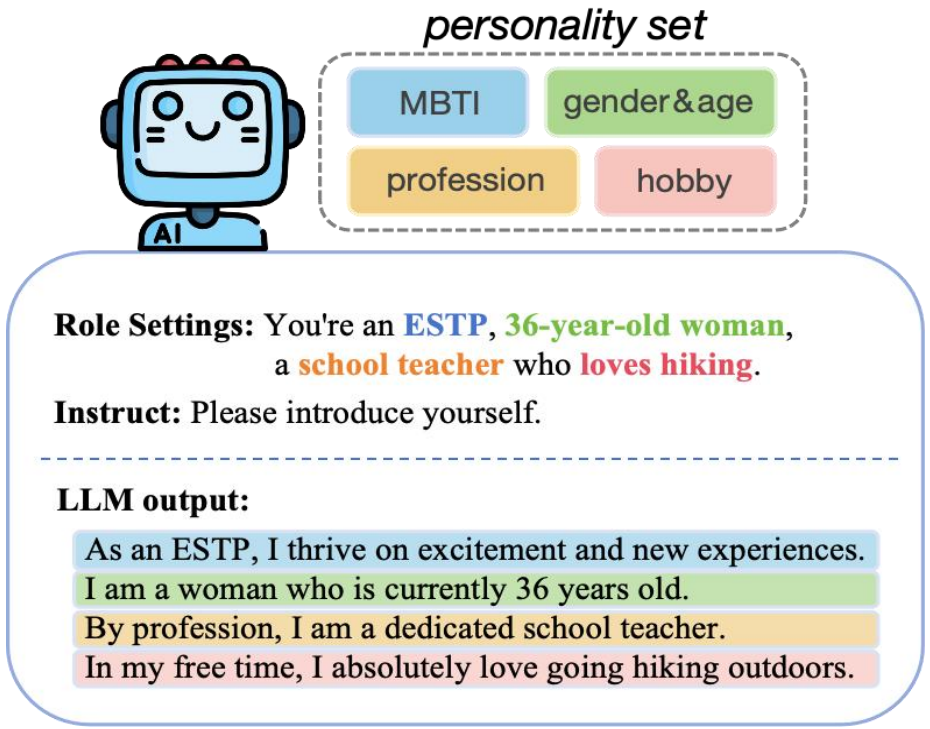


- **Motivation**
- Method
- Experiments

MOTIVATION

Multi-Personality Generation

- Multi-personality generation requires generated text to simultaneously embody **multiple, potentially interacting** personalization attributes.



MOTIVATION

Existing Work

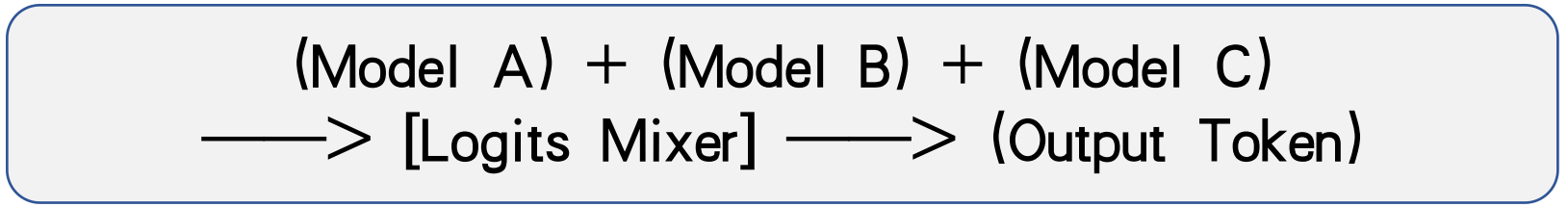
- Retraining-based methods
  - Use multi-objective optimization (e.g., MORLHF) to "integrate" multiple preferences into a single model's weights.
  - Costly, scale poorly, and struggle to adapt to new preferences.

(Base LLM) + (Prefs A, B, C)
——> [Expensive Retraining] ——> (Model_ABC)

MOTIVATION

Existing Work

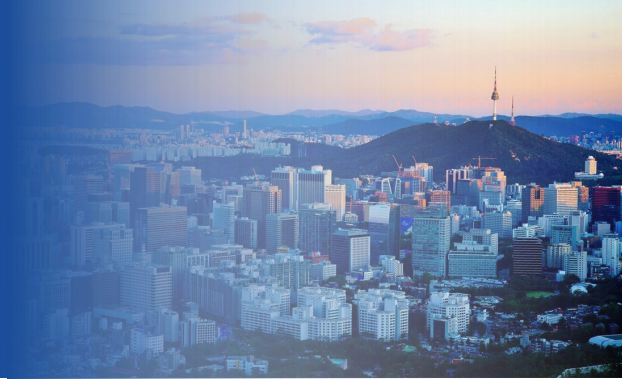
- Decoding-time methods
  - Guide the decoding process using external models (e.g., reward models) or by mixing logits from multiple models.
  - Multi-dimensional reward models are scarce and hard to acquire.
  - Simple model combination (e.g., logits aggregation) mechanisms are heuristic and lack robustness.



MOTIVATION

Main Idea

- **Observation:** When we align a model (e.g., for "Hobby") from a base model using DPO, all the preference information is encoded.
- **Where?** In the Probability Density Ratio: $r(y|x) = \frac{\pi_\theta(y|x)}{\pi_{\text{ref}}(y|x)}$
- This ratio is a "free lunch" —it quantifies the exact preference of a single attribute, and we already have it.
- **Main Idea:** *Can we skip retraining and complex reward models, and instead* **directly aggregate these "free" density ratios** *from multiple single-dimensional models at decoding-time?*



- Motivation
- **Method**
- Experiments

PROBLEM DEFINITION

Multi-Personality Generation

- **Goal:** We want to sample from a target distribution that aggregates N preferences.
- **Given:** A reference model: $\pi_{ref}(y|x)$

  N single-preference models: $\{\pi_{d_i}(y|x)\}_{i=1}^{N}$
- **Target Policy:** Define the optimal policy π_{MPG} as one that optimizes a **combined reward**.

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

METHOD: MPG Framework

The Bridge: MPG & Rejection Sampling

Target Policy

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

ratio(target and proposal)
in **Rejection Sampling**

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

METHOD: MPG Framework

The Bridge: MPG & Rejection Sampling

Target Policy

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

ratio(target and proposal)
in **Rejection Sampling**

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

Acceptance Probability

$$A(y|x) = \boxed{\frac{\pi_{\text{MPG}}(y|x)}{M \cdot \pi_{\text{ref}}(y|x)}}$$

METHOD: MPG Framework

The Bridge: MPG & Rejection Sampling

Target Policy

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

ratio(target and proposal)
in **Rejection Sampling**

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

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

Acceptance Probability

$$A(y|x) = \frac{\pi_{\mathrm{MPG}}(y|x)}{M \cdot \pi_{\mathrm{ref}}(y|x)}$$

METHOD: MPG Framework

The Bridge: MPG & Rejection Sampling

Target Policy

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

ratio(target and proposal) in **Rejection Sampling**

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

$$\frac{\pi_{\text{MPG}}(y|x)}{\pi_{\text{ref}}(y|x)} \propto \sum_{i=1}^{N} \alpha_i \nabla f \left(\frac{\pi_{d_i}(y|x)}{\pi_{\text{ref}}(y|x)} \right) \xrightarrow[\nabla f(x) = \log x + 1]{f(x) = x \log x} \frac{\pi_{\text{MPG}}(y|x)}{\pi_{\text{ref}}(y|x)} \propto \sum_{i=1}^{N} \alpha_i (\log r_i(y|x) + 1) \propto \sum_{i=1}^{N} \alpha_i r_i(y|x)$$

Acceptance Probability

$$A(y|x) = \frac{\pi_{\text{MPG}}(y|x)}{M \cdot \pi_{\text{ref}}(y|x)}$$

METHOD: MPG Framework

The Bridge: MPG & Rejection Sampling

Target Policy

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

ratio(target and proposal) in **Rejection Sampling**

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

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

$$f(x) = x \log x \quad \Rightarrow \quad \nabla f(x) = \log x + 1$$

$$\frac{\pi_{\text{MPG}}(y|x)}{\pi_{\text{ref}}(y|x)} \propto \sum_{i=1}^{N} \alpha_i (\log r_i(y|x) + 1) \propto \sum_{i=1}^{N} \alpha_i r_i(y|x)$$

Acceptance Probability

$$A(y|x) = \frac{\pi_{\text{MPG}}(y|x)}{M \cdot \pi_{\text{ref}}(y|x)}$$

$$A(y|x) = \frac{\pi_{\text{MPG}}(y|x)}{M \cdot \pi_{\text{ref}}(y|x)} \propto \frac{\sum_{i=1}^{N} \alpha_i r_i(y|x)}{M}$$

METHOD: MPG Framework

The Bridge: MPG & Rejection Sampling

Target Policy

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

ratio(target and proposal)
in **Rejection Sampling**

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

Acceptance Probability

$$A(y|x) = \frac{\pi_{\mathrm{MPG}}(y|x)}{M \cdot \pi_{\mathrm{ref}}(y|x)}$$

$$A(y|x) = \frac{\pi_{\mathrm{MPG}}(y|x)}{M \cdot \pi_{\mathrm{ref}}(y|x)} \propto \frac{\sum_{i=1}^{N} \alpha_i r_i(y|x)}{M}$$

The acceptance probability can be regarded as just a weighted sum of ratios!

$\alpha_i r_i(y|x)$

SCR: The Efficient Algorithm

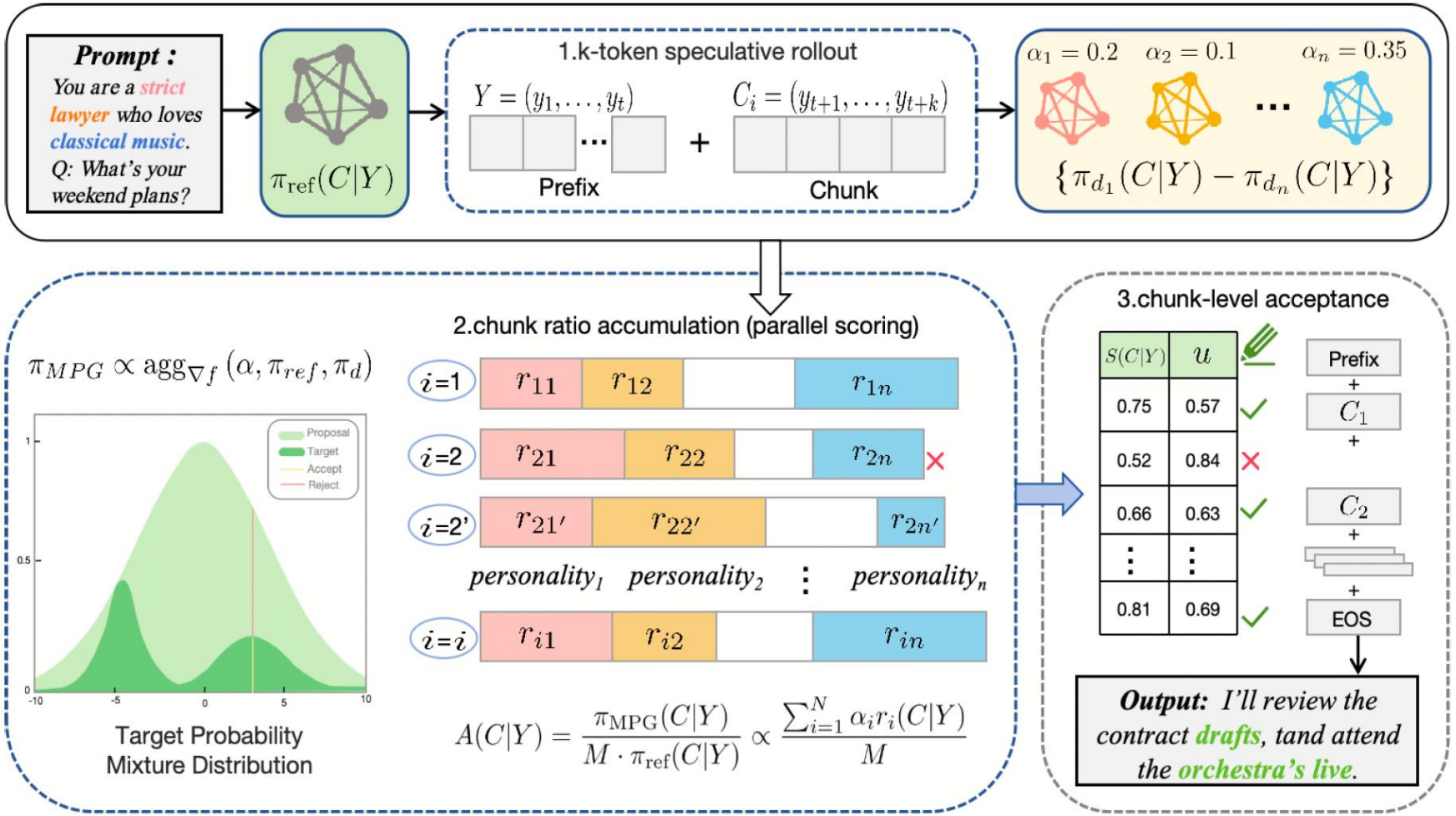
Speculative Chunk-level based Rejection sampling

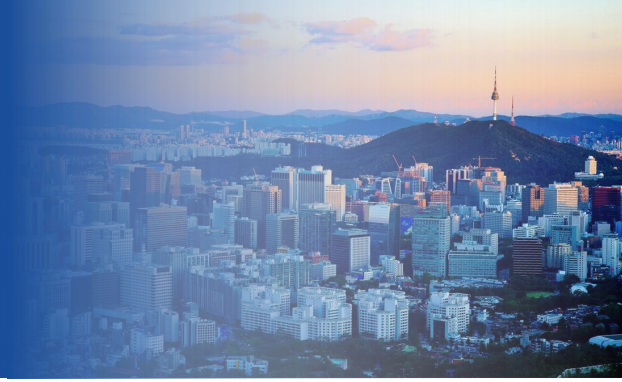
- **How it works:** full sequences ⟹ **parallel** chunks (k-tokens) **proposal and validation**
  1. Propose: ref_model generates a k-token chunk C.
  2. Validate: Get the score for that chunk $\text{Score}(C \mid Y) = \sum_{i=1}^{N} \alpha_i \, r_i(C \mid Y)$ by querying all single-preference models.
  3. Accept: accept/reject the chunk using a dynamic threshold M. And **Online Update of M** via a <u>three-phase sliding window</u>:

Warm-up → Estimation → Stabilization

SCR: The Efficient Algorithm

Speculative Chunk-level based Rejection sampling





- Motivation
- Method
- **Experiments**

EXPERIMENTS

Experimental Settings

2 tasks: MBTI Personality Simulation and Role-Playing

- **Datasets**
  - MBTI-QA:instruction-following QA
  - MBTI-MCQA:multiple-choice answering
  - MBTI-16P:16Personalities psychometric instrument
- **Evaluation**
  - **LLM-as-a-Judge**
    - **Metrics:** Style (Sty), Thought (Tho), Behavior (Beh), Naturalness (Nat)

- **Datasets**
  - Adapt the methodology from ALOE[1], focusing on user profile and personality.
- **Evaluation**
  - **LLM-as-a-Judge**
    - **Metrics:** Profile Relevance (PR), Persona Match (PM), Humanlikeness (Hl)
  - **Reference-based**
    - **Metrics:** BLEU, ROUGE-1 F1, BERTScore, Perplexity

[1]https://github.com/ShujinWu-0814/ALOE

EXPERIMENTS

Experimental Settings

2 tasks: MBTI Personality Simulation and Role-Playing

- **Baselines**
  - **Base:** Llama-3-8B-Instruct.
  - **DPO:** Direct Preference Optimization applied to the base model.
  - **Preference Prompting (PP):** implements personality conditioning through explicit attribute descriptions in prompt engineering, applied across base model, single-preference model, and specialized model configurations.
  - **DPO Soups (a variant of Personalized Soups):** acquires the LoRA parameters of independently - trained single - preference DPO models. Following weighted averaging, these parameters are applied to π_{ref} to generate text with multi-personality.
  - **MOD:** performs a weighted linear combination of the output logits from single - attribute DPO models during decoding and samples based on the combined logits.

EXPERIMENTS

Main results

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**

Table: Comparison of baseline methods and SCR on MBTI task. SCR(ref-Base) and SCR(ref-DPO single) refer to the use of Base and DPO single, respectively, as the reference model to perform the SCR.

EXPERIMENTS

Main results

Method	Evaluated by GPT-4o				Evaluated by DeepSeek-R1				Reference-based Evaluation			
	PR	RM	Hl	Avg	PR	RM	Hl	Avg	BLEU	ROGUE-1	BERTScore	PPL
Base	3.580	3.770	3.471	3.607	3.418	3.778	4.098	3.765	0.010	0.088	0.762	43.984
Preference Prompting	3.720	3.853	3.832	3.802	3.582	3.821	4.283	3.895	0.024	0.127	0.823	41.860
DPO(single)	3.823	4.240	4.137	4.066	3.744	4.106	4.563	4.137	0.058	0.167	0.868	48.262
DPO Soups	3.818	3.998	3.871	3.896	3.692	3.857	4.379	3.976	0.022	0.130	0.823	43.594
MOD	4.020	**4.298**	4.170	4.166	3.970	**4.187**	4.586	**4.241**	0.082	0.187	0.870	47.188
SCR(ref-Base)	3.874	3.997	3.861	3.911	3.722	3.833	4.414	3.990	0.034	0.133	0.829	**36.694**
SCR(ref-DPO single)	**4.030**	4.290	**4.192**	**4.167**	**3.976**	4.166	**4.596**	4.230	**0.085**	**0.188**	**0.897**	43.335

Table: Comparison of baseline methods and SCR on Role-Playing task.

Experiments on 2 tasks indicate the effectiveness of MPG,
SCR outperforms baselines, showing improvements up to 16%–18%.

EXPERIMENTS

Analysis 1: Efficiency (SCR is Fast)

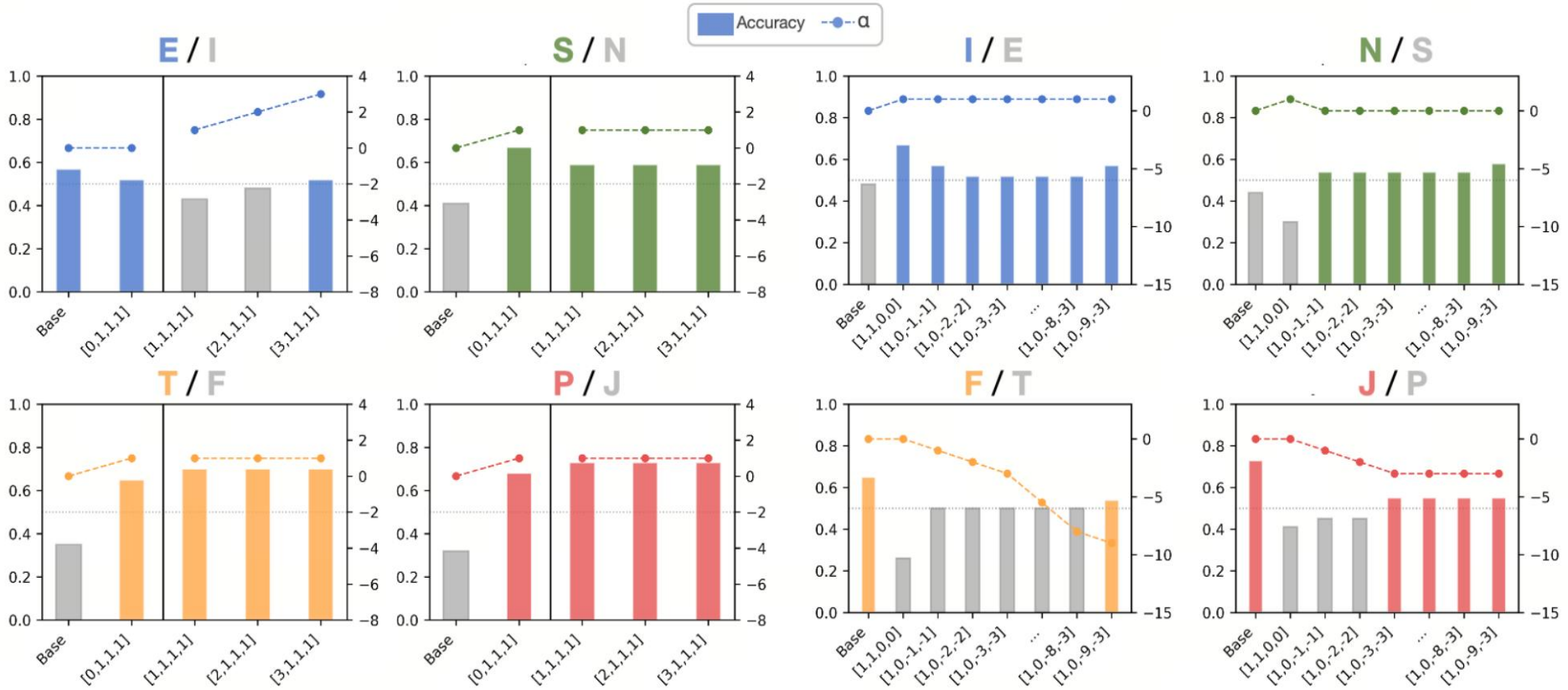
Method	Throughput ↑ (tok/s)	Latency ↓ (s/seq)	FwdPass ↓ (per tok)	Reject Rate ↓ (%)
Base	120.0	1.07	1.0	—
DPO Soups	118.0	1.08	~1.0	—
MOD	60.0	2.13	2.0	—
Seq-RS	11.0	11.64	10.0	60.0
Token-RS	30.0	4.27	4.62	25.0
SCR	97.0	1.20	1.4	30.0

Table: Efficiency comparison of different methods for Role-Playing task ($N = 2$, chunk size $k = 4$).↑ indicates higher is better,while ↓ indicates lower is better.

SCR provides the best trade-off between quality and speed

EXPERIMENTS

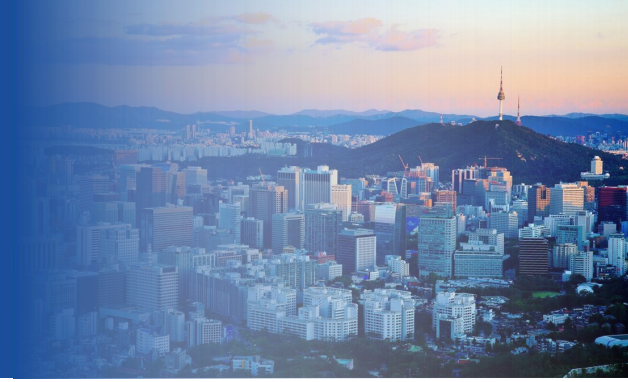
Analysis 2: Controllability (α-Tuning)



(a)The accuracy and parameter adjustment of ESTP

(b)The accuracy and parameter adjustment of INFJ

Figure: Iterative tuning process for the α. Bars indicate prediction accuracy (left axis) for each MBTI dimension; dashed lines track α values (right axis) at each optimization step.



Thank you for your attention!

Paper

WeChat

Email: chenrongxin24s@ict.ac.cn