USHB: A Unified Framework for Simulating Human Behaviors in Agent Society through User-and-Item Modeling

Renhuo Zhao*
Hailong Yang*
Mingxian Gu
zhaorenhuo@stu.jiangnan.edu.cn
yanghailong@stu.jiangnan.edu.cn
gumingxian@stu.jiangnan.edu.cn
Jiangnan University
Wuxi, Jiangsu, China

Jianqi Wang
Wu Long
Zhaohong Deng†
6243114031@stu.jiangnan.edu.cn
6243112036@stu.jiangnan.edu.cn
dengzhaohong@jiangnan.edu.cn
Jiangnan University
Wuxi, Jiangsu, China

ABSTRACT

Emulating real-world human behaviors within Artificial Intelligent (AI) agent systems continues to be a formidable challenge, as existing methodologies frequently have difficulty integrating the intricacies of real-world scenarios and personal preferences.To address this issue, we propose USHB, a mulit-agent framework that emphasizes advanced user and item modeling along with communication style simulation. USHB consists of 3 modules: a Knowledge-Mining Module (KMM), a User-and-Item Modeling Module (UIMM), and a Reasoning Module (RM). USHB utilizes Large Language Models (LLMs) to predict responses, reviews, and ratings tailored to individuals, guaranteeing results that are both coherent and contextually appropriate. USHB is capable of delivering precise, detailed simulations that closely mimic human behavior. We evaluated USHB using datasets from Yelp, Amazon, and Goodreads, where it consistently outperformed baseline methods. Moreover, USHB demonstrated strong generalization and maintained stable performance across a variety of model configurations. These advancements make USHB a valuable contribution to dynamic, context-aware behavior simulation, achieving a top-three ranking in the 2025 AgentSociety Challenge. Our codes are available at https://github.com/jnuaipr/AgentsChallenge.

CCS CONCEPTS

• **Computing methodologies → Multi-agent planning**; **Natural language generation**; *Information extraction.*

KEYWORDS

Agent, User-and-Item Modeling, Relationship Graph, IF-THEN Rules, Reasoning, Knowledge Mining, Large Language Model

*Both authors contributed equally to this research.
†Corresponding author.

WWW Companion '25, April 28-May 2, 2025, Sydney, NSW, Australia

ACM ISBN 979-8-4007-1331-6/2025/04
https://doi.org/10.1145/3701716.3719227

ACM Reference Format:
Renhuo Zhao, Hailong Yang, Mingxian Gu, Jianqi Wang, Wu Long, and Zhaohong Deng. 2025. USHB: A Unified Framework for Simulating Human Behaviors in Agent Society through User-and-Item Modeling. In *Companion Proceedings of the ACM Web Conference 2025 (WWW Companion '25), April 28-May 2, 2025, Sydney, NSW, Australia.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3701716.3719227

1 INTRODUCTION

Over the past several decades, Artificial Intelligence (AI) has achieved significant progress, transforming industries by allowing systems to understand, predict, and react to human actions [10]. Notably, the significant advancement is the development of Large Language Models (LLMs) such as Qwen2.5-72B-instruct [11], which have proven to be highly effective in both comprehending and generating texts that mimics human language. These LLMs are proficient in handling large datasets, which makes them suitable for applications like natural language processing, answering questions, and generating content [4]. However, LLMs still face numerous challenges. **First**, LLMs lack real-world context and deep understanding of user behaviors, which can hinder their ability to generate personalized and contextually relevant outputs [20]. **Second**, LLMs are incapable of replicating the intricate human-human, human-object, and object-object relationships that exist within human society.

The AgentSociety Challenge offers a distinctive chance to address these issues, with an emphasis on the User Modeling Track. Intelligent agents utilize user modeling to emulate user behaviors, allowing for dynamic and context-sensitive data interaction. This competition invites participants to design intelligent agents that simulate user behaviors through reviews and star ratings, leveraging historical data and contextual information. The AgentSociety Challenge competition utilizes large-scale datasets from prominent online platforms like Yelp, Amazon, and Goodreads, which provide rich, real-world user interactions and preferences. These datasets serve as a foundation for modeling user behavior, offering a diverse set of scenarios to test agent performance.

To address the LLM challenge, we propose a unified multi-agent framework for simulating human behaviors through User-and-Item modeling. Our framework incorporates the corresponding tools to simulate real user reviews and ratings. Our framework incorporates independent agents for user role modeling and item modeling, capturing the complexities of user-item interactions. We designed User-Item relationship graph to simulate human society complex relationship and speed knowledge mining up. Additionally, we incorporate writing style modeling to ensure that generated reviews reflect not only user preferences but also their tone and communication style. Our framework allows for detailed simulation of user behaviors, ensuring that generated reviews and ratings align with the realistic preferences and communication patterns of the users.

2 RELATED WORKS

Advancements in Language Models for User Modeling. Initially, the text was represented using the Bag-of-Words model [18]. Models dedicated to topics, like the Latent Dirichlet Allocation (LDA) [3], were developed to identify underlying themes. Despite their efforts, they encountered challenges in accurately grasping the subtleties of meaning. Subsequently, Word2Vec derives semantic embeddings via neural networks [9]. Nevertheless, its relatively shallow structure rendered it ineffective in handling long-text patterns adequately. The advent of the Transformer architecture and pre-trained language models (PLMs) significantly advanced the field of text understanding [14]. BERT4Rec [13] harnessed bidirectional self-attention for recommendation purposes. LLM-Rec utilizes ChatGPT to enhance recommendation models by generating user-preferred topics [5, 6, 8]. LLMs are potent deep-neural-network-based models that are pre-trained on extensive text corpora [7].

Applications of LLMs in User Modeling. LLMs have found extensive applications in user modeling. In recommendation systems, they are capable of predicting user interests. PALR leverages LLMs to generate personalized recommendations [16]. In user profiling, LLMs can effectively summarize user characteristics. Rao et al. utilized LLMs for personality analysis based on the MBTI framework [12]. Additionally, LLMs prove to be highly effective in rating prediction and the detection of suspicious behavior [17, 19].

3 METHOD

To enhance agent ability of the human-like expression and review, we propose a Unified framework for Simulating Human Behaviors (USHB). USHB have three modules: Knowledge-Mining Module (KMM), User-and-Item Modeling Module (UIMM) and Reasoning Module (RM), as illustrated in Figure 1.

3.1 Knowledge-Mining Module (KMM)

Human society is a complex relational network composed of people and objects, within which vast amounts of historical and contemporary knowledge and information are stored. In the user-item interacting scenarios, USHB delves into historical data and learns prior knowledge to enhance agents' cognitive and decision-making capabilities. USHB integrates the Knowledge-Mining Module (KMM) to extract valuable user and item information.

KMM builds a User-and-Item Relationship Graph (UIRG) to facilitate efficient information access. UIRG includes user nodes, item nodes, and review edges, combining three essential data categories: user information, user reviews, and item information, as shown in Figure 2. Upon verifying the user ID and item ID, KMM utilizes the UIRG to efficiently retrieve the user's profile, their past reviews, details of the target item, and reviews from other users about the target item, as illustrated in Equation (1).

$$\{Q_u, Q_p, C_u, C_p\} = G(ID_u, ID_p) \tag{1}$$

Where ID_u is the target user u ID, ID_p is the target item p ID, C_u is the full set of historical reviews authored by the user u, C_p is the full set of the item p reviews from the other users, Q_u is the user u information, Q_p is the item p information, $G(.)$ is the retrieval function of UIRG.

3.2 User-and-Item Modeling Module (UIMM)

3.2.1 User Modeling. The User Model (UM) includes user attributes and reviewing patterns for better understand their preferences and tendencies. By examining the user's review history and account information, the module identifies key factors that shape their unique reviewing style. For example, it considers the volume of reviews contributed over time, highlighting high activity levels that suggest a strong engagement with certain item categories. It also assesses sentiment trends—such as a predominance of positive reviews, balanced neutral feedback, or frequent negative critiques—to form a comprehensive picture of the user's reviewing disposition.

In addition to sentiment, the module analyzes specific rating patterns. For instance, the user may consistently assign four-star ratings to items they consider well-made but not exceptional, while reserving five-star ratings for items that surpass expectations. This pattern provides insight into the user's criteria for quality and value. Similarly, repeated two-star ratings might indicate dissatisfaction with a particular type of product or service. Through this detailed analysis, the character modeling module constructs an informed representation of the user's reviewing behavior, enabling more personalized interactions and recommendations.

For user modeling, we begin by selecting a random subset of the user's historical reviews. This subset is defined as Equation (2).

$$S_u = RandomSubset(C_u, N_u) \tag{2}$$

Where N_u is the number of reviews to be randomly selected, S_u is the random subset of the user reviews. This step allows the modeling process to consider diverse samples of the user's past behavior, rather than focusing exclusively on their most recent or most frequent interactions. Random selection ensures that the resulting model reflects a broader understanding of the user's overall preferences and reviewing habits. With this subset S_u, the user model is constructed as Equation (3).

$$M_u = BuildModel_u(Q_u, S_u) \tag{3}$$

Where $BuildModel_u(.)$is the user modeling agent function for the selected reviews and user information, M_u is the comprehensive profile of the user u. This profile includes patterns in sentiment, preferred product categories, and linguistic tendencies.

By leveraging a randomly chosen subset, the model gains a balanced view of the user's behavior, supporting applications such as personalized recommendations and targeted content delivery.

3.2.2 Item Modeling. The item modeling method is designed to systematically analyze items within the agent society, focusing on their fundamental characteristics, market reception, and user evaluations. The Item Model (IM) integrates structured item information with aggregated review data to develop a clear understanding of

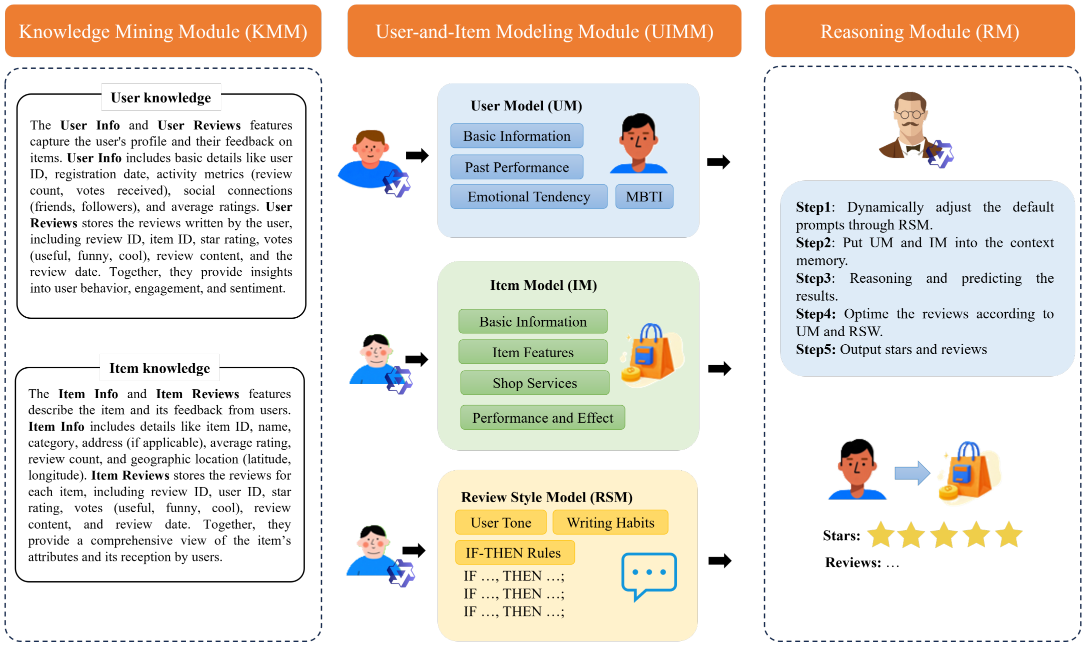

Figure 1: The Structure of USHB, including Knowledge-Mining Module (KMM), User-and-Item Modeling Module (UIMM) and Reasoning Module (RM).

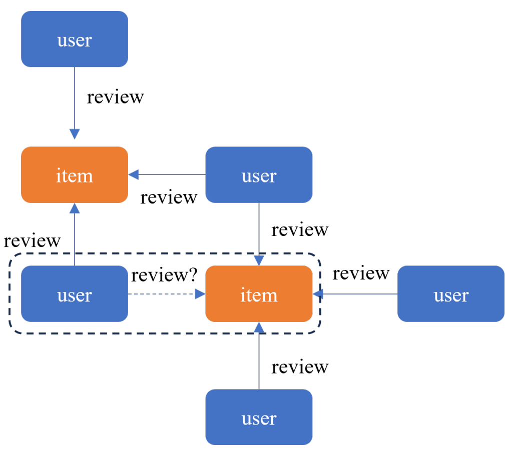

Figure 2: User-and-Item Relationship Graph (UIRG), including user node, item node and review edge.

each item's role and impact in the simulated environment.The modeling process begins with collecting detailed information about the item. This includes its name, origin, type, overall star rating, and the approximate number of user reviews. The model ensures that these foundational details are accurate and up-to-date, forming the basis for further analysis. For instance, a typical item might be a popular consumer electronic device, rated at 4.4 stars from several hundred reviews, and known for specific features such as durability, ease of use, or innovative functionality.

The item modeling approach analyzes user feedback to identify recurring trends in item strengths and weaknesses. It highlights key advantages, such as superior build quality or reliable performance, while also noting common criticisms, like high pricing or technical issues, quantifying their frequency and impact. This systematic summary offers a balanced and comprehensive view of each item.By combining detailed item attributes with nuanced user sentiment, the object modeling module provides a reliable foundation for various system components. It supports recommendation generation, quality assessment, and predictive analysis, enabling a wide range of research and application scenarios within the agent society through its structured representation of items.

Assuming there are C_p total reviews and we need to randomly pick N_p of them as shown in Equation (4).

$$S_p = RandomSubset(C_p, N_p) \tag{4}$$

Where $N_p$is the number of reviews to be randomly selected, S_p is the random subset of the item reviews from other users.

The item model is constructed as Equation (5).

$$M_p = BuildModel_p(Q_p, S_p) \tag{5}$$

Where $BuildModel_p(.)$ is the item modeling agent function, M_p is the comprehensive profile of the item p.

The IM modeling is to use a random number generator to pick N_p distinct reviews. And it is both simple and effective, making it suitable for small datasets or scenarios where reviews can be directly manipulated in memory.

3.2.3 Review Style Modeling. To generate human-like correct reviews, UIMM utilizes the review style modeling method to generate the Review Style Model (RSM). The user reviews are analyzed to determine personality traits by extracting specific features such as sentiment, length, vocabulary usage, and grammar patterns. These features are then evaluated against predefined thresholds to classify the user into one of multiple personality categories. Each category represents a distinct personality type, such as optimistic, critical, neutral, or unpredictable. The method is entirely IF-THEN rule-based: IF-parts are to measure the reviews from the users and compare the measure results with the thresholds of the corresponding personality category, and THEN-parts are to model the user review styles with the corresponding style function, as illustrated in Equation (6).

$$M_r = \begin{cases} IF\, f_1(x) > \tau_1, THEN\, P_1(x); \\ IF\, f_1(x) > \tau_2, THEN\, P_2(x); \\ \ldots \\ IF\, f_1(x) > \tau_K, THEN\, P_K(x); \end{cases} \tag{6}$$

Where K is the resulting personality category for the given review data x, $f_k(.)$ is the k-th feature metric derived from x, $k = 1 \ldots K$, τ_k is the k-th threshold for feature, $P_k(.)$ is the review style modeling agent function for k-th predefined personality categories, M_r is the user u review style model, consisting of the user stylistic patterns, sentiment trends, tone and selected past reviews.

3.3 Reasoning Module (RM)

In the Reasoning Module (RM), the reasoning agent predicts both the stars and the reviews for a given user and item, using a large language model guided by the structured prompts. The agent integrates information from the user model M_u, item model M_p, and review style model M_u, allowing it to produce outputs that closely reflect real-world user behavior as illustrated in Equation (7):

$$\{s_{u,p}, r_{u,p}\} = \phi(prompt, M_u, M_p, M_r) \tag{7}$$

Where $r_{u,p}$ represents the predicted user review, and $s_(u,p)$ predicted the number of stars, that user u would assign to item p, $\phi(.)$ is the reasoning agent function to guide the language model's predictions by the prompts. The inputs M_u, M_p, M_r provide the necessary context: M_u captures the user's historical preferences and reviewing tendencies, M_p escribes the item's aggregated features and feedback; and M_r contains the user stylistic patterns, sentiment trends, and tone. Combining these components, the agent generates realistic and personalized predictions.

The workflow of the reasoning agent is: **First**, dynamically adjust the default prompts through RSM; **Second**: put UM and IM into the context memory; **Third**, reasoning and predicting the results; **Forth**, optime the reviews according to UM and RSW; **Finally**, output stars and reviews.

4 EXPERIMENTS

4.1 Preliminaries

In our experiments, we used Qwen2.5-72B-instruct as the large language model for evaluating the submitted agents. The model was set to a temperature of 0. To measure the quality of the generated text, we employed a predefined embedding model and a predefined emotion classifier model. The baseline for comparison was the basic agent outputs provided by the competition organizers.

Table 1: The Performance Comparison Between USHB and Baseline Methods

Method	SRA	RGM	OQ
Baseline	80.14%	80.21%	80.17%
Our USHB	**88.55%**	**90.13%**	**89.34%**

4.2 Evaluation Metrics

We used multiple evaluation metrics to comprehensively measure the quality of the simulated reviews. These metrics included star rating accuracy, review generation quality, and overall performance.

Star Rating Accuracy (SRA). This metric quantifies the degree of alignment between the generated star ratings and the corresponding ground truth ratings, serving as an indicator of predictive accuracy and model performance [15], as illustrated in Equation (8).

$$SRA = 1/N \sum_{i=1}^{N} |\hat{s}_{ni} - s_{ni}| \tag{8}$$

where N is the total number of reviews, and $\hat{s}_{ni}$ and s_{ni} are the normalized predicted and ground truth star ratings, respectively, SRA is Star Rating Accuracy.

Review Generation Metric (RGM). The review generation metric evaluates how closely the generated reviews match the characteristics of real human reviews, as illustrated in Equation (9)

$$RGM = 1 - (0.25 * ETE + 0.25 * SAE + 0.5 * TRE) \tag{9}$$

Where ETE is the metric of Emotional Tone Error, SAE is the metric of Sentiment Attitude Error, TRE is the metric of Topic Relevance Error [1, 2].

Overall Quality (OQ). This metric is the mixed indicator of Star Rating Accuracy and Review Generation Metric, as defined in Equation (10).

$$OQ = 0.5 * SRA + 0.5 * RGM \tag{10}$$

Where OQ is Overall Quality.

4.3 Performance Experiments

To comprehensively evaluate the performance of our USHB model, we designed and executed two distinct experimental analyses: a comparative performance assessment and an investigation into the impact of varying LLM temperature parameters.

In the performance evaluation, we employ a single-agent as the baseline method. As illustrated in Table 1, our USHB demonstrates superior performance compared to the baseline, with improvements of 8.41 points in SRA, 9.92 points in RGM, and 9.17 points in OQ.

In the LLM temperature impact experiment, the mixed dataset (MIX) combined 40% simulated samples (SIM) and 60% real samples (REAL) and we conducted the evaluations of USHB under the LLM temperatures (T) $\{0, 0.3, 0.5, 0.7, 1.0\}$. when the temperature reaches 1.0, the processing time for reviews exceeds 120 minutes, which is identified as a timeout by the completion evaluation system. As evidenced by the data presented in Table 2, at a temperature of 0, the SRA and RGM of USHB reach their peak. As the temperature varies, the difference between the simulated OQ and the realOQ remains below 9.85. The smallest difference in OQ, which is 9.03, occurs

Table 2: The Performance of USHB based On Mixed Dataset and Different LLM Temperatures

T	SRA	RGM	OQ(SIM)	OQ(REAL)	OQ(MIX)
0	**85.32%**	**82.71%**	89.43%	**80.40%**	**84.01%**
0.3	85.05%	82.61%	89.74%	79.89%	83.83%
0.5	85.07%	82.31%	89.40%	79.88%	83.69%
0.7	84.43%	82.60%	**90.26%**	79.85%	**84.01%**

Table 3: The Overall Quality Statistics of USHB with Different LLM Temperatures

Metric	Mean	Var	Std
OQ (SIM)	89.70	0.12	0.35
OQ (REAL)	80.00	0.05	0.23
OQ (MIX)	83.89	0.02	0.13

at a temperature of 0, indicating that our USHB has the strongest generalization capability. From Table 3, the standard deviations (Std) of the simulated OQ, real OQ, and mixed OQ across different temperatures are all less than 0.35, demonstrating the robustness of our USHB under varying temperatures. In summary, our USHB method exhibits excellent generalization capability and robustness across different LLM temperatures.

5 CONCLUSION

We propose USHB, a framework designed to better understand and replicate human behavior in agent-based environments. It combines knowledge extraction, adaptive modeling, and advanced reasoning to address limitations like rigid context handling and limited personalization. Experiments show USHB outperforms single-agent methods in predictive accuracy and realistic review generation, while maintaining consistent performance across various language models. Future plans include expanding USHB for multi-agent collaboration, real-time interactions, and multimodal data integration, enabling more advanced behavioral simulations. These improvements will enhance applications in recommendation systems, virtual assistants, and social behavior modeling.

ACKNOWLEDGMENTS

This work was partially supported by the National key R & D plan under Grant (2022YFE0112400), the NSFC under Grant 62176105. The authors are grateful for resources from the Artificial Intelligence and Pattern Recognition Technology Research and Development Center of Jiangnan University. Prof. Zhaohong Deng is the corresponding author of the paper.

REFERENCES

[1] Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa Anke, and Leonardo Neves. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, Hong Kong, China, 1644–1650.

[2] Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa Anke, and Leonardo Neves. 2020. TweetEval: Unified Benchmark and Comparative Evaluation for Tweet Classification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. Association for Computational Linguistics, 1644–1650.

[3] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research* 3, 1 (2003). https://jmlr.csail.mit.edu/papers/v3/blei03a.html

[4] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 1877–1901.

[5] Yashar Deldjoo. 2024. Understanding Biases in ChatGPT-based Recommender Systems: Provider Fairness, Temporal Stability, and Recency. arXiv:2401.10545 [cs.DL]

[6] Dario Di Palma, Giovanni Maria Biancofiore, Vito Walter Anelli, Fedelucio Narducci, Tommaso Di Noia, and Eugenio Di Sciascio. 2023. Evaluating ChatGPT as a Recommender System: A Rigorous Approach. arXiv:2309.03613 [cs.DL]

[7] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep Learning. *Nature* 521, 7553 (2015). https://doi.org/10.1038/nature14539

[8] 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:2307.15780 [cs.DL]

[9] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc.

[10] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin Riedmiller, Andreas K. Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. 2015. Human-level control through deep reinforcement learning. *Nature* 518, 7540 (2015). https://doi.org/10.1038/nature14236

[11] Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2024. Qwen2.5 Technical Report. arXiv:2412.15115 [cs.DL]

[12] Haocong Rao, Cyril Leung, and Chunyan Miao. 2023. Can ChatGPT Assess Human Personalities? A General Evaluation Framework. In *Findings of the Association for Computational Linguistics: EMNLP 2023*. Association for Computational Linguistics, 1184–1194.

[13] Fei Sun, Jun Liu, Jian Wu, Changhua Pei, Xiao Lin, Wenwu Ou, and Peng Jiang. 2019. BERT4Rec: Sequential Recommendation with Bidirectional Encoder Representations from Transformer. In *CIKM '19: The 28th ACM International Conference on Information and Knowledge Management*. ACM, Beijing China, 1441–1450.

[14] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc.

[15] Yancheng Wang, Ziyan Jiang, Zheng Chen, Fan Yang, Yingxue Zhou, Eunah Cho, Xing Fan, Yanbin Lu, Xiaojiang Huang, and Yingzhen Yang. 2024. RecMind: Large Language Model Powered Agent For Recommendation. In *Findings of the Association for Computational Linguistics: NAACL 2024*. Association for Computational Linguistics, Mexico City, Mexico, 4351–4364.

[16] Fan Yang, Zheng Chen, Ziyan Jiang, Eunah Cho, Xiaojiang Huang, and Yanbin Lu. 2023. PALR: Personalization Aware LLMs for Recommendation. arXiv:2305.07622 [cs.DL]

[17] Bowen Zhang, Daijun Ding, Liwen Jing, Genan Dai, and Nan Yin. 2022. How would Stance Detection Techniques Evolve after the Launch of ChatGPT? arXiv:2212.14548 [cs.DL]

[18] Yin Zhang, Rong Jin, and Zhi-Hua Zhou. 2010. Understanding bag-of-words model: a statistical framework. *International Journal of Machine Learning and Cybernetics* 1, 1 (2010). https://doi.org/10.1007/s13042-010-0001-0

[19] Siyan Zhao, Mingyi Hong, Yang Liu, Devamanyu Hazarika, and Kaixiang Lin. 2025. Do LLMs Recognize Your Preferences? Evaluating Personalized Preference Following in LLMs. In *The Thirteenth International Conference on Learning Representations*.

[20] Zijun, Yanzhe Zhang, Peng Li, Yang Liu, and Diyi Yang. 2023. A Dynamic LLM-Powered Agent Network for Task-Oriented Agent Collaboration. arXiv:2310.02170 [cs.DL]