Externalizing Social-Cognitive Structures for User Modeling: Toward Theory-Driven Profiling with LLMs

Taehyung Noh
Hanyang University
Seoul, Republic of Korea
yestaehyung@hanyang.ac.kr

Seungwan Jin
Hanyang University
Seoul, Republic of Korea
seungwanjin@hanyang.ac.kr

Haein Yeo
Hanyang University
Seoul, Republic of Korea
haeinyeo@hanyang.ac.kr

Kyungsik Han*
Hanyang University
Seoul, Republic of Korea
kyungsikhan@hanyang.ac.kr

Abstract

In this paper, we propose TRIPLE (**T**PB-d**RI**ven **P**rofiling with **L**LM r**E**finement), a dynamic profiling framework that incorporates the Theory of Planned Behavior (TPB) into user profile modeling. Our method (1) extracts TPB components from historical text data to construct an initial user profile, (2) iteratively refines this profile by analyzing discrepancies between predicted and actual behaviors, and (3) continuously updates the user's state by incorporating newly arriving text. We evaluate TRIPLE on the LaMP datasets, focusing on rating prediction and personalized tweet paraphrasing tasks, using multiple open-source large language models. Experimental results demonstrate that TRIPLE consistently outperforms existing profiling methods across all evaluation settings. Qualitative analysis confirms that TRIPLE captures the psychological and social mechanisms underlying users' product evaluation and description. These findings provide empirical evidence that theory-driven user profiling can significantly improve personalization performance in recommender systems and related applications. Our implementation and examples of generated profiles are available at https://yestaehyung.github.io/cikm25-triple/.

CCS Concepts

• **Information systems → Personalization**.

Keywords

personalization; user modeling; large language model; dynamic profile refinement; theory of planned behavior

ACM Reference Format:
Taehyung Noh, Seungwan Jin, Haein Yeo, and Kyungsik Han. 2025. Externalizing Social-Cognitive Structures for User Modeling: Toward Theory-Driven Profiling with LLMs. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management (CIKM '25), November 10–14, 2025, Seoul, Republic of Korea.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3746252.3760965

*Corresponding author

CIKM '25, Seoul, Republic of Korea

ACM ISBN 979-8-4007-2040-6/2025/11
https://doi.org/10.1145/3746252.3760965

1 Introduction

Accurate user profiling remains a central challenge in personalized services [15, 16]. Research has suggested that behavioral intentions are shaped not only by personal attitudes but also by subjective norms and perceived behavioral control [1], and that behavioral decision-making may vary across contexts such as time, location, social companions, and mood [5]. Because user decisions reflect complex interactions among psychological, social, and contextual factors, capturing these latent dimensions is important for improving prediction accuracy and enhancing user experience [6, 8, 20].

Such latent characteristics have also been actively explored in personalization research, including recommendation systems [25]. Latent behavioral intentions are often inferred through embeddings of diverse consumer behavior data (e.g., purchasing behavior, purchase history, product features), which are then to train predictive models. However, these latent variables are typically represented as high-dimensional vectors computed through specific embedding algorithms and are often difficult to interpret, making it challenging to assess whether and how consumers' cognitive processes are effectively reflected in the modeling process [3]. The reliance on such algorithmically derived representations can constrain the incorporation of domain knowledge and socio-cognitive factors, reducing both model transparency and flexibility.

The emergence of Large Language Models (LLMs) offers new opportunities for extracting and modeling these latent variables. Pre-trained on massive text corpora, LLMs internalize extensive knowledge of human psychology, social norms, and behavioral patterns, allowing them to infer diverse user- and context-related latent factors with high confidence [11]. In line with this opportunity, recent studies have leveraged LLMs to generate user profiles from text data using LLMs and predict downstream decisions (e.g., product preferences, writing styles) [23].

Nevertheless, existing LLM-based profiling approaches still leave room for improvement. First, they often extract only surface-level preferences such as likes/dislikes or areas of interest, without systematically modeling the core latent variables that drive user behavior [26]. Second, they are somewhat limited in their ability to capture the dynamic nature of these latent variables. Since users' attitudes, social influences, and situational constraints can change over time, it becomes essential to continuously monitor and update these latent variables to ensure that predictive models remain accurate and reflective of user states [14, 24]. Existing approaches that

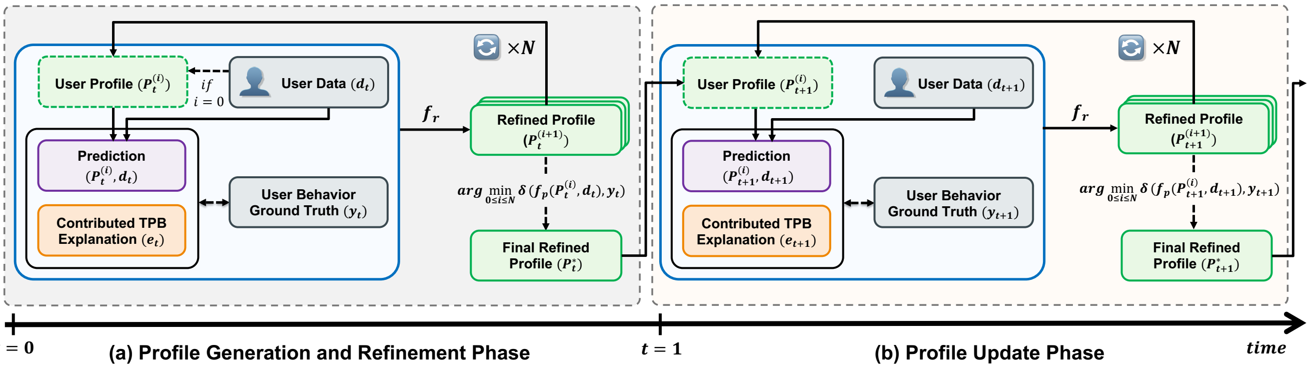

Figure 1: The architecture of TRIPLE. The *prediction-evaluation-refinement cycle* plays a crucial role in capturing user states.

simply regenerate entire profiles or incrementally aggregate new information are insufficient to address these dynamics.

To address these challenges, we propose TRIPLE (**T**PB-d**RI**ven **P**rofiling with **L**LM r**E**finement), a dynamic LLM-based profiling framework that incorporates the latent variable structure of the Theory of Planned Behavior (TPB) [1], one of the foundational theories explaining user behavior. TRIPLE leverages LLM reasoning to extract key TPB components—attitude, subjective norms, and perceived behavioral control—from input data and represent them in natural language. These latent variables are then continuously updated through a *prediction-evaluation-refinement cycle* to better reflect evolving user states. Errors identified during the evaluation phase are used to refine the latent variable representations, thereby improving behavioral prediction performance.

To evaluate the effectiveness of TRIPLE, we conducted experiments on the LaMP-3, 4 [18] for rating prediction and news headline generation tasks. Experimental results using various open-source LLMs, including Llama-3.1-8B-Instruct [7] and Qwen-2.5-32B-Instruct [21], consistently demonstrated superior performance of TRIPLE compared to existing methods.

This work presents the first systematic integration of social-psychological theory into LLM-based profiling, offering broad applicability across various LLMs without the need for additional training. TRIPLE demonstrates strong potential for application across diverse domains, offering insights to guide future research.

2 Background

While traditional methods (e.g., collaborative filtering [2, 19]) often lack interpretability and struggle to capture the evolving nature of user behavior, LLMs have opened up new opportunities for personalized recommendation and natural language generation [9, 12]. To enhance personalization, recent studies have explored ways to incorporate user data into LLM prompts. One approach involves summarizing a user's past interactions to create concise, informative prompts. For example, Richardson et al. (2023) proposed generating a user summary in advance and incorporating it into the prompt along with retrieved information to address both latency and token constraints when generating personalized responses [17].

Other methods focus on constructing intermediate user profiles from raw textual data. For example, GPG [23] extracts writing style, preferences, and intentions from user-generated content such as reviews or social media posts. These structured summaries serve as intermediate profiles that guide downstream tasks such as tweet paraphrasing and purchase prediction. Beyond static profiling, PURE [4] incrementally updates profiles based on extracted user preferences, such as likes, dislikes, and product features, from reviews to improve rating prediction accuracy.

Despite these advances, most existing LLM-based personalization techniques have focused on identifying favored keywords or writing styles, without modeling the deeper psychosocial mechanisms that underlie user decisions [13]. User profiles have been treated as static or naively updated, relying on the simple accumulation of data, which limits the ability to adapt to evolving user preferences or contextual shifts over time.

To overcome these limitations, we propose incorporating the Theory of Planned Behavior (TPB), a foundational framework in social psychology that explains the motivational drivers of human behavior, into LLM-based user profiling. This approach enables the extraction of deeper motivational factors, explainable behavior prediction, and temporal profile optimization–capabilities largely overlooked in prior work. To the best of our knowledge, this is the first study to successfully incorporate a theory-driven approach into LLM-based personalization, highlighting the potential of combining psychological theory with LLMs.

3 Methods

TRIPLE consists of four main stages: (1) generating an initial profile by extracting TPB components using LLM, (2) performing a personalization task using the generated profile, (3) refining the profile based on behavior prediction results, and (4) dynamically updating the user profile over time. The overall architecture of our framework is illustrated in Figure 1.

3.1 TPB-based Initial Profile Construction

To generate an initial user profile, we leverage the Theory of Planned Behavior (TPB) [1], which posits that human behavior is determined by three key components: *attitude*, *subjective norm*, and *perceived behavioral control*. We analyze the user's textual data using an LLM to extract these components and form the initial profile. At time $t = 0$, the initial profile P_0 is generated from the user's textual data d_0. The profile generation proceeds as follows:

$$P_0 = f_i(d_0) \tag{1}$$

where f_i is an LLM that takes prompts as input to extract TPB components from user text and generate an initial profile. The components are stored in JSON format for use in later modules.

3.2 Profile Refinement

To incrementally improve the user profile P_t, we adopt a dynamic refinement method based on an iterative cycle of prediction, evaluation, and refinement. This approach assesses how well the profile contributes to behavior prediction and continuously updates it to minimize prediction error.

Behavior prediction: Given the profile P_t and user data d_t, the model predicts the user behavior y_t and generates an explanation e_t of how each TPB component contributed to the prediction:

$$\hat{y}_t, e_t = f_p(P_t, d_t) \tag{2}$$

where f_p is an LLM that takes prompts as input to predict user behavior.

Profile refinement: The predicted behavior $\hat{y}_t$ is compared with the observed $\hat{y}_t$, and the profile is refined using the discrepancy and explanation e_t.

$$P_t^{(i+1)} = f_r(\hat{y}_t, y_t, P_t^{(i)}, d_t, e_t) \tag{3}$$

where f_r is an LLM that takes prompts as input to refine the user profile. This refinement is repeated N times, and the profile with the lowest prediction error is selected as the final refined profile P_t^*.

$$P_t^* = \arg\min_{0 \le i \le N} \delta(f_p(P_t^{(i)}, d_t), y_t) \tag{4}$$

where $P_t^{(i)}$ denotes the i^{th} candidate profile and δ depends on the specific task (e.g., MAE for rating prediction, ROUGE for text generation). We set $N = 5$ to balance refinement effectiveness and computational efficiency by limiting excessive LLM calls.

3.3 Profile Update

The profile update mechanism is designed to capture temporal changes in user behavior and gradually adapt to them. This process consists of three main steps. First, given the previous optimal profile P_t^* and the new textual data d_{t+1}, the model predicts the user's behavior at the next time step y_{t+1} and generates an explanation e_{t+1}, which describes the contribution of each TPB component to the prediction:

$$\hat{y}_{t+1}, e_{t+1} = f_p(P_t^*, d_{t+1}) \tag{5}$$

Next, the predicted behavior is compared with the ground truth y_{t+1}. Based on the discrepancy between the prediction and the ground truth, as well as the explanation e_{t+1} and the new data d_{t+1}, the profile is iteratively refined as follows:

$$P_{t+1}^{(i+1)} = f_r(\hat{y}_{t+1}, y_{t+1}, P_{t+1}^{(i)}, d_{t+1}, e_{t+1}) \tag{6}$$

Here, the update process starts from the previous optimal profile, using $P_{t+1}^{(0)} = P_t^*$ as the initialization. Through the iterative refinement, a set of $N+1$ candidate profiles $\{P_{t+1}^{(0)}, ..., P_{t+1}^{(N)}\}$ is generated. Each candidate is evaluated by computing the prediction error, and the profile with the lowest error is selected as the updated optimal profile:

$$P_{t+1}^* = \arg\min_{0 \le i \le N} \delta(f_p(P_{t+1}^{(i)}, d_{t+1}), y_{t+1}) \tag{7}$$

Through this mechanism, the user profile is progressively updated over time. The optimal profile P^* at each time step is used as the initialization for the consequent update, allowing the model to effectively capture the user's evolving preferences and behavioral patterns.

4 Experiments

4.1 Experiment Setup

We evaluate TRIPLE on the LaMP benchmark [18], focusing on LaMP-3 (product rating) and LaMP-4 (news headline generation) as representative tasks for classification and generation, respectively. LaMP-3 involves predicting product ratings on a 1-5 scale for Amazon items based on users' review histories (2,500 test samples), and is evaluated using MAE and RMSE. LaMP-4 involves generating personalized news headlines by modeling user writing styles (1,800 test samples), with performance evaluated using ROUGE-1 and ROUGE-L scores. To preserve the temporal structure of user behavior, we use a time-based separation approach and conduct all evaluations on the test set.

4.2 Baselines

We compare our method against baselines along two dimensions. First, for profile generation, we evaluate: (1) Zero-shot, which does not use any user profile; (2) ICL-Before, where raw user histories are directly used as profiles, specifically using only the most recent review per user. (3) PAG [17] and PURE [4], which construct structured profiles with positive/negative preferences and key features; and (4) ONCE [10], which generates summarized user profiles including users' interests in specific topics and regions. Second, for profile update, we compare: (1) regeneration using only new incoming data [26], and (2) sliding window [22], which maintains a fixed-size history. To ensure a comprehensive assessment, we evaluate four open-source LLMs.

5 Results

5.1 Evaluation of TPB-Based Profile

Table 1 presents a comprehensive comparison of personalization performance between TRIPLE and existing profiling approaches across two datasets. TRIPLE consistently achieves either the highest or second-highest performance across all evaluation models in both tasks, even without the refinement cycle.

Compared to existing LLM-based profiling methods, TRIPLE demonstrates improvements across all evaluation metrics. These results support our hypothesis that modeling core latent variables underlying user behavior—rather than merely capturing surface-level user preferences—yields more accurate behavioral predictions. Among the baselines, PAG emerges as the strongest competitor for models with larger parameter sizes; however, it still falls short of the performance achieved by our theoretically grounded approach.

The prediction-evaluation-refinement cycle further enhances performance, with refined profiles consistently outperforming the no-refinement configuration across all experimental settings. These results underscore the importance of continuously monitoring and updating latent variables to reflect evolving user states, as emphasized in our framework design. This iterative process enables

Table 1: Performance evaluation results. Bold: best performance, underlined: second-best performance.

Model	Method	LaMP-3		LaMP-4	
		MAE ↓	RMSE ↓	ROUGE-1 ↑	ROUGE-L ↑
Llama-3.2-3B-Instruct	zero-shot	0.5368	0.8564	0.1271	0.1106
	ICL-Before	0.5566	0.9696	0.0998	0.0862
	PAG [17]	0.4902	0.18346	0.0651	0.0565
	ONCE [10]	1.2908	1.8591	0.0939	0.0803
	PURE [4]	1.6472	2.1435	0.0648	0.0558
	TRIPLE (no refine)	0.4288	0.7859	0.1282	0.1111
	TRIPLE (refine)	0.4000	0.7430	0.1294	0.1113
	TRIPLE (refine+update)	**0.3744**	**0.6664**	**0.1340**	**0.1159**
Llama-3.1-8B-Instruct	zero-shot	0.3992	0.7462	0.1045	0.0914
	ICL-Before	0.3938	0.7463	0.1312	0.1152
	PAG [17]	0.3472	0.6794	0.0867	0.0745
	ONCE [10]	1.2916	1.8682	0.0851	0.0728
	PURE [4]	1.1888	1.8665	0.0689	0.0599
	TRIPLE (no refine)	0.3470	0.6696	0.1344	0.1174
	TRIPLE (refine)	0.3304	0.6518	0.1353	0.1193
	TRIPLE (refine+update)	**0.3297**	**0.6269**	**0.1370**	**0.1201**
Gemma-3-27B-Instruct	zero-shot	0.5216	0.8000	0.0604	0.0521
	ICL-Before	0.4584	0.8173	0.0617	0.0534
	PAG [17]	0.3848	0.7122	0.0482	0.0412
	ONCE [10]	1.8424	2.3519	0.0379	0.0329
	PURE [4]	1.3996	2.0247	0.0400	0.0345
	TRIPLE (no refine)	0.4316	0.7785	0.1395	0.1255
	TRIPLE (refine)	0.4368	0.8020	0.1411	0.1260
	TRIPLE (refine+update)	**0.3780**	**0.7023**	**0.1444**	**0.1292**
Qwen2.5-32B-Instruct	zero-shot	0.4504	0.7424	0.1359	0.1182
	ICL-Before	0.4348	0.7481	0.1308	0.1138
	PAG [17]	0.3832	0.7105	0.1195	0.1026
	ONCE [10]	1.8364	2.3586	0.0934	0.0791
	PURE [4]	0.4912	0.7940	0.0927	0.0792
	TRIPLE (no refine)	0.4100	0.7108	**0.1403**	0.1235
	TRIPLE (refine)	0.3764	0.6708	0.1317	0.1152
	TRIPLE (refine+update)	**0.3586**	**0.6616**	0.1374	**0.1246**

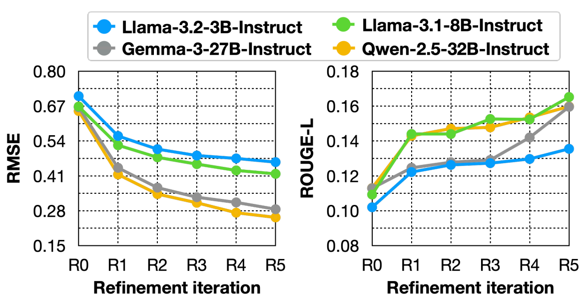

Figure 2: Impact of profile refinement iterations.

TRIPLE to adapt to evolving user states, a clear advantage over static profiling approaches.

5.2 Role of Profile Refinement

Figure 2 shows how our prediction–evaluation–refinement cycle effectively captures and refines users' latent variables across all four models and both datasets. The continuous performance improvements observed across refinement iterations validate that the errors identified during the evaluation phase successfully guide adjustments to the representations of attitude, subjective norms, and perceived behavioral control.

Notably, models with relatively larger parameter sizes, such as Gemma-3-27B-Instruct and Qwen-2.5-32B-Instruct exhibit more substantial improvements on the LaMP-3 dataset, suggesting their greater capacity to model the complex interactions between TPB components and observed user behaviors. These results confirm that the refinement process progressively enhances the model's ability to capture users' core latent variables. Each refinement session allows the framework to better align the TPB representations with actual users' latent variables, demonstrating that our theory-driven approach can effectively extract and model the psychological and social factors underlying user decisions through iterative updates.

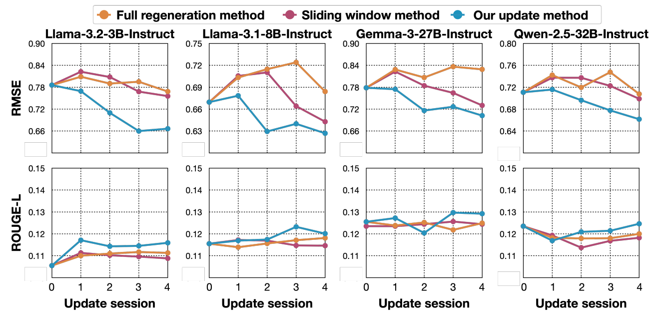

Figure 3: Impact of TRIPLE's profile update.

5.3 Role of Profile Update

Figure 3 presents the results of comparing different update strategies for TPB profile generation across the two datasets. Our proposed TPB-based update method (in blue) achieved the highest performance across all models and metrics, showing progressive improvements in both MAE and RMSE as the update rounds proceeded. In contrast, the Full regeneration method (in orange) tended to stagnate or worsen over successive rounds, suggesting that recreating the profile from scratch fails to capture the evolving users' latent variables. The Sliding window approach (in purple) initially exhibited competitive performance but declined after Round 3, indicating that focusing solely on recent information is insufficient to model the long-term latent variables of users.

6 Conclusion

In this paper, we introduced TRIPLE, a dynamic profiling framework that integrates the Theory of Planned Behavior into LLM-based user modeling. By extracting and refining TPB's core latent variables—attitude, subjective norms, and perceived behavioral control, TRIPLE addresses key limitations of existing profiling methods focused on surface-level preferences. Experiments demonstrated consistent performance gains across multiple LLMs, with the *prediction-evaluation-refinement cycle* playing a crucial role in capturing evolving user states. As the first work to bridge established behavioral theory with modern LLM capabilities for user profiling, TRIPLE opens new avenues for more sophisticated personalization systems. Future work may extend TRIPLE to multi-modal data, additional psychological theories, and broader application domains.

Acknowledgments

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No.RS-2020-II201373, Artificial Intelligence Graduate School Program(Hanyang University))

7 GenAI Usage Disclosure

No generative AI tools were used in any aspect of this research, including idea formulation, experimentation, data analysis, or paper writing.

References

[1] I Ajzen. 1991. The theory of planned behavior. *Organizational Behavior and Human Decision Processes* 50, 2 (1991), 179–211.

[2] Mohammed Fadhel Aljunid, DH Manjaiah, Mohammad Kazim Hooshmand, Wasim A Ali, Amrithkala M Shetty, and Sadiq Qaid Alzoubah. 2025. A collaborative filtering recommender systems: Survey. *Neurocomputing* 617 (2025), 128718.

[3] Hind I Alshbanat, Hafida Benhidour, and Said Kerrache. 2025. A survey of latent factor models in recommender systems. *Information Fusion* 117 (2025), 102905.

[4] Seunghwan Bang and Hwanjun Song. 2025. LLM-based User Profile Management for Recommender System. *arXiv preprint arXiv:2502.14541* (2025).

[5] Russell W Belk. 1975. Situational variables and consumer behavior. *Journal of Consumer research* 2, 3 (1975), 157–164.

[6] Moshe Ben-Akiva, Joan Walker, Adriana T Bernardino, Dinesh A Gopinath, Taka Morikawa, and Amalia Polydoropoulou. 2002. Integration of choice and latent variable models. *Perpetual motion: Travel behaviour research opportunities and application challenges* 2002 (2002), 431–470.

[7] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783* (2024).

[8] Lorena Torres Lahoz, Francisco Camara Pereira, Georges Sfeir, Ioanna Arkoudi, Mayara Moraes Monteiro, and Carlos Lima Azevedo. 2023. Attitudes and latent class choice models using machine learning. *Journal of choice modelling* 49 (2023), 100452.

[9] Jinming Li, Wentao Zhang, Tian Wang, Guanglei Xiong, Alan Lu, and Gerard Medioni. 2023. GPT4Rec: A generative framework for personalized recommendation and user interests interpretation. *arXiv preprint arXiv:2304.03879* (2023).

[10] Qijiong Liu, Nuo Chen, Tetsuya Sakai, and Xiao-Ming Wu. 2024. Once: Boosting content-based recommendation with both open-and closed-source large language models. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*. 452–461.

[11] Zizhou Liu, Ziwei Gong, Lin Ai, Zheng Hui, Run Chen, Colin Wayne Leach, Michelle R Greene, and Julia Hirschberg. 2025. The Mind in the Machine: A Survey of Incorporating Psychological Theories in LLMs. *arXiv preprint arXiv:2505.00003* (2025).

[12] Hanjia Lyu, Song Jiang, Hanqing Zeng, Yinglong Xia, Qifan Wang, Si Zhang, Ren Chen, Christopher Leung, Jiajie Tang, and Jiebo Luo. 2023. Llm-rec: Personalized recommendation via prompting large language models. *arXiv preprint arXiv:2307.15780* (2023).

[13] Lucie Charlotte Magister, Katherine Metcalf, Yizhe Zhang, and Maartje ter Hoeve. 2024. On the Way to LLM Personalization: Learning to Remember User Conversations. *arXiv preprint arXiv:2411.13405* (2024).

[14] Nusrat Jahan Prottasha, Md Kowsher, Hafijur Raman, Israt Jahan Anny, Prakash Bhat, Ivan Garibay, and Ozlem Garibay. 2025. User Profile with Large Language Models: Construction, Updating, and Benchmarking. *arXiv preprint arXiv:2502.10660* (2025).

[15] Erasmo Purificato, Ludovico Boratto, and Ernesto William De Luca. 2024. User modeling and user profiling: A comprehensive survey. *arXiv preprint arXiv:2402.09660* (2024).

[16] Hongjin Qian, Zhicheng Dou, Yutao Zhu, Yueyuan Ma, and Ji-Rong Wen. 2021. Learning implicit user profile for personalized retrieval-based chatbot. In *proceedings of the 30th ACM international conference on Information & Knowledge Management*. 1467–1477.

[17] Chris Richardson, Yao Zhang, Kellen Gillespie, Sudipta Kar, Arshdeep Singh, Zeynab Raeesy, Omar Zia Khan, and Abhinav Sethy. 2023. Integrating summarization and retrieval for enhanced personalization via large language models. *arXiv preprint arXiv:2310.20081* (2023).

[18] Alireza Salemi, Sheshera Mysore, Michael Bendersky, and Hamed Zamani. 2024. LaMP: When Large Language Models Meet Personalization. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 7370–7392.

[19] Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl. 2001. Item-based collaborative filtering recommendation algorithms. In *Proceedings of the 10th international conference on World Wide Web*. 285–295.

[20] Jyrki Suomala. 2020. The consumer contextual decision-making model. *Frontiers in Psychology* 11 (2020), 570430.

[21] An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zhihao Fan. 2024. Qwen2 Technical Report. *arXiv preprint arXiv:2407.10671* (2024).

[22] Fan Yang, Zheng Chen, Ziyan Jiang, Eunah Cho, Xiaojiang Huang, and Yanbin Lu. 2023. Palr: Personalization aware llms for recommendation. *arXiv preprint arXiv:2305.07622* (2023).

[23] Jiarui Zhang. 2024. Guided profile generation improves personalization with llms. *arXiv preprint arXiv:2409.13093* (2024).

[24] Weixiang Zhao, Xingyu Sui, Yulin Hu, Jiahe Guo, Haixiao Liu, Biye Li, Yanyan Zhao, Bing Qin, and Ting Liu. 2025. Teaching Language Models to Evolve with Users: Dynamic Profile Modeling for Personalized Alignment. *arXiv preprint arXiv:2505.15456* (2025).

[25] Xiangyu Zhao, Maolin Wang, Xinjian Zhao, Jiansheng Li, Shucheng Zhou, Dawei Yin, Qing Li, Jiliang Tang, and Ruocheng Guo. 2023. Embedding in recommender systems: A survey. *arXiv preprint arXiv:2310.18608* (2023).

[26] Joyce Zhou, Yijia Dai, and Thorsten Joachims. 2024. Language-based user profiles for recommendation. *arXiv preprint arXiv:2402.15623* (2024).