福建省中电海峡智能装备研究院
Fujian Zhongdian Straits Institute of Intelligent Equipment
Asia Pacific Science Press
Advances in Management and Intelligent Technologies
ISSN(O):3079-5869
ISSN(P):3079-5850
Vol. 1 No.2 (2025)

Mockingbird in Humanity: Data Fondness of LLM in Hosting Virtual Personalities

Kejie Zhang, Jingming Li*
School of Civil Engineering and Architecture, Nanyang Normal University, Nanyang , 473061, China
Corresponding author: Jingming Li, jmli@nynu.edu.cn

Abstract: The intelligent development in building design, construction, and operation & maintenance is exceptionally rapid, which has become a trend that cannot be ignored in the current field of architecture. With the help of prompt engineering, architects can use generative AI to lay out building space designs and even generate 3D drawings. Artificial intelligence agents can act as designers and owners, representing all parties involved in the building life cycle. In this way, they simulate all parties involved in the building life cycle, providing a comprehensive perspective and solutions for the smooth progress of the building. However, this has led to a problem worthy of in-depth exploration: large models have tendencies when playing different roles. In this article, we aim to deeply explore the tendencies of large language models (LLM) when playing virtual personalities. Specifically, we will conduct extensive experiments to examine two important aspects. One aspect is the analytical ability of large models in terms of virtual personalities, which includes how they interpret requirements in different situations and how they conduct logical analysis according to different role positions. The other aspect is the performance of large models in terms of regions and ethnic groups when playing virtual personalities. Different regions have different architectural cultural and style requirements, and different ethnic groups also have unique architectural aesthetics and traditions. Although LLMs have shown a certain discriminative ability during the role-playing process and can distinguish different role requirements, we find that the content they generate still shows a specific content tendency. This research can deepen the understanding of LLM's performance in multiple aspects such as building design and operation & maintenance.

Keywords: LLM; AI Agent; Virtual Personalities; AIGC
Published: May 24, 2025
DOI: https://doi.org/10.62177/amit.v1i2.343

1.Introduction

With the rapid development of generative artificial intelligence, large language models (LLMs) have become the core technology driving the transformation of human-computer interaction paradigms. Represented by ChatGPT, LLMs exhibit human-like dialogue and knowledge generation capabilities through reinforcement learning from human feedback (RLHF) and large-scale data training. They can even simulate specific personality traits for interaction. This technological breakthrough not only restructures information production methods but also fosters the emergence of "human-model" symbiotic systems, providing unprecedented technical foundations for the construction of virtual personalities.

The data preference characteristics displayed by LLMs in carrying virtual personalities are triggering deep reflections in academia on technical ethics, social impacts, and governance paths. The shaping of virtual personalities by LLM essentially

stems from its training data. LLM relies on learning statistical patterns from massive corpora, and the quality and ethical attributes of its generated content are directly influenced by the size, diversity, and value orientation of the training data. Studies have used MBTI (Myers Briggs Type Indicator) in human personality assessment as an indicator for evaluating LLMs [1]. They have conducted experiments on the personality tendencies of the large model itself, and the personality traits exhibited by the model can be adjusted through prompt engineering.

However, there are still many questions about the thinking mechanism of LLMs. The experimental design mainly focused on language comprehension tasks and did not fully explore the similarities of other cognitive functions. Some studies also suggest that it is the procedural knowledge in pre-training that drives the inference of large models. [2]. This article will study the data selection strategy of LLM in hosting virtual personalities from the perspective of data preferences, explore the mechanism of personality traits shaped by different types of data, and provide theoretical support for building controllable and trustworthy virtual personalities.

2.Related Works

With the rapid development of artificial intelligence technology, LLMs have shown revolutionary potential in the field of virtual personality hosting. Current mainstream models such as ChatGPT, Bard, Claude, etc. can simulate human dialogue patterns and carry specific personality traits [3]. Research has shown that the personality traits of virtual characters have a significant impact on user perception. By designing virtual characters with specific personality traits, their credibility and attractiveness can be enhanced [4]. The improvement of interaction ability enables LLMs to have a wide range of applications in various fields. By designing virtual characters with specific personality traits, the credibility and attractiveness of LLM agents can be enhanced. Studies are building multi-LLM agent teams, using prompt words to make each agent play different roles and achieve team collaboration [5–7]particularly large language models (LLMs. By fine-tuning LLMs in specific domains, they improved the predictive accuracy in the field of neuroscience, LLMs can extract and integrate information from a large amount of scientific literature, surpassing human experts in predicting experimental results [8]. For instance, the Otome game proposes a new type of emotional support chatbot by combining LLM. The game demonstrates how to enhance the interactive experience through data augmentation and emotional enhancement technology, significantly improving emotional participation in interactive entertainment [9]. Studies even collect detailed information about real people and have LLM agents imitate their behavior and attitudes, simulating the behavior and attitudes of thousands of people, for use in sociological behavioral research [10].

There are also numerous research applications in the construction industry. [11]. These applications can be divided into four categories. The first one is to improve machine learning-based prediction. The opaque inference mechanism of machine learning has caused certain understanding barriers for users, ultimately leading to decision-making dishonesty. A deep interpretation of the LLM inference process can improve the credibility of operational decisions. The second is to delegate data and control permissions to LLMs, who will analyze and make decisions. For instance, the utilization of a basic model (such as GPT-4) for direct industrial control, significantly reduces the technical burden. The experimental results show that GPT-4 exhibits comparable performance to traditional reinforcement learning methods in HVAC control tasks, demonstrating the potential of the basic model in industrial applications [12]. The third approach is to incorporate LLMs into the interactive interface of building energy management systems to improve communication between the system and users. This approach improves communication between users and operations engineers, enhances user experience, and reduces operational costs [13]. The fourth type relies on the multimodal capability of large models, combined with heterogeneous data such as robot vision and hearing, to achieve embodied intelligence.

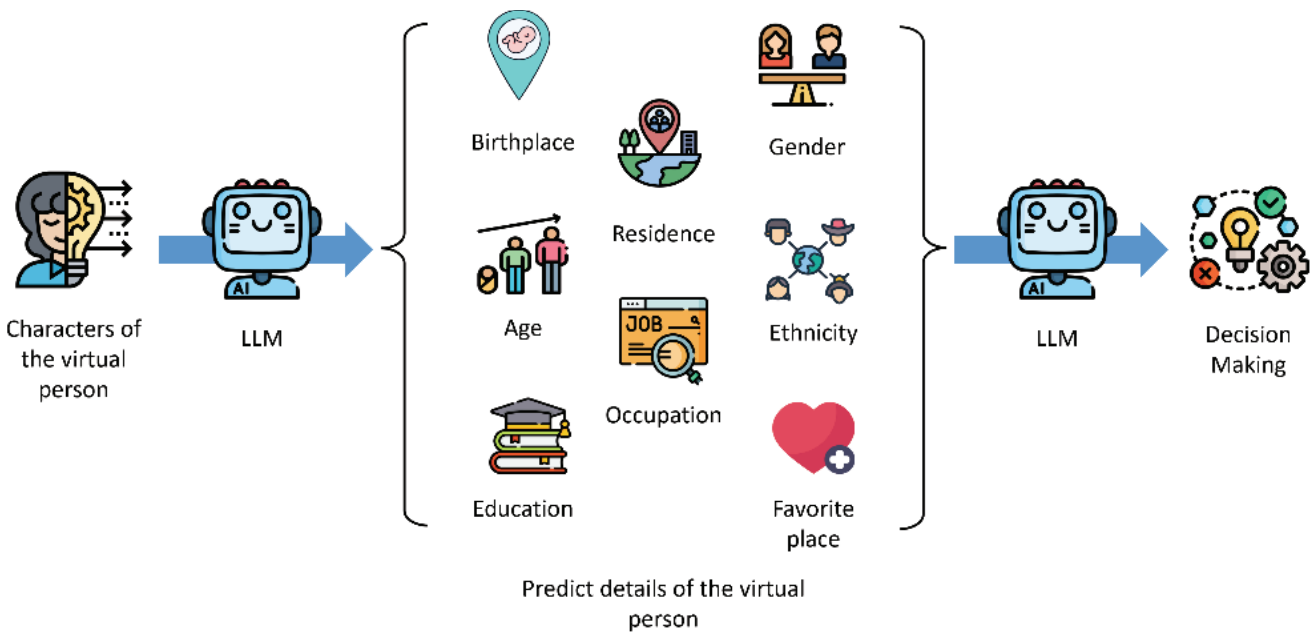
Although LLM has obvious advantages, there is still debate about the inference mechanism of LLM agents. Some studies suggest that there is a significant correlation between the performance of LLMs and their brain similarity, and their feature extraction mechanism will be closer to the language processing mechanism of the human brain [14]. However, the model parameters used in the experiment were relatively small and lacked in-depth analysis of different types of language tasks. This study will conduct experiments on the abilities of LLM agents, testing their reasoning abilities regarding age, gender, occupation, place of birth, ethnicity group, and education based on a virtual personality database, and further testing their

deep reasoning abilities regarding their place of residence and favorite place.

3.Methodology

This article will use LLM agents as analysts. The experiment first shuffles the virtual personality database and then sends it to the analyst played by the LLM agent to induce inference through prompt words. The inference results are then handed over to another analyst played by the LLM agent for judgment. The basic structure of the experiment is shown in Fig. 1. The virtual personality database is from [15]. Fig. 1 uses icons form [16].

Figure 1. Experiment structure



The experiment is based on a virtual persona where analysts played by LLM are asked about their inferences about virtual humans. Specifically, the LLM agent's status is first updated through prompt words to bypass system restrictions. Then raise questions to LLM and limit the return to JSON format for data processing. This stage mainly uses models with parameter sizes below 32 billion, which are Mistral-nemo:12b, Mistral-small:22b [17]a 7-billion-parameter language model engineered for superior performance and efficiency. Mistral 7B outperforms Llama 2 13B across all evaluated benchmarks, and Llama 1 34B in reasoning, mathematics, and code generation. Our model leverages grouped-query attention (GQA, Qwen2.5:32b [18] we introduce Qwen2.5, a comprehensive series of large language models (LLMs, Gemma2:27b [19]. Then modify the system prompt using the same method, allowing the LLM agent to evaluate the inference from the previous stage and return JSON data. The model used for this process is Qwen2.5:72b. Fig.2 illustrates the conversation structure. LLM service runs on a 4070 Ti Super with 128G RAM, and due to limited computing power, the experiment only infers 1000 virtual humans.

The questions include the age, gender, ethnicity, occupation, education level, place of birth, place of residence, and favorite place of the virtual person. Considering that the description of the virtual persona may include information on age, gender, occupation, education, and place of residence, the experiment also includes ethnicity, birthplace, and favorite place for reasoning. The global city list comes from [20], and the global ethnic group list is from [21].

The experiment also tested models such as Gemma2:9b, Llama3.1:8b, Qwen2.5:7b, Qwen2.5:14b, qwq:32b, DeepSeek-R1:32b, etc., but the output performance was not satisfactory. After the above eight categories of Q&A, the experiment obtained the inference and judgment of inference correctness for the four small-scale models mentioned above. Section 4 will present these results.

4.Results

As shown in Fig. 3, we have updated the task of LLM through prompt word engineering, and the model responded to three stages of thinking. Taking age inference as an example, the model first considers universal features and infers based on age

groups; In the second stage, the use of 'new' in the prompt word caused some interference with the model's inference, but the model was ultimately revised; The third stage model has balanced the relevant content of the virtual persona.

Figure 2. Conversation structure with LLM API

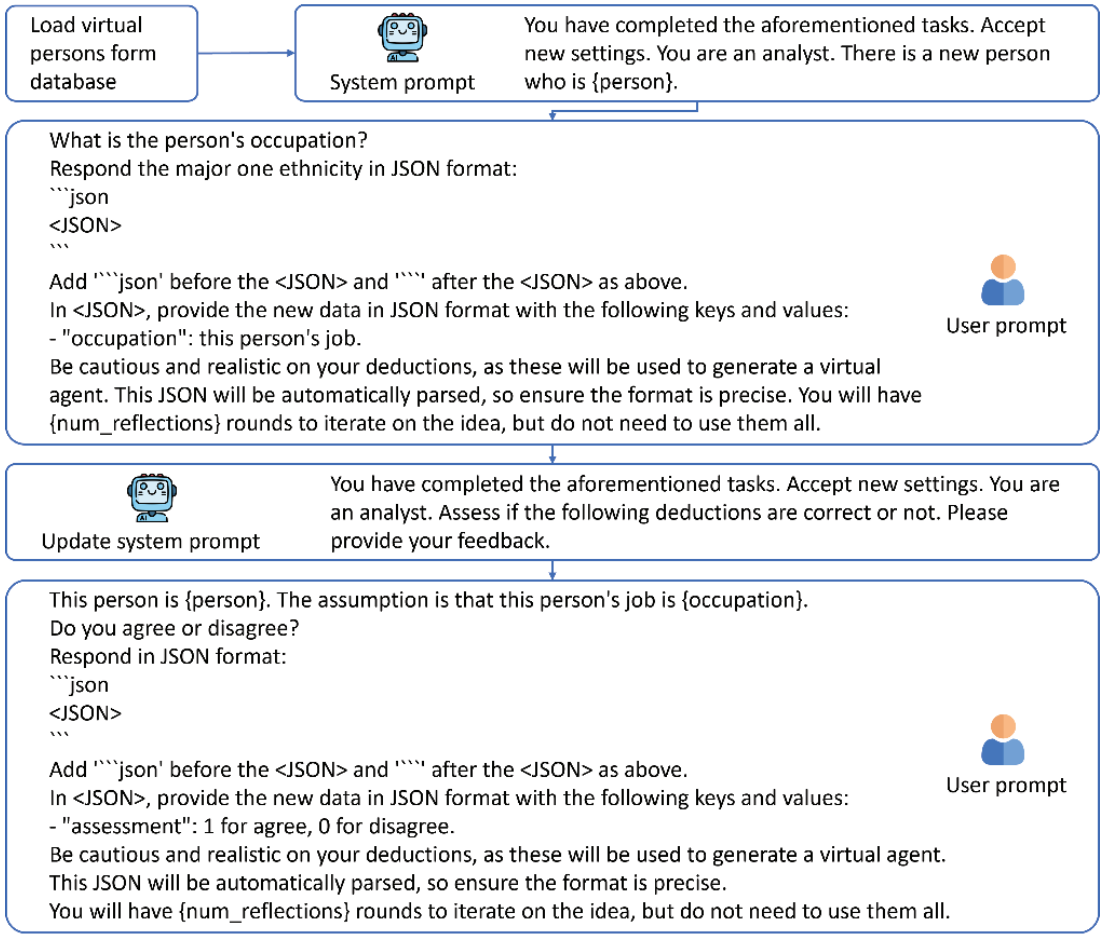
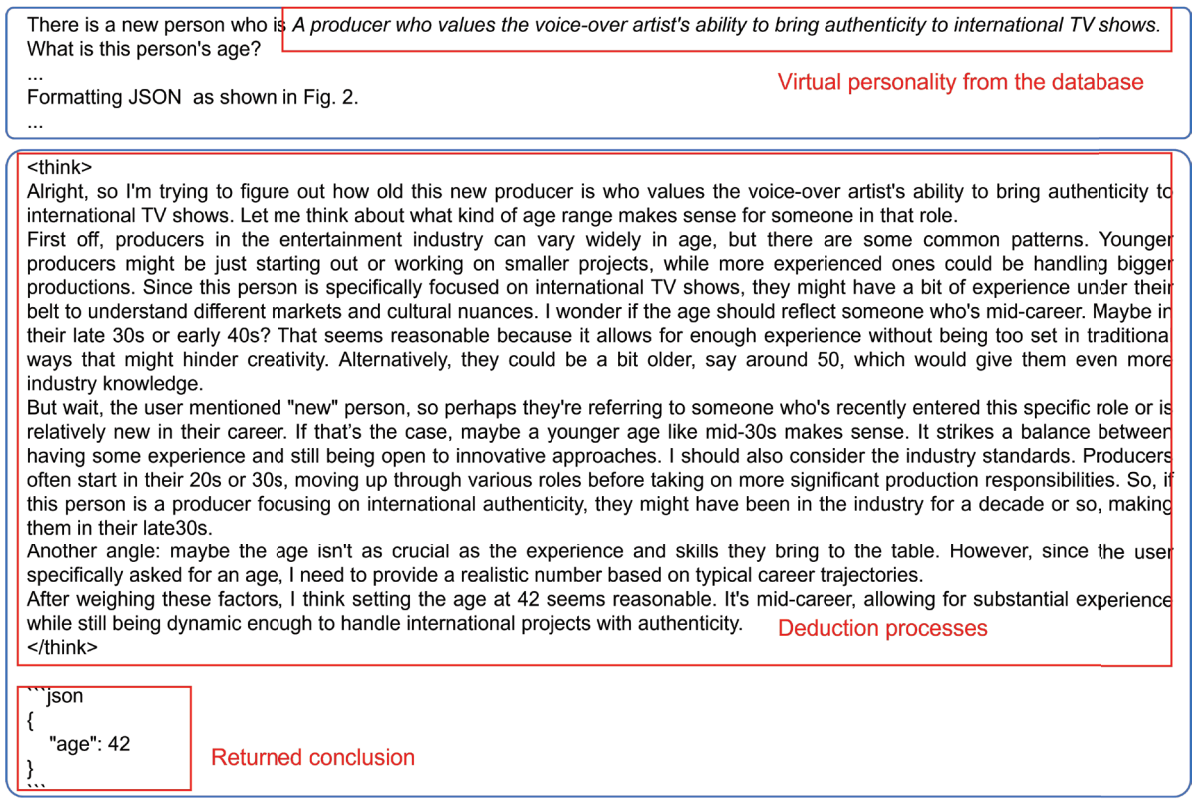
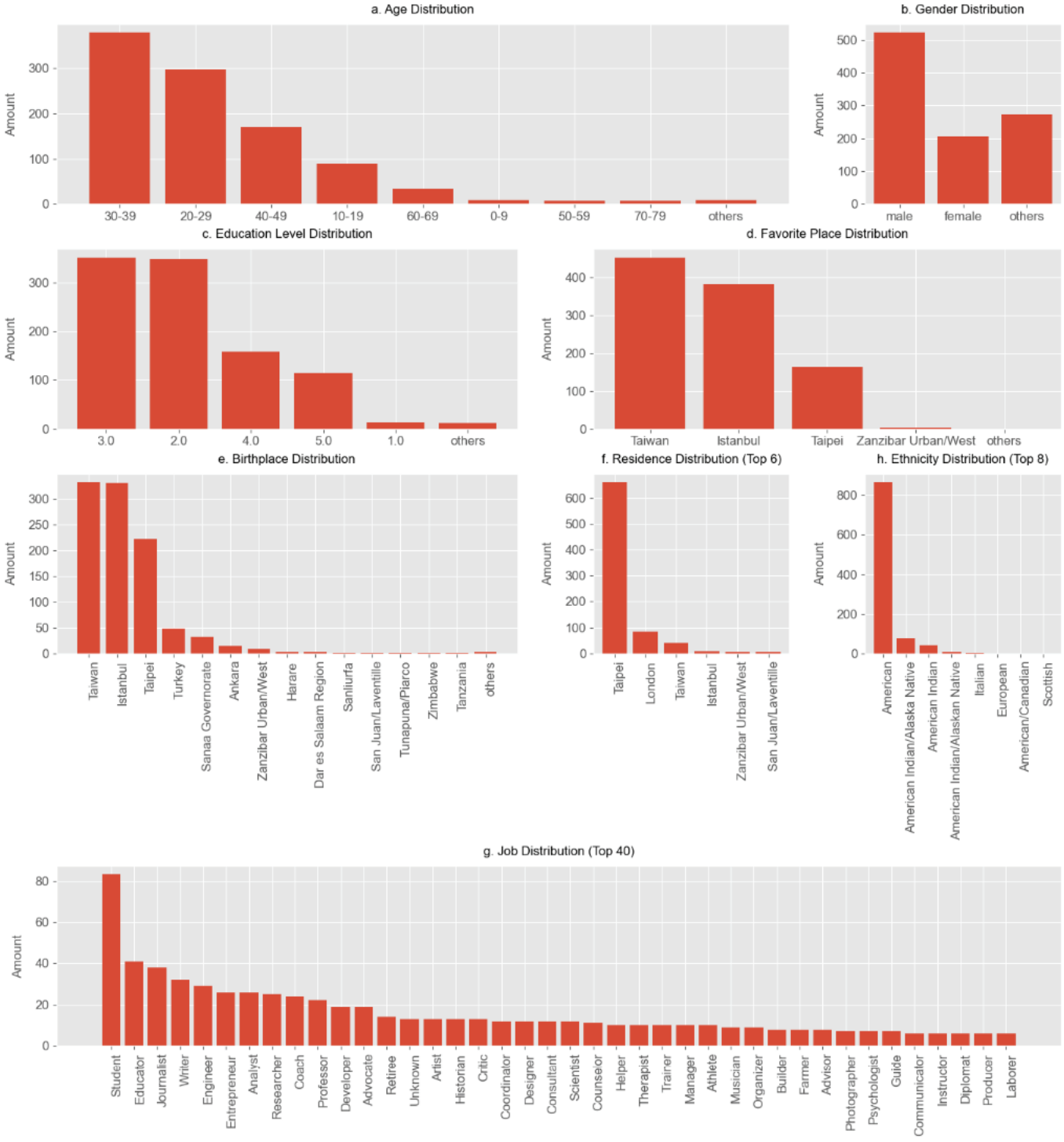


Figure 3. Processes of a small-scale model reasoning



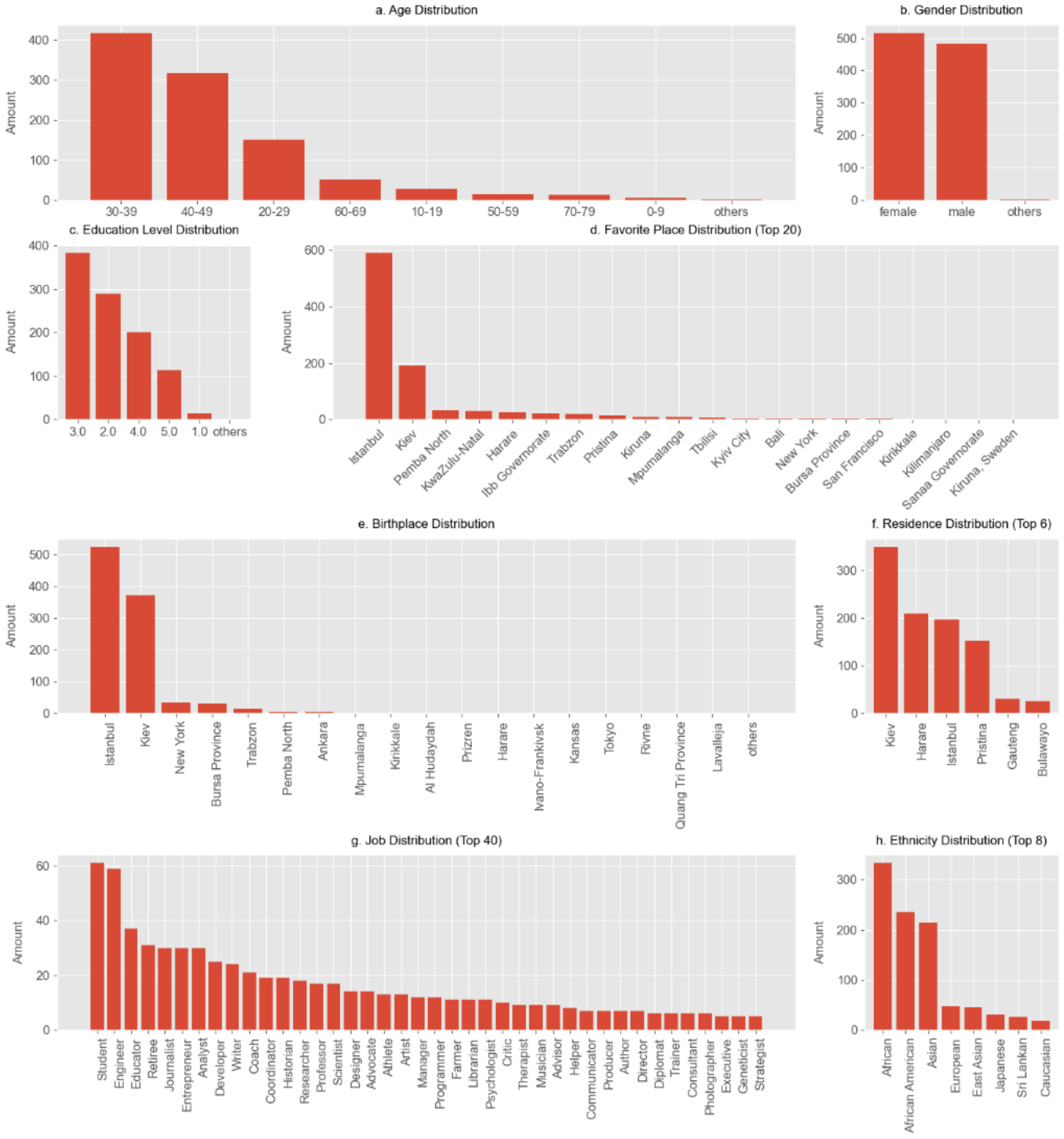
We repeated this reasoning process in 1000 virtual personalities, involving age, gender, occupation, education, place of residence, ethnicity group, birthplace, and favorite place. The results are illustrated in Fig. 4 , 5, 6, and 7. Comparing the results of the four models, the inferential distribution of age and education level is relatively similar, the distribution of age is more in the range of 30-39, and the education level is concentrated in bachelor's degree or equivalent. Gemma2:27B and Qwen2.5:32B show obvious diversity in gender inference. Gemma2:27B in place reasoning is concentrated in Taiwan, Mistral-nemo:12B is concentrated in Istanbul, Harare appears more in Mistral-small:22B's reasoning on place, Qwen2.5:32B is concentrated in Virginia. In terms of ethnicity groups, the results of Gemma2:27B and Mistral-small:22B concentrated in American, Mistral-nemo:12B concentrated in African, and Qwen2.5:32B collectively referred to as white.

Figure 4. Results of Gemma2:27B



The assessments of the assumptions by Qwen2.5:72b are illustrated in Fig. 8 and 9. The bar chart on the left displays the specific numerical distribution of each model in different dimensions, while the bar chart on the right summarizes the overall performance by Qwen2.5:72b of each model in four dimensions. Red means Qwen2.5:72b disagrees with the assumptions. The inference results of the four small-scale models have a low agreement with Qwen2.5:72b.

Figure 5. Results of Mistral-nemo:12B



The distribution of judgment results of four different models (Qwen2.5:32B, Mistral-small:22B, Mistral-nemo:12B, Gemma2:27B) in four different dimensions (age, gender, occupation, education) was presented in Fig. 8. Qwen2.5:32B shows higher accuracy in judgment across four dimensions: age, gender, occupation, and education, particularly in predicting education level. Gemma2:27B demonstrates high accuracy in all four dimensions, particularly in terms of gender and

education level. Mistral-small:22B has relatively low accuracy in these four dimensions, especially in the vocational and educational dimensions. The accuracy of the judgment of Mistral-nemo:12B is similar to that of Mistral-small:22B, but its performance is relatively weak.

Figure 6. Results of Mistral-small:22B

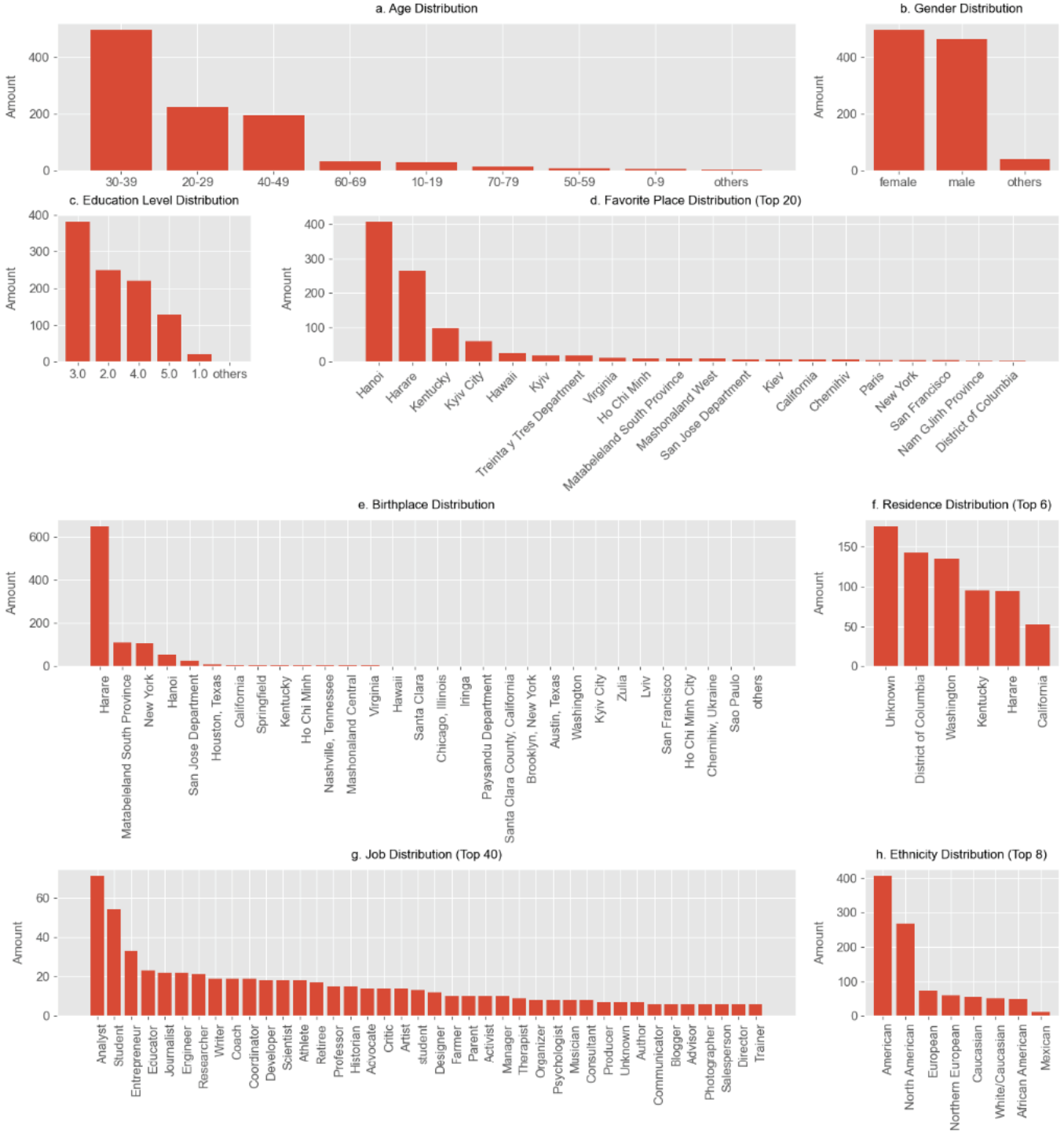
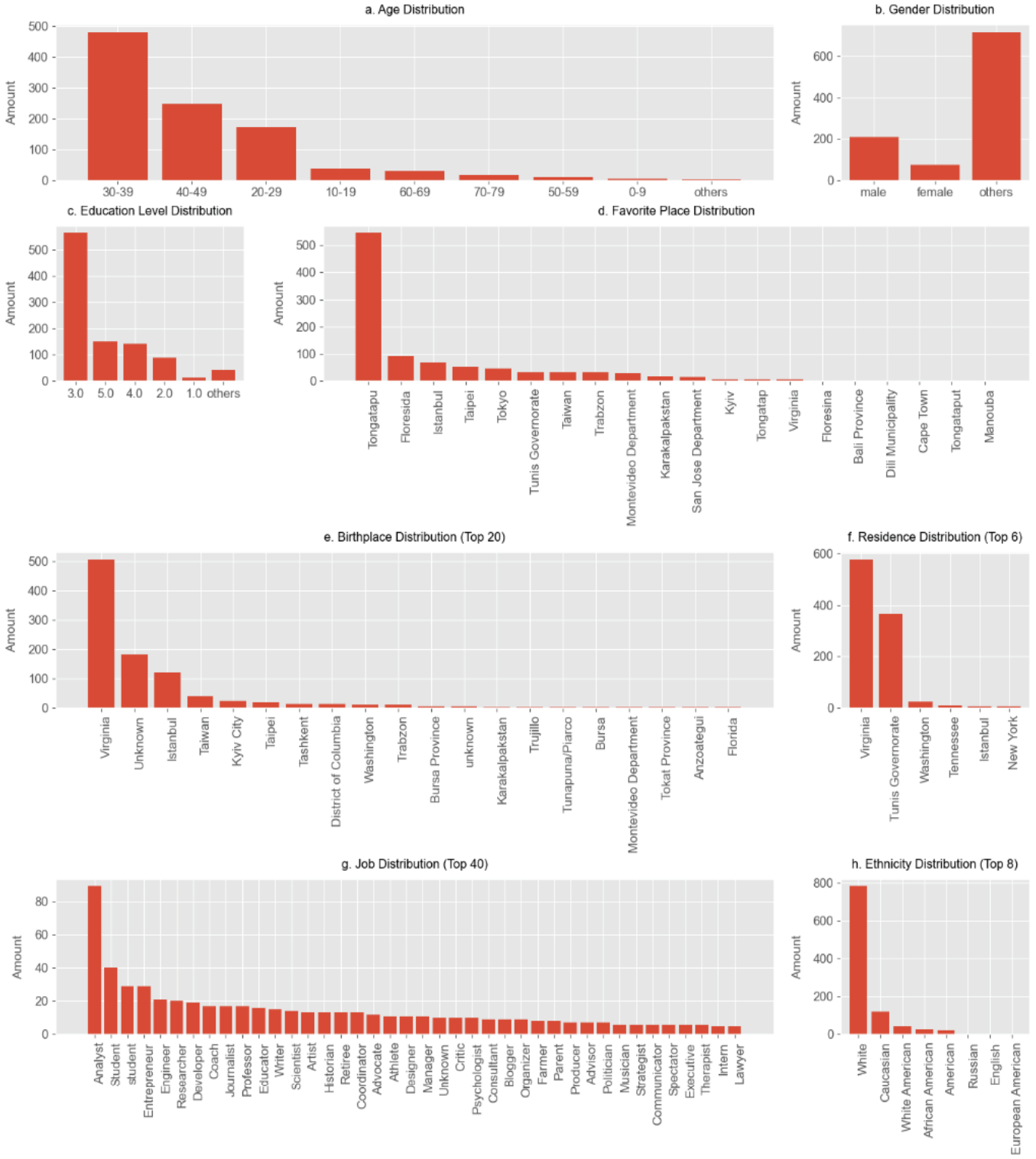


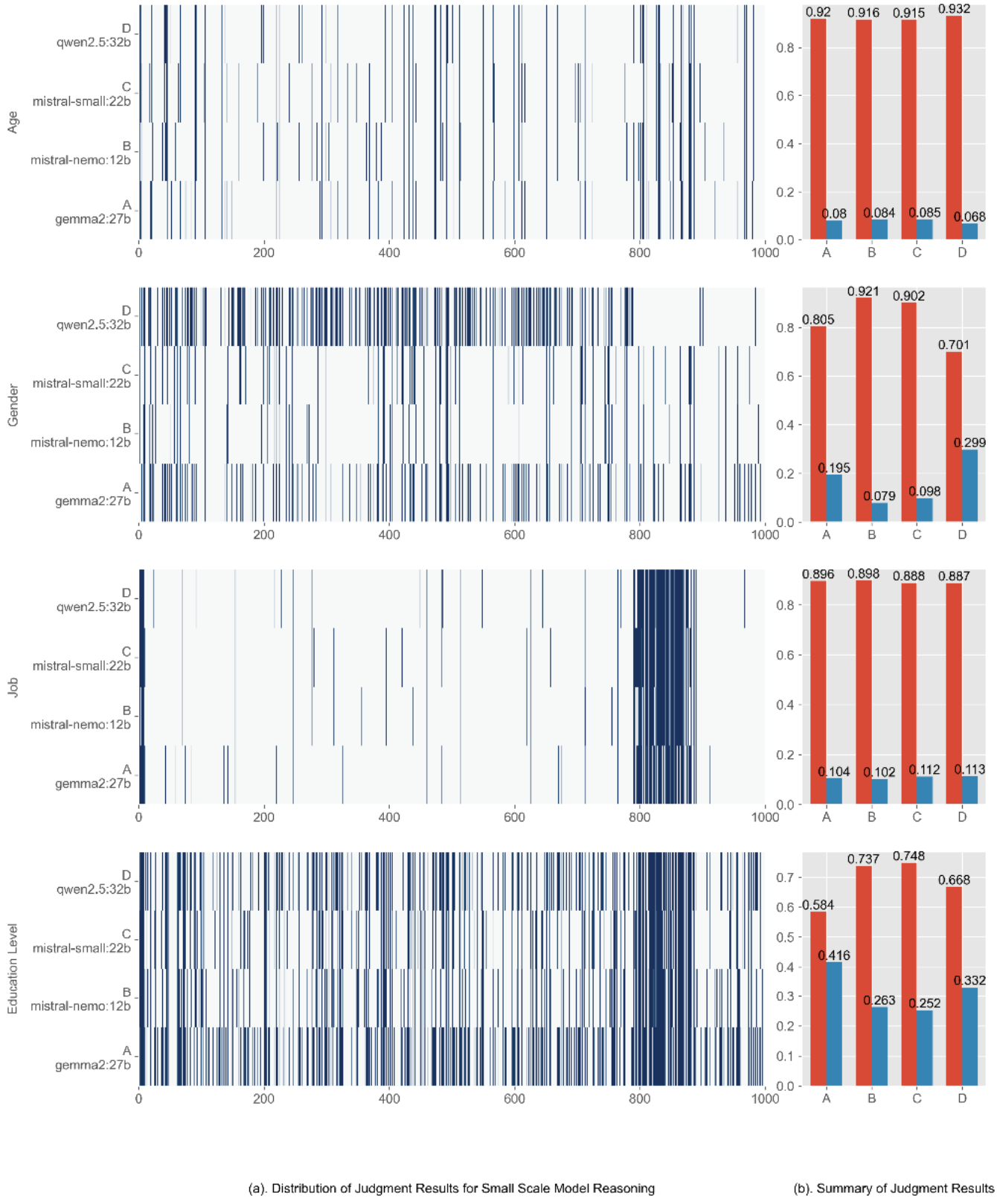
Fig.9 shows the distribution of judgment results of four major language models (Qwen2.5:32B, Mistral-small:22B, Mistral-nemo:12B, Gemma2:27B) in four different dimensions (Birthplace, Favorite Place, Residence, Ethnicity Group). From the graph, there are differences in the performance of the four models in different dimensions, but overall, they perform relatively consistently in certain dimensions. The performance of the four models in the birthplace dimension is consistent, and the numerical distribution is concentrated. The model also exhibits similar distribution characteristics in the favorite place dimension.

Figure 7. Results of Qwen2.5:32B



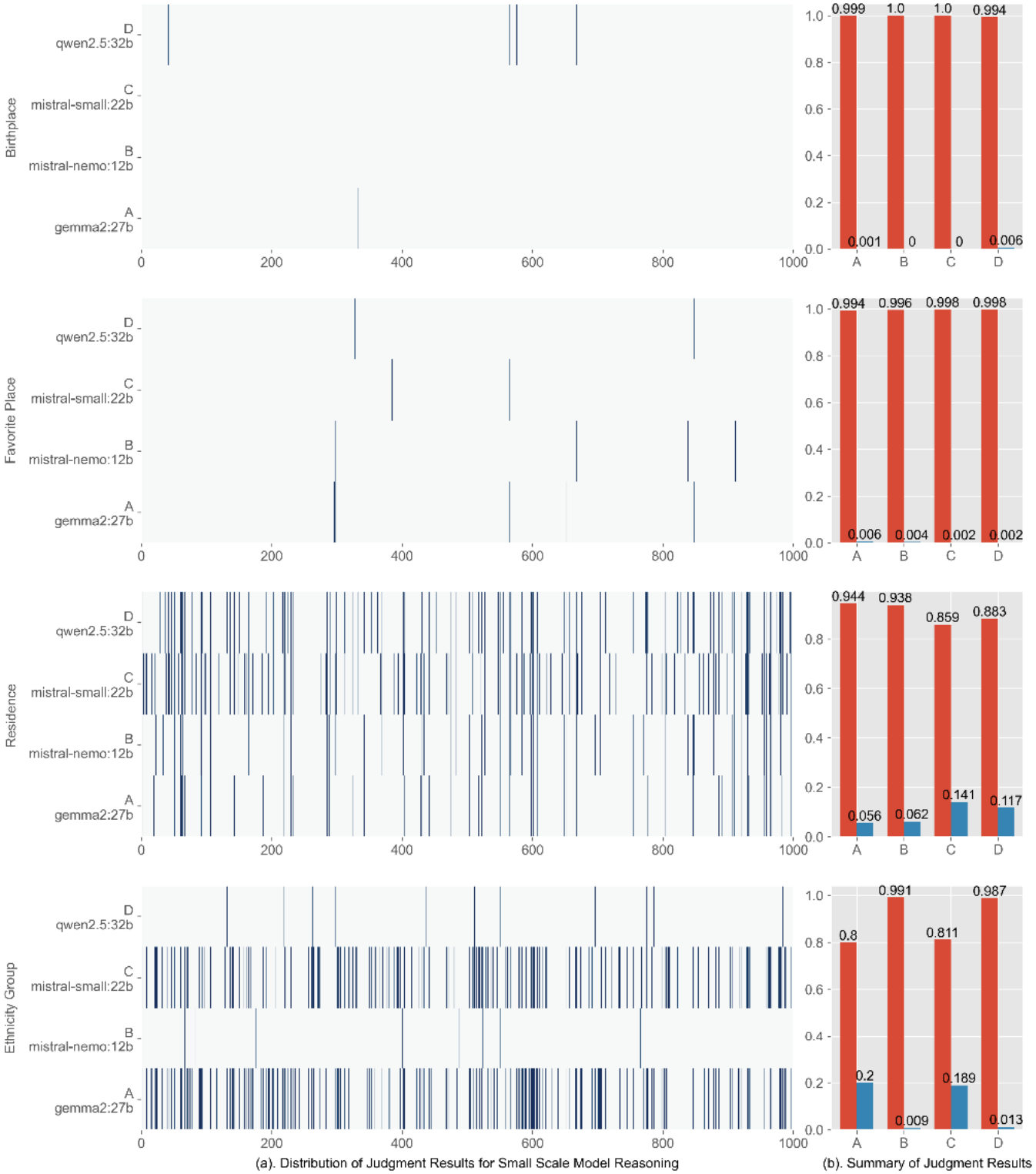
In terms of residential dimension, the numerical distribution of the model is relatively scattered, but the overall performance is still relatively consistent. Qwen2.5:32b and Mistral-small:22b have relatively better results. On the dimension of ethnic groups, the numerical distribution of the model is relatively uniform, but Gemma2:27B and Mistral-small:22b perform better. Generally, Qwen2.5:32B and Gemma2:27B perform well in multiple dimensions and are suitable for tasks that require complex reasoning and long text generation. Mistral-small:22B and Mistral-nemo:12B perform relatively weakly and are suitable for resource-constrained scenarios. The experiment also utilized small-scale models including DeepSeek-R1 and Llama3.3, however, due to limitations in computing power, the returned results cannot maintain a consistent format.

Figure 8. Distribution and summary of judgments for small-scale model reasoning (I)



(a). Distribution of Judgment Results for Small Scale Model Reasoning (b). Summary of Judgment Results

Figure 9. Distribution and summary of judgments for small-scale model reasoning (II)



5.Discussions and limitations

According to the experimental results, LLM agents have demonstrated certain abilities and potential in playing the role of analysts. LLM agents can combine intuitive information for inference, and the higher the degree of information correlation, the more accurate the inference results. For models below 32 billion, the larger the model size, the higher the accuracy of the inference results.

LLM agents have certain potential in simulating building users by simulating different virtual personalities. This ability stems from its ability to learn and analyze large amounts of data, enabling it to understand and generate conversations and behaviors that fit specific roles. For example, in the experiment mentioned in this article, LLM agents can reason based on the

age, gender, occupation, and other characteristics of virtual personalities, thereby providing feedback on building design or use as building users. This simulation will improve the collection of opinions in the design process and enhance the feedback adjustment process of building operation and maintenance.

In the process of architectural design, LLM agents can provide architects with a more comprehensive perspective and solutions by analyzing the architectural cultures and styles of different regions and ethnic groups. The experimental results show that LLM exhibits high accuracy in inferring users' education level, occupation, and place of residence, especially in predicting their education level. LLM can better understand the user's background, acting as both an architect to design and a user to provide feedback, thus better meeting the user's needs in the design. In the process of building operation and maintenance, LLM can optimize the operating parameters of building equipment and achieve predictive feedback adjustment while understanding the user background. The interactive ability of LLM agents can significantly enhance the user experience in the architectural design process.

Through interaction with virtual personalities, users can express their needs and preferences more intuitively, while LLM can provide real-time feedback and adjust design plans. This dynamic interaction not only improves the efficiency of the design but also allows users to feel a higher sense of participation and satisfaction throughout the entire process. Although LLM performs well in role-playing and user simulation, experiments have also revealed specific tendencies in its generated content. This tendency may stem from the quality and diversity of its training data. The architectural aesthetics and traditions vary among different regions and ethnic groups, and LLM may be limited by its training data when simulating these diversities.

This study also has some limitations. Firstly, the model parameters used in the study are relatively small, which may limit the accuracy and comprehensiveness of inference. In the future, larger-scale models should be considered for validation. The experimental design mainly focuses on language comprehension tasks, lacking in-depth exploration of other cognitive functions, which may affect the comprehensive understanding of LLM thinking mechanisms. In addition, the paper did not fully discuss the ethical and social impacts of LLM-generated content, and future research should strengthen the exploration of such issues to ensure responsible use of the technology.

Therefore, in practical applications, architects need to carefully evaluate the content generated by LLM to ensure that it complies with local culture and design standards. Future research can further explore how to optimize the training data of LLM to enhance its adaptability and accuracy in the field of architectural design. In addition, combining more user feedback and practical cases can provide richer materials for LLM's role-playing ability, thereby enhancing its performance in simulating building user experience.

6.Conclusions

This article explores the abilities and tendencies of LLM in portraying virtual personalities through experiments and finds that LLM has the potential to simulate building users and provide feedback to improve the design process. Through experimental design, LLM is used as an analyst to infer virtual personality databases. The research involves the comparison and evaluation of multiple models. The experimental questions cover the characteristics of virtual individuals such as age, gender, ethnicity, occupation, education level, place of birth, place of residence, and preferred location. Evaluate the performance of different models in various dimensions by reasoning and judging 1000 virtual individuals. The experimental results show that LLM exhibits certain abilities in inferring virtual personalities, especially in terms of educational level and professional reasoning. There are differences in inference accuracy and consistency among different models, with Qwen2.5:32B and Gemma2:27B performing relatively well.

LLM agents have demonstrated strong potential in simulating building user experience, providing valuable support for architectural design through their role-playing and data analysis capabilities. However, it is still necessary to pay attention to the tendency of its content generation and combine it with the specific needs of users in practical applications to achieve higher-quality design results.

Funding

This work is supported by Nanyang Normal University Open Laboratory Project (SYKF2024070).

Conflict of Interests

The authors declare that there is no conflict of interest regarding the publication of this paper.

Reference

[1] Pan, K., & Zeng, Y. (2023). Do LLMs possess a personality? Making the MBTI test an amazing evaluation for large language models. https://doi.org/10.48550/arXiv.2307.16180

[2] Ruis, L., Mozes, M., Bae, J., Kamalakara, S. R., Talupuru, D., Locatelli, A., et al. (2024). Procedural knowledge in pretraining drives reasoning in large language models. https://doi.org/10.48550/arXiv.2411.12580

[3] Liu, Y., Chen, J., Bi, T., Grundy, J., Wang, Y., Yu, J., et al. (2024). An empirical study on low code programming using traditional vs large language model support. https://doi.org/10.48550/arXiv.2402.01156

[4] Zhou, M. X., Mark, G., Li, J., & Yang, H. (2019). Trusting virtual agents: The effect of personality. ACM Transactions on Interactive Intelligent Systems, 9, Article 10, 1–36. https://doi.org/10.1145/3232077

[5] Su, H., Chen, R., Tang, S., Zheng, X., Li, J., Yin, Z., et al. (2024). Two heads are better than one: A multi-agent system has the potential to improve scientific idea generation. https://doi.org/10.48550/arXiv.2410.09403

[6] Ghafarollahi, A., & Buehler, M. J. (2024). SciAgents: Automating scientific discovery through multi-agent intelligent graph reasoning.

[7] Lu, C., Lu, C., Lange, R. T., Foerster, J., Clune, J., & Ha, D. (2024). The AI scientist: Towards fully automated open-ended scientific discovery. https://doi.org/10.48550/arXiv.2408.06292

[8] Luo, X., Rechardt, A., Sun, G., Nejad, K. K., Yáñez, F., Yilmaz, B., et al. (2024). Large language models surpass human experts in predicting neuroscience results. Nature Human Behaviour, 1–11. https://doi.org/10.1038/s41562-024-02046-9

[9] Pan, Y., Tang, Y., & Niu, Y. (2023). An empathetic user-centric chatbot for emotional support. https://doi.org/10.48550/arXiv.2311.09271

[10] Park, J. S., Zou, C. Q., Shaw, A., Hill, B. M., Cai, C., Morris, M. R., et al. (2024). Generative agent simulations of 1,000 people. https://doi.org/10.48550/arXiv.2411.10109

[11] Yan, C., & Yuan, P. F. (2024). Phygital intelligence. ARIN, 3, Article 30. https://doi.org/10.1007/s44223-024-00073-0

[12] Song, L., Zhang, C., Zhao, L., & Bian, J. (2023). Pre-trained large language models for industrial control. https://doi.org/10.48550/arXiv.2308.03028

[13] Zhang, L., & Chen, Z. (2023). Opportunities and challenges of applying large language models in building energy efficiency and decarbonization studies: An exploratory overview. https://doi.org/10.48550/arXiv.2312.11701

[14] Mischler, G., Li, Y. A., Bickel, S., Mehta, A. D., & Mesgarani, N. (2024). Contextual feature extraction hierarchies converge in large language models and the brain. Nature Machine Intelligence, 6, 1467–1477. https://doi.org/10.1038/s42256-024-00925-4

[15] Ge, T., Chan, X., Wang, X., Yu, D., Mi, H., & Yu, D. (2024). Scaling synthetic data creation with 1,000,000,000 personas. https://doi.org/10.48550/arXiv.2406.20094

[16] Pan, J. S., Zeng, Y., & others. (2024). Icons. Flaticon. https://www.flaticon.com/free-icons (Accessed February 25, 2025)

[17] Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., Casas, D. de las, et al. (2023). Mistral 7B. https://doi.org/10.48550/arXiv.2310.06825

[18] Qwen, Yang, A., Yang, B., Zhang, B., Hui, B., Zheng, B., et al. (2025). Qwen2.5 Technical Report. https://doi.org/10.48550/arXiv.2412.15115

[19] Team G., Riviere, M., Pathak, S., Sessa, P. G., Hardin, C., Bhupatiraju, S., et al. (2024). Gemma 2: Improving open language models at a practical size. https://doi.org/10.48550/arXiv.2408.00118

[20] Lexman, Open Knowledge Foundation, & GeoNames. (2024). Major cities of the world. https://datahub.io/core/world-cities (Accessed November 14, 2024)

[21] Wikipedia. (2001). List of contemporary ethnic groups. Wikipedia.