*Table 1.* LLMs' Self-knowledge - Action Congruence Performance with Reference of Human Respondents' Performance (AVG, SD, MIN and MAX represents the average number, standard deviation, minimum and maximum.)

| LLMs & Human Respondents | Cosine Similarity | Spearman Rank Correlation Coefficient | Value Mean Difference | Proportion of Consistent Pairs |
|---|---|---|---|---|
| ChatGLM3 | 0.24 | 0.23 | 1.58 | 47.22% |
| GPT-3.5-turbo | 0.17 | 0.19 | 1.74 | 50.56% |
| GPT-4 | 0.52 | 0.56 | 1.02 | 78.89% |
| Vicuna-13b | 0.08 | 0.07 | 1.57 | 52.78% |
| Vicuna-33b | 0.18 | 0.06 | 1.68 | 52.22% |
| LLMs(AVG $\pm$ SD) | **0.24 $\pm$ 0.15** | **0.22 $\pm$ 0.18** | **1.52 $\pm$ 0.26** | **56.78 $\pm$ 11.25%** |
| Human(AVG $\pm$ SD) | **0.76 $\pm$ 0.09** | **0.78 $\pm$ 0.08** | **0.69 $\pm$ 0.27** | **84.69 $\pm$ 8.22%** |
| Human(MIN) | 0.61 | 0.66 | 1.08 | 73.78% |
| Human(MAX) | 0.95 | 0.96 | 0.07 | 99.44% |

to each other one-to-one.

**Spearman's Rank Correlation Coefficient** A non-parametric measure of rank correlation, assessing how well the relationship between two variables can be described using a monotonic function. Its value ranges from -1 to 1, where 1 means a perfect association of ranks. Specifically, we rank the responses on two questionnaires of the LLMs based on their numerical values separately. Then, we calculate the difference in rankings for each personality knowledge – scenario & action pair. Afterwards, we use the following formula to calculate the coefficient $r_s$.

$$r_s = 1 - \frac{6\sum d_i^2}{n(n^2-1)}, \tag{2}$$

where $d_i$ is the difference in rankings of each pair and $n$ is the total count of pairs.

**Value Mean Difference (VMD)** Value Mean Difference is the average difference in responses across all paired items in the questionnaires, as shown in the formula below.

$$\mathrm{VMD} = \frac{\sum \mathrm{d_i}}{\mathrm{n}}, \tag{3}$$

where $d_i$ is the difference of responses in each pair.

**Proportion of Consistent Pairs** Recognizing that minor discrepancies are natural when comparing psychological tendencies with actual actions, this metric quantifies the proportion of item pairs with a response difference of 1 or less, focusing on the consistency of tendencies rather than exact matches.

$$P_c = \frac{N_c}{N_t}, \tag{4}$$

where $N_c$ is the number of consistent pairs, $N_t$ is the total number of pairs.

For this study, we recruited 16 participants, comprising 8 males and 8 females, all native Chinese speakers with an English proficiency level of CEFR C1. As shown in Table 1, the analysis of their response data yielded an average Cosine Similarity and Spearman's Rank Correlation Coefficient above 0.75, with a Value Mean Difference around 0.68, and a Proportion of Consistent Pairs exceeding $84\%$. These results indicate a high degree of similarity and strong correlation between responses to the two types of questionnaires, suggesting a basic consistency in human self-knowledge and an ability to align self-knowledge with action in real-life scenarios.

The same questionnaires were administered to the 5 LLMs selected in Section B, and their responses were analyzed using the aforementioned metrics. Compared to human respondents, the similarity in LLMs' responses is notably lower, and the corresponding significance test is shown in Appendix F. Specifically, the average Cosine Similarity and Spearman's Rank Correlation Coefficient for LLMs are substantially below those of human respondents, with a huge difference exceeding 0.42. The Value Mean Difference for LLMs averages around 1.52, indicating a substantial divergence in self-knowledge between the two types of questionnaires for LLMs. And as for most LLMs, the proportion of consistent pairs falls below $55\%$, raising questions about LLMs' ability to achieve self-knowledge-action unity in practice.

## 4. Conclusion

We demonstrate that while LLMs exhibit some capacity to mimic human-like tendencies, there are significant gaps in the coherence between their stated personality and exhibited behaviors. This disparity probably suggests a limitation in LLMs' ability to authentically replicate human personality dynamics. Our study underscores the importance of further exploration into enhancing LLMs' ability to perform more genuinely human-like interactions, suggesting avenues for future research in improving the psychological realism of

LLM outputs.

## Limitations

In this study, we delve into the alignment between what Large Language Models (LLMs) claim and their actions, aiming to discern if there's a consistency in their self-knowledge and their actual behavior tendency. This observation is merely one among several hypotheses exploring the root causes of this inconsistency, underscoring the need for further investigation into the fundamental reasons behind it. Moreover, the scope of our initial experiments was limited to a selection of several LLMs. Future endeavors will expand this investigation to encompass a broader array of models. Additionally, our study has yet to identify an effective strategy for enhancing the congruence between LLMs' self-knowledge and action. As we move forward, our efforts will focus on leveraging the insights gained from this research to improve the performance and reliability of LLMs, paving the way for models that more accurately mirror human thought and behavior.

## Impact Statement

Our personality knowledge survey leverages the TDA-100, BFI-44, and the 16 Personalities Test, which are extensively recognized and employed within the personality knowledge domain. These tests, available in both Chinese and English, are backed by thorough reliability and validity analyses. We ensured the integrity of these instruments by maintaining their original content without any modifications. The design of every questionnaire intentionally avoids any bias related to gender and is free from racial content, fostering an inclusive approach. Participants' anonymity was strictly preserved during the survey process. Moreover, all individuals were fully informed about the purpose of the study and consented to their responses being utilized for scientific research, thereby arising no ethical issues.

## Acknowledgement

This paper is partially supported by SMP-Zhipu.AI Large Model Cross-Disciplinary Fund.

## References

Allport, G. W. Personality: A psychological interpretation. 1937. URL https://psycnet.apa.org/record/1938-01964-000.

Brown, A. and Maydeu-Olivares, A. Ordinal factor analysis of graded-preference questionnaire data. *Structural Equation Modeling: A Multidisciplinary Journal*, 25(4):516–529, 2018. URL https://www.tandfonline.com/doi/abs/10.1080/10705511.2017.1392247.

Costa, P. T. and McCrae, R. R. The revised neo personality inventory (neo-pi-r). *The SAGE handbook of personality theory and assessment*, 2(2):179–198, 2008.

Gabriel, I. Artificial intelligence, values, and alignment. *Minds and machines*, 30(3):411–437, 2020. URL https://link.springer.com/article/10.1007/s11023-020-09539-2.

Goldberg, L. R. Language and individual differences: The search for universals in personality lexicons. *Review of personality and social psychology*, 2(1):141–165, 1981. URL https://www.scienceopen.com/document?vid=3cdca9a2-ab50-48bf-97b5-0c2236e65098.

Goldberg, L. R. The development of markers for the big-five factor structure. *Psychological assessment*, 4 (1):26, 1992. URL https://psycnet.apa.org/record/1992-25730-001.

Golfried, M. R. and Kent, R. N. Traditional versus behavioral personality assessment: A comparison of methodological and theoretical assumptions. *Psychological Bulletin*, 77(6):409, 1972. URL https://psycnet.apa.org/record/1972-29191-001.

GU, X.-Y. and Hu, S. Mbti: New development and application. *Advances in Psychological Science*, 20(10):1700, 2012. URL https://journal.psych.ac.cn/xlkxjz/EN/10.3724/SP.J.1042.2012.01700.

Huang, J., Wang, W., Lam, M. H., Li, E. J., Jiao, W., and Lyu, M. R. Revisiting the reliability of psychological scales on large language models, 2023. URL https://arxiv.org/abs/2305.19926v3.

Huang, J., Wang, W., Li, E. J., Lam, M. H., Ren, S., Yuan, Y., Jiao, W., Tu, Z., and Lyu, M. R. Who is chatgpt? benchmarking llms' psychological portrayal using psychobench, 2024. URL https://arxiv.org/abs/2310.01386.

John, O. P., Donahue, E. M., and Kentle, R. L. Big five inventory. *Journal of personality and social psychology*, 1991. URL https://psycnet.apa.org/doiLanding?doi=10.1037%2Ft07550-000.

Johnson, W. L., Johnson, A. M., Murphy, S. D., Weiss, A., and Zimmerman, K. J. A third-order component analysis of the myers-briggs type indicator. *Educational and psychological measurement*, 58(5):820–831, 1998. URL https://journals.sagepub.com/doi/abs/10.1177/0013164498058005007.

Joshi, A., Kale, S., Chandel, S., and Pal, D. K. Likert scale: Explored and explained. *British journal of applied science & technology*, 7(4):396–403, 2015. URL http://research.sdpublishers.net/id/eprint/2464/.

Makwana, K. and Dave, D. G. B. Confirmatory factor analysis of neris type explorer® scale–a tool for personality assessment. *International Journal of Management*, 11(9), 2020. URL https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3709640.

Miotto, M., Rossberg, N., and Kleinberg, B. Who is GPT-3? an exploration of personality, values and demographics. In Bamman, D., Hovy, D., Jurgens, D., Keith, K., O'Connor, B., and Volkova, S. (eds.), *Proceedings of the Fifth Workshop on Natural Language Processing and Computational Social Science (NLP+CSS)*, pp. 218–227, Abu Dhabi, UAE, November 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.nlpcss-1.24. URL https://aclanthology.org/2022.nlpcss-1.24.

Myers, I. B. The myers-briggs type indicator: Manual (1962). 1962. URL https://psycnet.apa.org/record/2013-29682-000?doi=1.

Nardi, D. Neuroscience of personality. *Neuroscience*, 2: 10–2012, 2011. URL https://core.ac.uk/pdf/aaa287819128.pdf.

Roberts, B. W. and Yoon, H. J. Personality psychology. *Annual review of psychology*, 73:489–516, 2022. URL https://www.annualreviews.org/doi/abs/10.1146/annurev-psych-020821-114927.

Roberts, B. W., Kuncel, N. R., Shiner, R., Caspi, A., and Goldberg, L. R. The power of personality: The comparative validity of personality traits, socioeconomic status, and cognitive ability for predicting important life outcomes. *Perspectives on Psychological science*, 2(4):313–345, 2007. URL https://journals.sagepub.com/doi/abs/10.1111/j.1745-6916.2007.00047.x.

Safdari, M., Serapio-García, G., Crepy, C., Fitz, S., Romero, P., Sun, L., Abdulhai, M., Faust, A., and Matarić, M. Personality traits in large language models. *arXiv preprint arXiv:2307.00184*, 2023. URL https://arxiv.org/abs/2307.00184.

Saucier, G. and Goldberg, L. R. Lexical studies of indigenous personality factors: Premises, products, and prospects. *Journal of personality*, 69(6):847–879, 2001.

Sisson, E. D. Forced choice—the new army rating 1. *Personnel Psychology*, 1(3):365–381, 1948. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1744-6570.1948.tb01316.x.

Zhang, B., Luo, J., and Li, J. Moving beyond likert and traditional forced-choice scales: A comprehensive investigation of the graded forced-choice format. *Multivariate Behavioral Research*, pp. 1–27, 2023. URL https://www.tandfonline.com/doi/abs/10.1080/00273171.2023.2235682.

Zhang, X. *Preliminary revision of the Big Five Personality Inventory (IPIP NEO-PI-R)*. PhD thesis, Yangzhou University, 2012.