PUB: An LLM-Enhanced Personality-Driven User Behaviour Simulator for Recommender System Evaluation

Anonymous Author(s)

Abstract

Traditional offline evaluation methods for recommender systems struggle to capture the complexity of modern platforms due to sparse behavioural signals, noisy data, and limited modelling of user personality traits. While simulation frameworks can generate synthetic data to address these gaps, existing methods fail to replicate behavioural diversity, limiting their effectiveness. To overcome these challenges, we propose the **P**ersonality-driven **U**ser **B**ehaviour Simulator (PUB), an LLM-based simulation framework that integrates the Big Five personality traits to model personalised user behaviour. PUB dynamically infers user personality from behavioural logs (e.g., ratings, reviews) and item metadata, then generates synthetic interactions that preserve statistical fidelity to real-world data. Experiments on the Amazon review dataset show that logs generated by PUB closely align with real user behaviour and reveal meaningful associations between personality traits and recommendation outcomes. These results highlight the potential of the personality-driven simulator to advance recommender system evaluation, offering scalable, controllable, high-fidelity alternatives to resource-intensive real-world experiments. [1]

CCS Concepts

• **Information systems** → **Recommender systems**.

Keywords

LLM-based Agent, Simulation, Recommender Systems Evaluation, User Behaviour, Big Five Personality Traits

ACM Reference Format:

Anonymous Author(s). 2018. PUB: An LLM-Enhanced Personality-Driven User Behaviour Simulator for Recommender System Evaluation. In *Woodstock '18: ACM Symposium on Neural Gaze Detection, June 03–05, 2018, Woodstock, NY*. ACM, New York, NY, USA, 5 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Modern recommender systems (RSs) face critical challenges in evaluation methodologies, particularly as traditional offline datasets struggle to capture the dynamic complexity of user interactions on contemporary platforms. These datasets often lack granular behavioural signals (e.g., personality-driven decision-making) and exhibit biases from sparse or noisy logs, limiting their utility for robust system optimisation. While user studies and real-world datasets provide valuable insights, they are resource-intensive, suffer from uncontrollable confounding variables, and fail to reproduce differentiable evaluation results for modern RSs (e.g., short video platforms). Recent advances in large language models (LLMs) offer promising avenues for simulating user behaviour [5, 27], yet existing frameworks remain inadequate in modelling individual differences, such as personality traits, which significantly influence preferences and engagement patterns.

The integration of personality-driven modelling into RSs [4] has demonstrated benefits, including improved accuracy and cold-start mitigation. The Big Five personality model [7, 22] (i.e., Openness (O), Conscientiousness (C), Extraversion (E), Agreeableness (A), Neuroticism (N)) has emerged as a gold standard due to its psychological validity and cross-cultural reproducibility. However, simulating these personality traits at scale remains challenging. Prior work in personality computing primarily focuses on inferring traits from static user data (e.g., social media posts [1, 17]), while LLM-based simulations often prioritise generic behavioural patterns over trait-specific dynamics. For instance, research [14, 23, 25] reveal that LLMs like GPT-4 exhibit distinct Big Five profiles under varying prompts, but their stability and applicability to recommendation tasks remain understudied.

Existing simulation tools face two key limitations: (1) *Low fidelity*: Generated behaviours often deviate from real-world statistical distributions, undermining the reliability of recommender system evaluation [5, 19]. (2) *Oversimplified personalisation*: Collaborative filtering hybrids or Markov decision processes based methods fail to capture the nuanced correlations between user traits and behaviours [13, 27]. Despite their potential to mitigate ethical and logistical challenges associated with real-user experiments, such as privacy concerns and experimental costs, these limitations hinder the widespread adoption of simulation tools for large-scale recommender system evaluation [5, 27].

To address these challenges, we propose **P**ersonality-driven **U**ser **B**ehaviour Simulator (PUB), which synergises LLM-driven personality inference with dynamic behaviour simulation. Our approach is based on two key insights. First, user interactions encode rich psychological signals that serve as proxies for personality traits [1, 17, 26]. PUB leverages these digital footprints along with item metadata to enhance LLM-based trait inference, improving the fidelity of simulated behaviours. Second, recent studies [14, 22, 23, 25] show that LLMs can simulate stable personality profiles when guided by structured prompts. PUB extends this by grounding simulations in empirically validated Big Five correlations, ensuring psychologically coherent behavioural outputs. In summary, our paper makes the following contributions:

[1] The source code can be found at https://anonymous.4open.science/r/sigir2025-EBF6/.

Conference acronym 'XX, June 03–05, 2018, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/18/06
https://doi.org/XXXXXXX.XXXXXXX

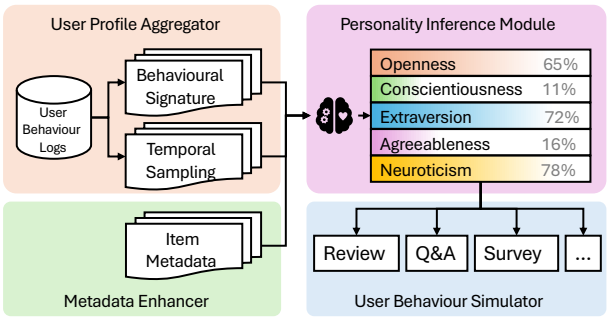


Figure 1: Overview of the proposed PUB architecture.

- **Personality-driven simulation**: We introduce a personality-based user behaviour simulation framework that dynamically maps inferred Big Five traits to probabilistic behaviour models.
- **Evaluation robustness**: Experiments show PUB-generated logs can effectively replicate performance trends of various recommendation algorithms when compared to real data.
- **Fine-grained analysis**: Our analysis reveals significant correlations between personality traits and recommendation susceptibility, offering new insights into user behaviour dynamics.

2 Methodology

The proposed PUB is a hybrid architecture designed to simulate personality-driven user behaviour for recommender system evaluation. As illustrated in Fig. 1, the framework operates in four phases: (1) **User Profile Aggregator**: Extracts user behavioural logs and computes aggregated features (e.g., purchase frequency, category preferences) using statistical functions. (2) **Metadata Enhancer**: Embeds item metadata (e.g., title, description, price) to enrich contextual information for downstream processing. (3) **Personality Inference Module**: Maps user behaviour logs and item metadata to Big Five personality traits, leveraging prompt-guided LLMs and psychometric mapping. (4) **User Behaviour Simulator**: Generates synthetic interactions conditioned on inferred traits to preserve statistical fidelity to real-world patterns in downstream tasks.

The details of each module are provided below. To illustrate the framework more clearly, we use Amazon review dataset [11] as an example in the following description. However, PUB is designed to be dataset-agnostic and applicable across various domains.

2.1 User Profile Aggregator

The User Profile Aggregator employs a hierarchical approach to construct statistical user profiles from heterogeneous behavioural data. We formalise the process through two primary stages: Behavioural Signature Extraction and Temporal Stratified Sampling.

Behavioural Signature Extraction. The core behavioural patterns are formalised through statistical feature engineering. We first aggregate all category-specific interactions into a unified dataset, denoted as $\mathcal{D} = \bigcup \{\mathcal{D}_c\}_{c \in C}$, where C represents the complete set of Amazon categories. Subsequently, we extract user behavioural logs, comprising sequential interactions $d = (i_k, t_k, r_k, c_k)$, where i_k denotes the item ID, t_k the timestamp, r_k the rating (1–5 stars), and c_k the item category. The module computes aggregated features (e.g., purchase frequency, purchase rhythm, category preferences, and price tier distribution) via domain-specific functions.

For example, to model purchase rhythm, we employ circular statistics to capture periodic behavioural patterns based on interaction timestamps. Each interaction timestamp t_i is mapped onto a unit circle using an angular transformation: $\theta_i = \frac{2\pi t_i}{y \times 24 \times 60 \times 60}$, where y denotes the number of days in the observed cyclic behaviour. This transformation normalises timestamps within a 2π range, ensuring that interactions occurring at the same relative position but on different cycles are mapped to similar angular positions. The overall purchase rhythm is then computed as the mean resultant vector:

$$\gamma = \frac{1}{n} \sum_{i=1}^{n} e^{j\theta_i}, \tag{1}$$

where n represents the number of interactions.

The interval entropy is computed to measure regularity using Shannon entropy [18] over purchase intervals:

$$H_{\text{int}} = -\sum_{t \in \mathcal{T}} p(t) \log p(t), \tag{2}$$

where $p(t)$ denotes the probability of purchase interval t, and $\mathcal{T}$ represents the set of all intervals [2].

Finally, the module computes user-specific statistical features, denoted as $\mathbf{S}_u = \{\gamma, H_{\text{int}}, \dots\} \in \mathbb{R}^{d_u}$, where d_u represents the feature dimensionality. This module acts as a bridge between raw behavioural logs and trait inference, providing interpretable user representations.

Temporal Stratified Sampling. Given the longitudinal nature of user interactions, we partition the interaction records into K temporal bins using exponentially increasing time windows. This adaptive sampling strategy maintains chronological integrity while efficiently reducing computational complexity:

$$\mathcal{D}' = \bigcup_{k=1}^{K} \text{Sample}\Big(\mathcal{D}|_{\Delta t_k}, \eta\Big), \tag{3}$$

where $\mathcal{D}|_{\Delta t_k}$ denotes the subset of $\mathcal{D}$ corresponding to the temporal bin of length Δt_k, η represents the sampling length for each bin, and K is the total number of bins.

The adaptive time windows are defined as:

$$\Delta t_k = \begin{cases} 1 \text{ week}, & \text{if } t_{\text{end}} - t_{\text{start}} \le 1 \text{ year}, \\ 1 \text{ month}, & \text{if } 1 \text{ year} < (t_{\text{end}} - t_{\text{start}}) \le 3 \text{ years}, \\ 1 \text{ quarter}, & \text{otherwise}. \end{cases} \tag{4}$$

The thresholds in Eq. (4) are determined empirically.

2.2 Metadata Enhancer

The Metadata Enhancer module enriches item metadata (e.g., title, description, price) by incorporating it into LLM prompt-guided contexts. The metadata is then modulated by user-specific statistical features to enhance contextual information. Formally, let $\mathbf{m}_i \in \mathbb{R}^{d_i}$ denote the raw metadata for item i, where $i \in \mathcal{D}'$, and let $\mathbf{M}_u = [\mathbf{m}_i]_{i \in \mathcal{I}_u}$, where $\mathcal{I}_u$ represents the set of items interacted with by user u. The enhanced metadata context $\mathbf{C}_u$ is then obtained through a prompt-guided fusion function $g(\cdot)$ as follows:

$$\mathbf{C}_u = g\left(\mathbf{M}_u, \phi(\mathbf{S}_u)\right), \tag{5}$$

[2] Please refer to our code for other feature definitions.

where $\phi(\cdot)$ transforms statistical features by normalising and scaling them for compatibility with the raw metadata. The function $g(\cdot)$, implemented via an LLM prompt, integrates enriched metadata with transformed statistical cues. The whole process yields a dynamic, task-specific representation $\mathbf{C}_u$ that emphasises personality-relevant signals.

2.3 Personality Inference Module

The Personality Inference Module employs a psychometric mapping function to estimate users' Big Five personality traits from the enhanced context: $\mathbf{T}_u = f(C_u) = \{O, C, E, A, N\}$. To disentangle the model design from its dependency on specific LLM reasoning abilities [28], we leverage well-established psycholinguistic correlates to guide the trait inference process. Drawing upon established correlations between digital footprints and psychological constructs [1, 7, 17, 26], we formalise this mapping as follows:

- **Openness**: Category entropy [26] and metaphor density.
- **Conscientiousness**: Review length consistency, rating deviation from category averages, and purchase rhythm regularity.
- **Extraversion**: Social reference frequency (e.g., "we", "gift" via the LIWC-22 lexicon [3, 6]).
- **Agreeableness**: Positive sentiment ratio (e.g., VADER [12]) and politeness markers [6].
- **Neuroticism**: Negative emotion volatility [6].

The prompt is structured, to encapsulate the user's behavioural statistics and metadata context. The response contains the inferred trait scores, which subsequently guide the behaviour simulation process, enabling diverse and personalised interactions.

2.4 User Behaviour Simulator

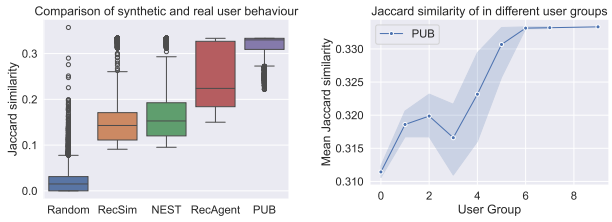
The User Behaviour Simulator generates synthetic interactions conditioned on inferred trait distributions, preserving statistical fidelity to real-world patterns. We assign the inferred personality to the LLM-based agent, which then performs various downstream tasks, such as completing Q&A tasks, filling out questionnaires, and providing feedback on recommendations.

In this study, we focus on the recommendation task, where the agent generates synthetic interactions based on the inferred personality. These interactions are then evaluated against real-world data using standard metrics (e.g., nDCG). While our scope is limited to this application, the proposed PUB framework is broadly applicable and can be extended to other domains.

3 Experiments

3.1 Experimental Setup

Our experiments use the Amazon Review Dataset [11], which contains 571.54 million interactions across 30 distinct categories from 1996 to 2023. To ensure methodological rigour, we apply a three-stage preprocessing pipeline informed by established practices in sparse recommender systems [24]. First, interactions across all categories are aggregated to model cross-domain behavioural diversity. Second, users and items with fewer than 20 interactions are filtered to reduce sparsity-induced noise. Finally, to preserve temporal dynamics [20], the dataset is chronologically partitioned into training set $\mathcal{D}_{\text{tr}}$ and testing set $\mathcal{D}_{\text{te}}$, following an 80%-20% split.



(a) (b)

Figure 2: (a) Comparison of synthetic and real user behaviour sequences; (b) Jaccard similarity of different user groups.

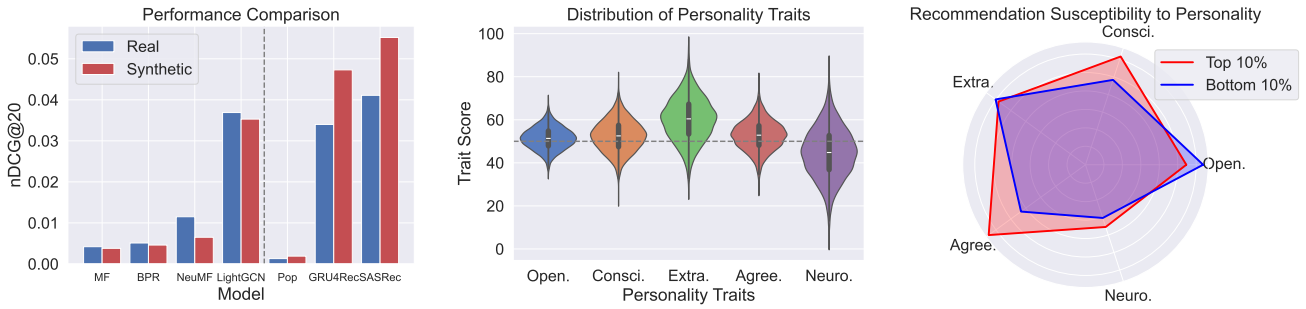
3.2 Research Questions and Analysis

3.2.1 ***RQ 1:*** *Can the proposed PUB generate synthetic behaviour sequences that closely resemble real user interactions?* Building on [27], we derive each user's initial personality distribution and shopping behaviour patterns from their interaction history $\mathcal{I}_u$ in the training dataset $\mathcal{D}_{\text{tr}}$. The test set interactions $\mathcal{P}_u$ serve as the ground truth. To isolate the effects of the recommendation algorithm, we employ a proxy recommender. At each iteration t, this model constructs a mock recommendation list $\mathcal{L}_t$ for user u: $\mathcal{L}_t = \{i_p, i_{n1}, i_{n2}, \ldots, i_{nk}\}, k = 9$. Here, i_p is the positive sample, chronologically drawn from $\mathcal{P}_u$, while each i_{nk} is a negative sample randomly selected from $\mathcal{I}_u^- = \{i \mid i \notin \mathcal{I}_u\}$. The user agent then selects the most relevant item from $\mathcal{L}_t$ based on its inferred personality, forming the synthetic sequence $\mathcal{S}_u$. We measure the similarity between $\mathcal{S}_u$ and the true interaction sequence $\mathcal{P}_u$ using Jaccard similarity [2], where a higher value indicates better alignment with real behaviour.

We compare PUB with four baseline simulation frameworks: *Random Sampling*, which randomly selects items without user preferences; *RecSim* [13], a Markov Chain-based simulation; *NEST* [19], an extension of RecSim incorporating user need states; *RecAgent* [27], an LLM-based simulation that randomly assigns user profiles (e.g., age, gender, occupation).

The results in Fig. 2a show that PUB-generated sequences align closely with real user interactions, achieving an average Jaccard similarity of 0.31. While RecAgent performs well in top cases, its overall performance is less stable, with an average Jaccard similarity of 0.24. These findings suggest that PUB effectively replicates real-world user behaviour, providing a reliable foundation for further evaluation. To assess the impact of interaction frequency, we group users into 10 clusters (0 is the least frequent and 9 is the most frequent) and compute Jaccard similarity for each. Fig. 2b shows a clear trend: Jaccard similarity increases with interaction frequency, indicating that richer interaction histories contribute to more accurate personality and behaviour modelling.

3.2.2 ***RQ 2:*** *Can the proposed PUB accurately evaluate the performance of different recommendation algorithms?* In this experiment, synthetic user behaviour sequences are generated following the aforementioned methodology and then divided into a training set, $\mathcal{S}_{\text{tr}}$, and a test set, $\mathcal{S}_{\text{te}}$. To ensure a fair comparison, the synthetic test set is matched in size to the original test data. A variety of recommendation algorithms are evaluated, including *Pop*, *MF* [16],



(a) (b) (c)

Figure 3: (a) Performance comparison; (b) Distribution of personality traits; (c) Recommendation susceptibility to personality.

BPR [21], *NeuMF* [9], *LightGCN* [8], *GRU4Rec* [10], and *SASRec* [15]. Their performance is assessed using ranking metrics (e.g., nDCG@20) on both real and synthetic test sets.

As shown in Fig. 3a, the performance of each algorithm on the synthetic test set closely mirrors that on the real test set, suggesting that the proposed simulation model is a viable alternative for evaluating recommendation algorithms. Notably, MF, BPR, NeuMF, and LightGCN generally perform worse on the synthetic test set, while Pop, GRU4Rec, and SASRec perform better. The former group primarily relies on collaborative filtering, leveraging the user-item interaction matrix, whereas the latter (except Pop) focuses on sequential recommendation, capturing sequential or attentional patterns in user interactions.

This discrepancy likely arises because the synthetic data is generated based on users' inferred *personality traits* rather than the *collaborative behaviour* observed in real-world data. Pop achieves higher performance on the synthetic test set as PUB links users' personality traits with their preference for item popularity during profile construction, aligning with the recommendation strategy of Pop. These findings suggest a potential enhancement for our simulation framework by incorporating social influences and collaborative signals among users.

3.2.3 ***RQ 3:*** *What are the distributions of the Big Five personality traits, and how do these traits relate to shopping behaviour?* An association analysis correlates the inferred personality trait distributions with users' shopping behaviour patterns. This analysis seeks to determine which personality traits are more indicative of users who are prone to leaving detailed shopping traces and reviews. The results are expected to provide insights into the behavioural tendencies associated with different personality profiles.

As shown in Fig. 3b, the distribution of the Big Five personality traits within the Amazon Review Dataset appears relatively balanced, with Extraversion being the most prevalent trait. Notably, the average Neuroticism score is significantly lower than that of other traits, suggesting that users who leave reviews on Amazon tend to be less emotionally stable. This aligns with the hypothesis that writing a review represents a stronger interaction signal than making a purchase, as it requires greater effort and emotional investment. Users with *extreme* purchase experiences are more likely to leave reviews and express emotions in their feedback, potentially correlating with lower Neuroticism scores. Conversely, users with *moderate* purchase experiences may be more inclined to provide ratings or refrain from leaving feedback altogether, making it more challenging to infer their personality traits. This poses a challenge for recommender systems in ensuring fair, consistent, and high-quality recommendations.

3.2.4 ***RQ 4:*** *Which personality traits are favoured by recommendation algorithm?* In this analysis, we investigate the relationship between personality traits and recommendation outcomes by selecting users with extreme performance (top 10% and bottom 10% in terms of nDCG@20). Note that we employ GRU4Rec as the recommendation algorithm. As shown in Fig. 3c, users with high Agreeableness and Conscientiousness scores tend to receive better recommendations, whereas, surprisingly, those with high Openness scores exhibit lower recommendation accuracy. This suggests that users who are more open to new experiences may disrupt the algorithm's inference process by deviating from established behavioural patterns. In contrast, individuals with higher levels of agreeableness and conscientiousness are more likely to adhere to recommendations and provide detailed feedback, thereby enhancing the system's performance.

4 Conclusion

In this paper, we introduce the Personality-driven User Behaviour Simulator (PUB), a simulation framework that integrates the Big Five personality traits into user behaviour modelling. Our experiments show that the synthetic logs generated by PUB closely resemble real user interactions, providing a robust and scalable platform for evaluating recommendation algorithms. The model effectively measures the performance of various recommendation approaches, offering a high-fidelity alternative to costly real-world experiments. Furthermore, our analysis of personality distributions and recommendation susceptibility provides valuable insights into how different personality profiles influence recommendation outcomes. Future research focuses on refining these methodologies and exploring their applications in live recommendation environments, with an emphasis on incorporating social influence models to further enhance the simulation framework.

References

[1] Danny Azucar, Davide Marengo, and Michele Settanni. 2018. Predicting the Big 5 personality traits from digital footprints on social media: A meta-analysis. *Personality and individual differences* 124 (2018), 150–159.

[2] Sujoy Bag, Sri Krishna Kumar, and Manoj Kumar Tiwari. 2019. An efficient recommendation generation using relevant Jaccard similarity. *Information Sciences* 483 (2019), 53–64. https://doi.org/10.1016/j.ins.2019.01.023

[3] Ryan L Boyd, Ashwini Ashokkumar, Sarah Seraj, and James W Pennebaker. 2022. The development and psychometric properties of LIWC-22. *Austin, TX: University of Texas at Austin* 10 (2022).

[4] Sahraoui Dhelim, Liming Chen, Nyothiri Aung, Wenyin Zhang, and Huansheng Ning. 2023. A hybrid personality-aware recommendation system based on personality traits and types models. *J. Ambient Intell. Humaniz. Comput.* 14, 9 (2023), 12775–12788. https://doi.org/10.1007/s12652-022-04200-5

[5] Chen Gao, Xiaochong Lan, Nian Li, Yuan Yuan, Jingtao Ding, Zhilun Zhou, Fengli Xu, and Yong Li. 2024. Large language models empowered agent-based modeling and simulation: A survey and perspectives. *Humanities and Social Sciences Communications* 11, 1 (2024), 1–24.

[6] Lewis R Goldberg. 1981. Language and individual differences: The search for universals in personality lexicons. *Review of Personality and Social Psychology/Sage* (1981).

[7] Lewis R Goldberg. 1992. The development of markers for the Big-Five factor structure. *Psychological assessment* 4, 1 (1992), 26.

[8] Xiangnan He, Kuan Deng, Xiang Wang, Yan Li, Yong-Dong Zhang, and Meng Wang. 2020. LightGCN: Simplifying and Powering Graph Convolution Network for Recommendation. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020*. 639–648. https://doi.org/10.1145/3397271.3401063

[9] Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. 2017. Neural Collaborative Filtering. In *Proceedings of the 26th International Conference on World Wide Web, WWW 2017, Perth, Australia, April 3-7, 2017*. 173–182. https://doi.org/10.1145/3038912.3052569

[10] Balázs Hidasi, Alexandros Karatzoglou, Linas Baltrunas, and Domonkos Tikk. 2016. Session-based Recommendations with Recurrent Neural Networks. In *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*. http://arxiv.org/abs/1511.06939

[11] Yupeng Hou, Jiacheng Li, Zhankui He, An Yan, Xiusi Chen, and Julian J. McAuley. 2024. Bridging Language and Items for Retrieval and Recommendation. *CoRR* abs/2403.03952 (2024). https://doi.org/10.48550/arXiv.2403.03952

[12] Clayton J. Hutto and Eric Gilbert. 2014. VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. In *Proceedings of the Eighth International Conference on Weblogs and Social Media, ICWSM 2014, Ann Arbor, Michigan, USA, June 1-4, 2014*. http://www.aaai.org/ocs/index.php/ICWSM/ICWSM14/paper/view/8109

[13] Eugene Ie, Chih-Wei Hsu, Martin Mladenov, Vihan Jain, Sanmit Narvekar, Jing Wang, Rui Wu, and Craig Boutilier. 2019. RecSim: A Configurable Simulation Platform for Recommender Systems. *CoRR* abs/1909.04847 (2019). http://arxiv.org/abs/1909.04847

[14] Guangyuan Jiang, Manjie Xu, Song-Chun Zhu, Wenjuan Han, Chi Zhang, and Yixin Zhu. 2023. Evaluating and Inducing Personality in Pre-trained Language Models. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*. http://papers.nips.cc/paper_files/paper/2023/hash/21f7b745f73ce0d1f9bcea7f40b1388e-Abstract-Conference.html

[15] Wang-Cheng Kang and Julian J. McAuley. 2018. Self-Attentive Sequential Recommendation. In *IEEE International Conference on Data Mining, ICDM 2018, Singapore, November 17-20, 2018*. 197–206. https://doi.org/10.1109/ICDM.2018.00035

[16] Yehuda Koren, Robert M. Bell, and Chris Volinsky. 2009. Matrix Factorization Techniques for Recommender Systems. *Computer* 42, 8 (2009), 30–37. https://doi.org/10.1109/MC.2009.263

[17] Renaud Lambiotte and Michal Kosinski. 2014. Tracking the digital footprints of personality. *Proc. IEEE* 102, 12 (2014), 1934–1939. https://doi.org/10.1109/JPROC.2014.2359054

[18] Jianhua Lin. 1991. Divergence measures based on the Shannon entropy. *IEEE Transactions on Information theory* 37, 1 (1991), 145–151. https://doi.org/10.1109/18.61115

[19] Chenglong Ma, Yongli Ren, Pablo Castells, and Mark Sanderson. 2022. NEST: Simulating Pandemic-like Events for Collaborative Filtering by Modeling User Needs Evolution. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management, Atlanta, GA, USA, October 17-21, 2022*. 1430–1440. https://doi.org/10.1145/3511808.3557407

[20] Chenglong Ma, Yongli Ren, Pablo Castells, and Mark Sanderson. 2024. Temporal Conformity-aware Hawkes Graph Network for Recommendations. In *Proceedings of the ACM on Web Conference 2024, WWW 2024, Singapore, May 13-17, 2024*. 3185–3194. https://doi.org/10.1145/3589334.3645354

[21] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. 2009. BPR: Bayesian Personalized Ranking from Implicit Feedback. In *UAI 2009, Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence, Montreal, QC, Canada, June 18-21, 2009*. 452–461. https://www.auai.org/uai2009/papers/UAI2009_0139_48141db02b9f0b02bc7158819ebfa2c7.pdf

[22] Sonia Roccas, Lilach Sagiv, Shalom H Schwartz, and Ariel Knafo. 2002. The big five personality factors and personal values. *Personality and social psychology bulletin* 28, 6 (2002), 789–801.

[23] Mustafa Safdari, Greg Serapio-García, Clément Crepy, Stephen Fitz, Peter Romero, Luning Sun, Marwa Abdulhai, Aleksandra Faust, and Maja J. Mataric. 2023. Personality Traits in Large Language Models. *CoRR* abs/2307.00184 (2023). https://doi.org/10.48550/arXiv.2307.00184

[24] Monika Singh. 2020. Scalability and sparsity issues in recommender datasets: a survey. *Knowledge and Information Systems* 62, 1 (2020), 1–43. https://doi.org/10.1007/s10115-018-1254-2

[25] Aleksandra Sorokovikova, Sharwin Rezagholi, Natalia Fedorova, and Ivan Yamshchikov. 2024. LLMs Simulate Big5 Personality Traits: Further Evidence. In *Proceedings of the 1st Workshop on Personalization of Generative AI Systems (PERSONALIZE 2024)*. 83–87.

[26] Wen-Chin Tsao and Hung-Ru Chang. 2010. Exploring the impact of personality traits on online shopping behavior. *African journal of business management* 4, 9 (2010), 1800.

[27] Lei Wang, Jingsen Zhang, Hao Yang, Zhi-Yuan Chen, Jiakai Tang, Zeyu Zhang, Xu Chen, Yankai Lin, Hao Sun, Ruihua Song, et al. 2024. User Behavior Simulation with Large Language Model-based Agents for Recommender Systems. *ACM Transactions on Information Systems* (2024).

[28] Qian Wang, Jiaying Wu, Zhenheng Tang, Bingqiao Luo, Nuo Chen, Wei Chen, and Bingsheng He. 2025. What Limits LLM-based Human Simulation: LLMs or Our Design? *arXiv preprint arXiv:2501.08579* (2025).