LLMs + Persona-Plug = Personalized LLMs (ACL 2025)

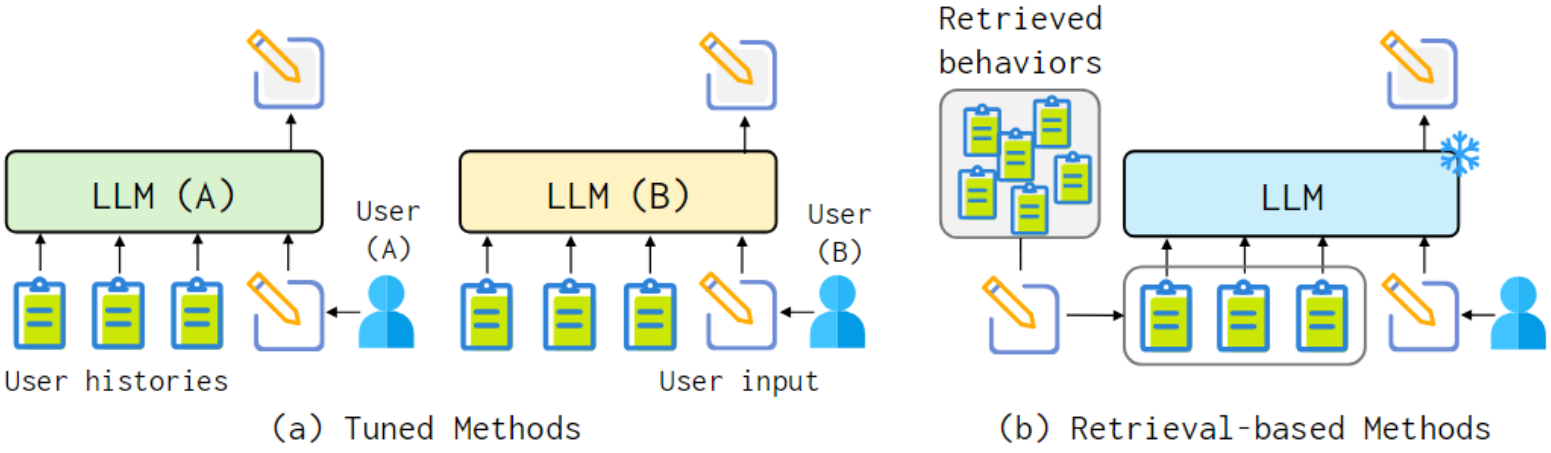
Jiongnan Liu, Yutao Zhu, Shuting Wang, Xiaochi Wei, Erxue Min, Yu Lu, Shuaiqiang Wang, Dawei Yin and Zhicheng Dou

October 26, 2025

Seoul National University - IDEA Lab.
Presenter: Sehyun Park

Introduction

- Personalization plays a crucial role in various language tasks and applications, as users with the same requirements may prefer diverse outputs depending on their individual interests.
- There are two primary approaches to achieving a personalized LLM for each user: the *fine-tuned* method and the *retrieval-based* method.



Introduction

- Limitations of previous methods
  - *Fine-tuned method*
    - Some approaches fine-tune a unique personalized LLM for each user, which is too expensive for widespread application.
  - *Retrieval-based method*
    - This strategy may break the continuity of user history and fail to capture the user' s overall styles and patterns, leading to sub-optimal performance.
- Goal
  - This paper proposes a method that encodes all historical behaviors of each user and aggregates them into a single personalized embedding, which is then integrated into the LLM.
- Notation
  - i: the index of the user, where $i = 1, 2, \ldots, m$.
  - x_i: the task input of the i-th user.
  - $H_i = \{h_1, h_2, \ldots, h_{n_i}\}$: the history (or profile) of the i-th user.

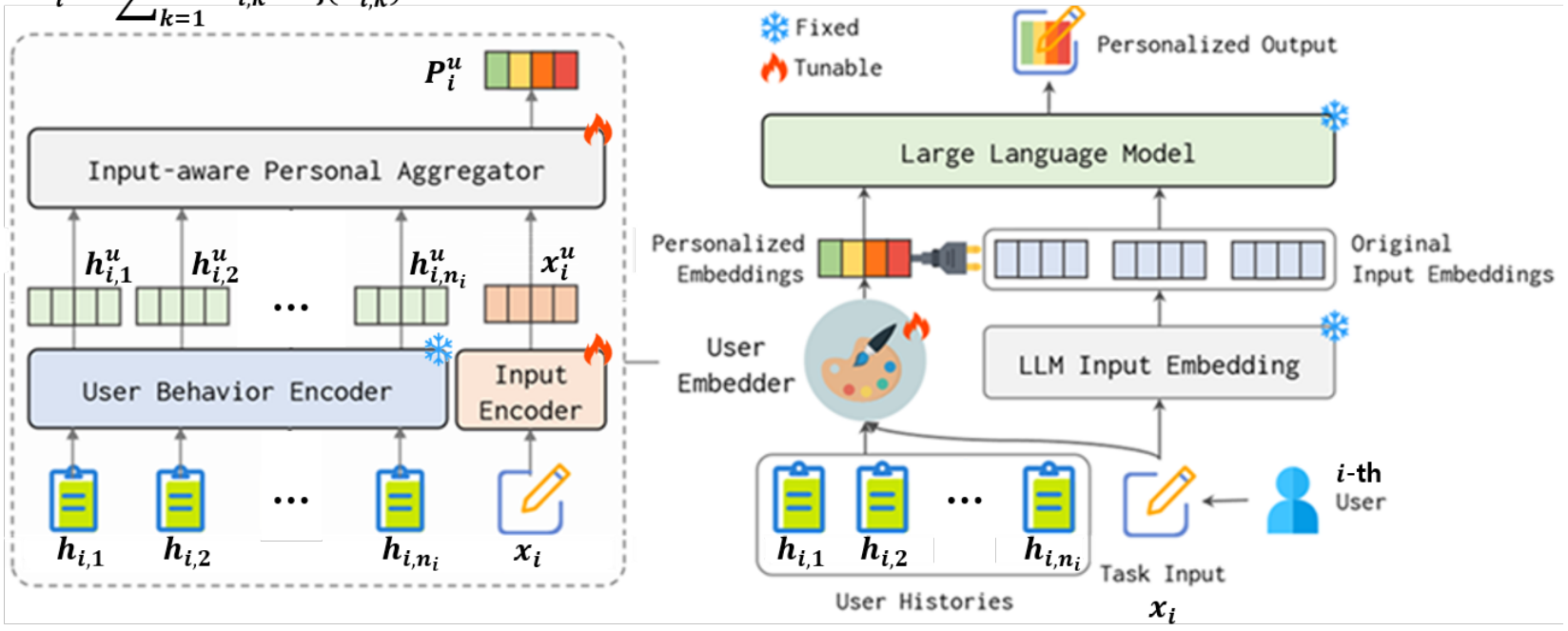
Overall framework

$$w_{i,k} = \frac{\exp(x_i^{u\top} h_{i,k}^u)}{\sum_{l=1}^{n_i} \exp(x_i^{u\top} h_{i,l}^u)}$$

$$P_i^u = \sum_{k=1}^{n_i} w_{i,k} \mathrm{Proj}(h_{i,k}^u)$$

$$X_i^u = [I; P_i^u; \mathrm{Emb}_{LLM}(x_i)]$$

$$\mathcal{L} = -\sum_i \sum_t \log p_{LLM}\left(y_{i,t} | X_i^u, \mathrm{Emb}_{LLM}(y_{i,<t})\right)$$



▶ where Proj(·) transforms the user embeddings from the encoder output space into the LLM input space through a two-layer MLP.

▶ I : a learnable embedding token to guide the LLM with task-specific information.

Experimental Results

Table 1: Performance of all models on six LaMP tasks. The best results are in **bold**.

Dataset	Metric	Ad-hoc	Naive RBP			Optimized RBP				PPlug
		FlanT5-XXL	BM25	Recency	Contriever	ROPG-RL	ROPG-KD	RSPG-Pre	RSPG-Post	
LaMP-1	Accuracy ↑	0.498	0.629	0.639	0.641	**0.682**	0.676	0.672	0.670	0.680
LaMP-2	Accuracy ↑	0.326	0.345	0.361	0.362	0.365	0.365	0.391	0.416	**0.565**
	F1 ↑	0.255	0.282	0.291	0.282	0.292	0.291	0.312	0.337	**0.501**
LaMP-3	MAE ↓	0.335	0.293	0.305	0.297	0.273	0.274	0.266	0.246	**0.231**
	RMSE ↓	0.639	0.585	0.596	0.592	0.561	0.566	0.560	0.539	**0.534**
LaMP-4	ROUGE-1 ↑	0.173	0.192	0.194	0.190	0.190	0.193	0.195	0.207	**0.216**
	ROUGE-L ↑	0.157	0.175	0.177	0.174	0.174	0.176	0.179	0.188	**0.197**
LaMP-5	ROUGE-1 ↑	0.472	0.467	0.469	0.471	0.473	0.472	0.479	0.480	**0.487**
	ROUGE-L ↑	0.419	0.419	0.422	0.421	0.425	0.423	0.429	0.429	**0.435**
LaMP-7	ROUGE-1 ↑	0.454	0.451	0.452	0.440	0.458	0.451	0.460	0.468	**0.536**
	ROUGE-L ↑	0.401	0.401	0.402	0.391	0.407	0.402	0.409	0.416	**0.484**

- LLM model : FlanT5-XXL (11B)
  - ⇒ Input dim : 4096
- Encoder model : BGE-baseen-v1.5 (0.1B)
  - ⇒ Output dim : 768

End