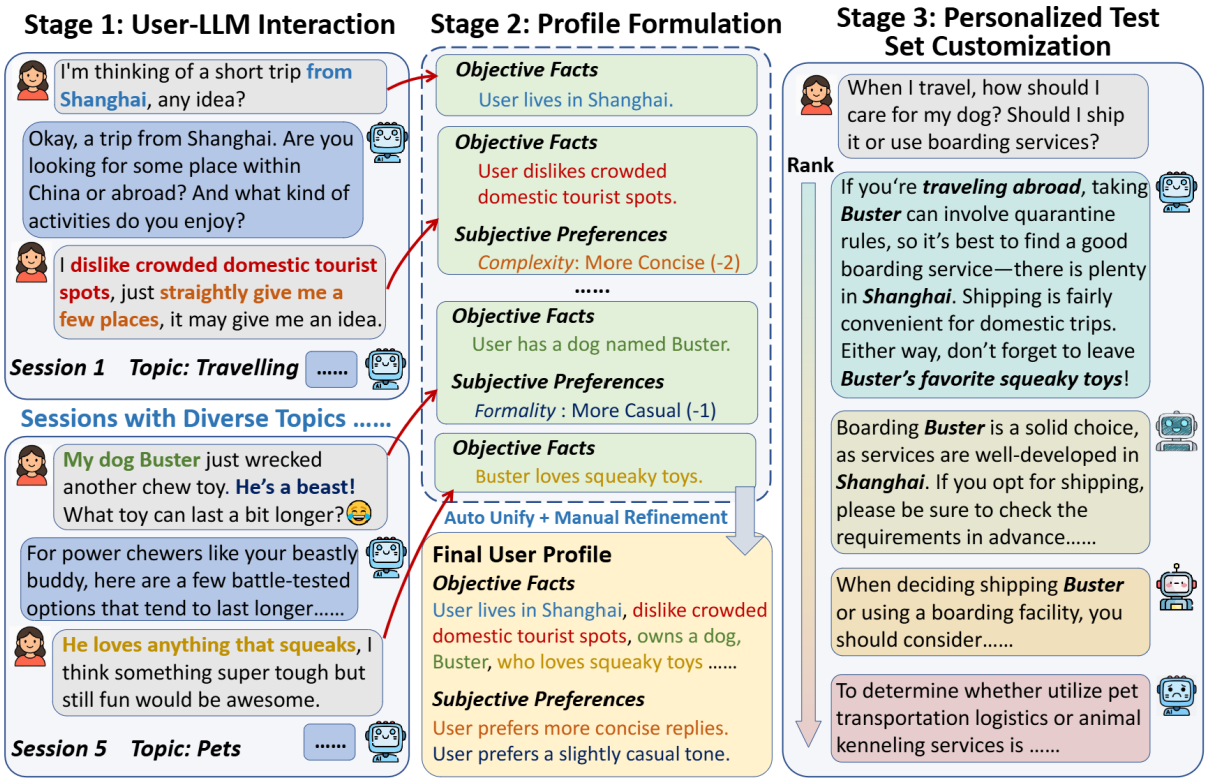
ColdChat: Benchmarking Large Language Model Personalization using Limited Real-User Interaction History

Wenbin Zhang, Zheni Zeng, Zhiyuan Liu, Wanxiang Che

Frontiers of Computer Science, FCS-251716

Problems & Ideas

- Problems of Personalization for LLM-based Chatbots:
  - Facing a "cold-start" scenario with only limited interaction history.
  - A critical lack of data resources for this challenge.
  - Existing methods struggle to personalize LLMs in this scenario.
- Ideas: Introduce ColdChat, the first benchmark for this scenario, alongside EPIC framework for effective personalization.



Main Contributions

- Contributions:
  - Constructed the first benchmark specifically designed to evaluate LLM personalization in “cold-start” scenarios, featuring real-world multi-session dialogues and self-annotated user profiles;
  - Proposed a two-stage framework for Extracting user Profiles from Interaction Context (EPIC), which significantly outperforms baselines in both personalized ranking (+27.2%) and generation tasks (+12.7%).

	Methods	*GPT-4o*	*Claude-3.5-Sonnet*	*Llama-3.1-70B-Instruct*	*Llama-3.1-8B-Instruct*	*Qwen-2.5-72B-Instruct*	*Qwen-2.5-7B-Instruct*
	Random Ranking	0.00 / 0.25					
Baselines	**Zero-Shot**	0.11 / 0.33	0.16 / 0.35	0.07 / 0.30	0.06 / 0.25	0.10 / 0.31	0.12 / 0.30
	Chat-Prompting	0.24 / 0.37	0.25 / 0.38	0.16 / 0.34	0.07 / 0.29	0.12 / 0.33	0.09 / 0.30
	Self-inferred Profiling	0.19 / 0.34	0.23 / 0.37	0.18 / 0.35	0.10 / 0.29	0.13 / 0.33	0.15 / 0.32
	FSPO	——	——	0.33 / 0.41	0.25 / 0.34	0.37 / 0.41	0.26 / 0.33
EPIC-Based Methods	**Profile-Prompting**	0.52 / 0.52	0.50 / 0.53	0.32 / 0.42	0.26 / 0.35	0.34 / 0.40	0.24 / 0.31
	Profile-Refine	**0.60 / 0.58**	**0.55 / 0.57**	0.39 / 0.45	**0.31** / 0.39	0.41 / 0.44	0.30 / 0.38
	Few-Shot CoT	0.54 / 0.55	0.48 / 0.53	0.35 / 0.42	0.27 / 0.34	0.34 / 0.41	0.29 / 0.34
	Personalized RM	——	——	**0.46 / 0.52**	0.30 / **0.40**	**0.43 / 0.51**	**0.31 / 0.41**

Table 9 The results of the Personalized Ranking Task, values are presented as **Spearman's ρ / Top-1 Accuracy**. The EPIC-based methods consistently outperform baseline approaches, with Profile-Refine and Personalized RM yielding the best results. FSPO and Personalized Reward Models are only applied on the open-sourced Llama and Qwen models.

	Methods	*GPT-4o*	*Claude-3.5-Sonnet*	*Llama-3.1-70B-Instruct*	*Llama-3.1-8B-Instruct*	*Qwen-2.5-72B-Instruct*	*Qwen-2.5-7B-Instruct*	**Average**
Baselines	**Zero-Shot**	22.47	21.72	19.45	17.28	19.77	15.89	19.43
	Chat-Prompting	22.97	21.52	20.95	16.64	18.53	15.01	19.27
	Self-inferred Profiling	22.91	22.23	19.96	17.76	19.44	16.62	19.82
	LD-Agent	23.98	22.14	21.64	18.10	19.91	15.24	20.17
EPIC-Based Methods	**Profile-Prompting**	25.34	**24.15**	22.13	19.20	21.22	16.68	21.45
	Profile-Refine	**26.59**	23.40	**22.63**	18.90	**21.30**	**17.44**	**21.71**
	Few-Shot CoT	23.57	21.80	19.24	**19.62**	20.49	15.48	20.03

Table 10 The results of the Personalized Generation Task, the values are multi-dimensional LLM-as-a-judge scores designed in Section 5.2, with a maximum of 30. The EPIC-based methods outperform the approaches that rely on interaction history and those focus on LLM memory, with the Profile-Refine setup yields the best results.