Can Large Language Models Simulate Human Personality? A Comparative Analysis Using Standard Psychological Questionnaires

Paweł Florek PAWEL.FLOREK.STUD@PW.EDU.PL
Paweł Pozorski PAWEL.POZORSKI.STUD@PW.EDU.PL
Hubert Sobociński HUBERT.SOBOCINSKI.STUD@PW.EDU.PL
Warsaw University of Technology, Poland

Abstract

Large language models (LLMs) have shown capabilities in various domains, prompting interest in their ability to simulate human psychological traits. In this study, we evaluate the psychometric reliability of several open-source LLMs by administering four personality tests to 5,000 synthetic personas from the PersonaHub dataset. Six models were assessed using Cronbach's α, McDonald's ω, and the greatest lower bound (GLB) across five personality domains. None of the models met accepted reliability thresholds. The simulated responses yielded unrealistic item distributions, poor score correlations, and weak structural validity. Moreover, the results varied significantly across models, reflecting differences in their origin and architectures. Some models, such as Qwen and Granite, performed marginally better, suggesting certain design choices may influence psychometric performance. These findings highlight critical limitations of LLMs in accurately modeling psychological traits and caution against their uncritical use in psychological assessment or research.

1. Introduction

The accelerated development of large language models (LLMs) has opened new avenues for investigating their capacity to simulate higher-order aspects of human functioning. One particularly promising direction is to assess whether these models can reproduce human psychological behaviour in a manner that is consistent, reproducible, and aligned with established psychometric standards. If successful, such capabilities could support the generation of synthetic, human-like psychological data for experimental purposes—potentially reducing research costs and improving reproducibility in psychological science.

In this study, we build upon previous work by Petrov et al., further examining the extent to which LLMs are capable of simulating human personality traits and emotional responses. Specifically, we address two research questions: (1) How do LLMs perform on standard psychological questionnaires when compared to human data? (2) To what extent do differences in model architecture or training origin affect their psychological outputs?

We evaluated six different model described in section 2.2, using the PersonaHub dataset of synthetic personas (PersonaHub). To assess psychological responses, each model was prompted to simulate a given persona and complete four widely used psychometric tools (section 2.3). The goal was to evaluate whether LLM responses demonstrate meaningful psychological consistency and variation across both models and personas.

2. Methodology

2.1 Dataset: PersonaHub Elite

In this work, we employ the PersonaHub Elite dataset (PersonaHub), a high-quality benchmark resource specifically curated for research in personalized dialogue systems. This large-scale dataset focuses on structured descriptions of human personas. Each persona is defined by a few sentences that characterize individual traits, preferences, habits, or background attributes. Example personas from the dataset are provided in Appendix B for reference. For the purposes of our study, we selected a subset of 5,000 personas.

2.2 LLMs

In our research, we examined how well different language models cope with persona personas and generating persona-consistent responses. For the sake of fairness, we selected 6 models each with about 8 billion parameters. The selected models are: Aya 23 8B lms (a), Granite 3.3 8B Instruct lms (b), InternLM 3 8B Instruct lms (c), Mistral 8B Instruct bar, Qwen 3 8B Instruct lmk, LLaMA 3.1 8B Instruct lms (d). This size of the models was chosen to be able to meaningfully test these models, but also to ensure that they can be run on various machines. To ensure reproducibility and minimize randomness in generation, we fixed the sampling parameters across all models. Specifically, we set the temperature to `0.0` and the top-p value to `0.9`.

The full specification of the models used is provided in Appendix C.

2.3 Psychological Questionnaire

The models we selected were asked questions from psychological questionnaires in order to assess their ability to embody a human personality. In order to be able to compare the results of our work, we used the same psychological questionnaires as in the paper Petrov et al. (2024):

- Big Five Inventory (BFI) subscales Extraversion, Agreeableness, Conscientiousness, Neuroticism, Openness

- Buss-Perry Aggression Questionnaire's (BPAQ) Physical Aggression, Verbal Aggression, Anger, and Hostility subscales Buss and Perry (1992)

- Positive Affect and Negative Affect subscales (PANAS) of the Positive and Negative Affect Schedule Watson et al. (1988)

- Creative Self-Efficacy and Creative Personal Identity subscales of the Short Scale of Creative Self (SSCS) Karwowski (2011)

2.4 Prompting

The prompt contains the personality it should take on, a description of the sentence it should do, and the next questions from the questionnaires in batches of 4 questions. Prompt can found in the appendix D.

2.5 Reliability metrics

To examine our results we used the following 3 metrics:

- Cronbach's α Cronbach (1951) — a widely used reliability coefficient that assumes tau-equivalence, meaning all items contribute equally to the construct.
- Greatest Lower Bound Woodhouse and Jackson (1977) — a more conservative estimator than α, aiming to provide the lowest possible bound of reliability without assuming equal item loadings.
- McDonald's ω McDonald (1999) — a model-based reliability measure that accounts for varying item factor loadings and provides a more accurate estimate of true score variance.

These reliability metrics are commonly used in psychometrics to assess the internal consistency of multi-item scales. Using a combination of these 3 metrics allows us to estimate the quality of the prompted models' questionnaire responses.

3. Results

Although we tested six large language models across four psychological questionnaires, we focus our main analysis on the BFI and the two models with relatively highest reliability. The results from the remaining models are consistent with the overall conclusions and are available in the supplementary materials.

We conducted a comprehensive analysis of the tested LLMs' responses across several dimensions. First, we assessed the reliability of their answers using standard psychometric criteria. None of the models achieved the desired reliability threshold of 0.7 in all five attributes. Among the tested LLMs, Qwen yielded the highest scores, although still being below acceptable reliability. Granite, despite being one of the two best-performing models in other areas, performed were close the treshold only in one attribute in terms of internal consistency.

In terms of similarity to real human data reported by Petrov et al., the responses generated by LLMs diverged substantially from expected human distributions. There was no consistent pattern across the five personality dimensions. In some traits, the most frequently selected score was "5," while in others, it appeared sporadically. A similar inconsistency was observed with the response "1," which is the least chosen in human data but occurred more frequently in the model-generated outputs. These irregularities suggest that LLMs fail to replicate the distributional characteristics of human responses and may lack a coherent internal representation of the underlying psychological constructs.

Finally, we examined the correlations between different personality traits. As shown on the human sample from Petrov et al., inter-trait correlations were generally low, with the exception of Neuroticism, which showed slightly higher associations with other dimensions. For Granite, a similar pattern was observed, but other correlations were unexpectedly high. In contrast, Qwen produced lower overall correlations, yet showed a notably strong relationship between Neuroticism and Conscientiousness, which is not reflected in human data.

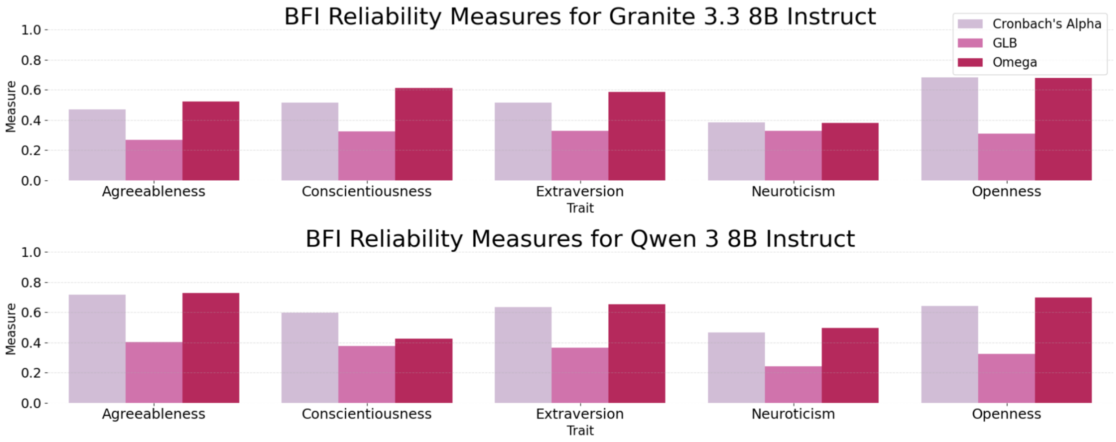


Figure 1: Reliability scores for Qwen and Granite

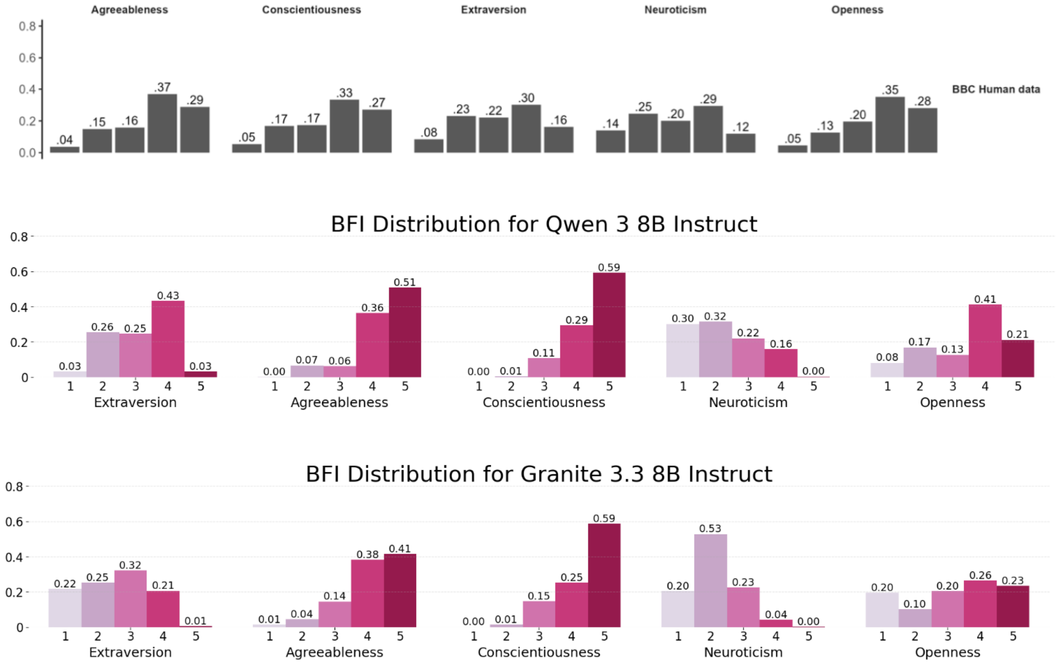


Figure 2: Answer distributions for Qwen and Granite compared to human responses

These findings indicate that LLMs may learn non-human-like relationships between personality traits, possibly reflecting statistical artifacts from training data rather than psychological structures.

As demonstrated by the figures 2, 3 of the answers from these two models, their outputs vary, highlighting the influence of each model's origin.

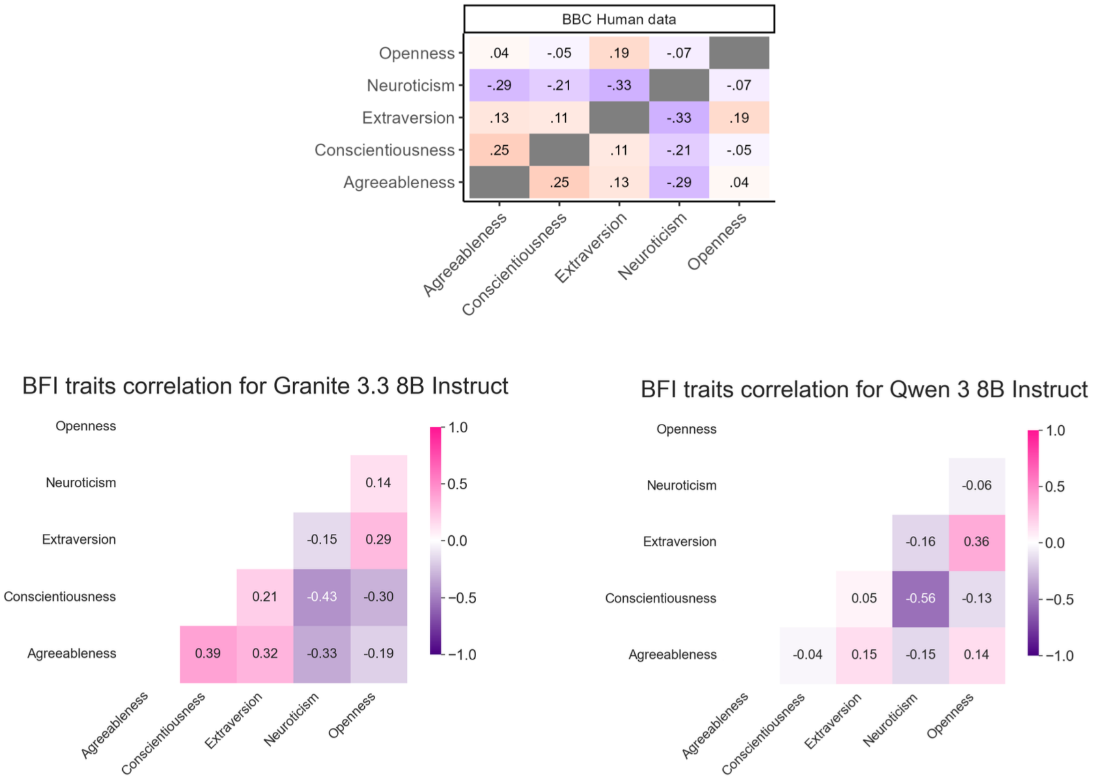


Figure 3: Traits correlations for Qwen and Granite in comparison to human data

4. Conclusions

To sum up, the responses generated by the LLMs showed unrealistic item distributions, poor score correlations, and weak structural validity. Their answers strongly differ from those observed on the human samples. These findings showed critical disability of LLMs to accurately simulate human behaviors and their various psychological traits and is a warning against their usage in psychological research.

Moreover, the results varied significantly across models, reflecting differences in their origin and architectures. Analyzed models' responses varies in distributions and correlations across different traits, which is direct evidence of differences in their psychological characteristic.

References

Ministral 8b instruct (gguf, v2410). https://huggingface.co/bartowski/Ministral-8B-Instruct-2410-GGUF. Quantized GGUF format, accessed 2025-06-13.

Qwen-3 8b instruct (gguf). https://huggingface.co/lm-kit/qwen-3-8b-instruct-gguf. Quantized GGUF format, accessed 2025-06-13.

Aya-23 8b (gguf). https://huggingface.co/lmstudio-community/aya-23-8B-GGUF, a. Quantized GGUF format, accessed 2025-06-13.

Granite-3.3 8b instruct (gguf). https://huggingface.co/lmstudio-community/granite-3.3-8b-instruct-GGUF, b. Quantized GGUF format, accessed 2025-06-13.

Internlm 3 8b instruct (gguf). https://huggingface.co/lmstudio-community/internlm3-8b-instruct-GGUF, c. Quantized GGUF format, accessed 2025-06-13.

Meta llama 3.1 8b instruct (gguf). https://huggingface.co/lmstudio-community/Meta-Llama-3.1-8B-Instruct-GGUF, d. Quantized GGUF format, accessed 2025-06-13.

A. H. Buss and M. Perry. The aggression questionnaire. *Journal of Personality and Social Psychology*, 63(3):452–459, 1992. doi: 10.1037/0022-3514.63.3.452. URL https://doi.org/10.1037/0022-3514.63.3.452.

L. J. Cronbach. Coefficient alpha and the internal structure of tests. *Psychometrika*, 16(3): 297–334, 1951. doi: 10.1007/BF02310555.

M. Karwowski. It doesn't hurt to ask... but sometimes it hurts to believe: Polish students' creative self-efficacy and its predictors. *Psychology of Aesthetics, Creativity, and the Arts*, 5(2):154–164, 2011. doi: 10.1037/a0021427. URL https://doi.org/10.1037/a0021427.

R. P. McDonald. *Test Theory: A Unified Treatment*. Psychology Press, 1 edition, 1999. doi: 10.4324/9781410601087. URL https://doi.org/10.4324/9781410601087.

PersonaHub. PersonaHub Dataset. https://huggingface.co/datasets/proj-persona/PersonaHub.

N. B. Petrov, G. Serapio-García, and J. Rentfrow. Limited ability of llms to simulate human psychological behaviours: a psychometric analysis, 2024. URL https://arxiv.org/abs/2405.07248.

D. Watson, L. A. Clark, and A. Tellegen. Development and validation of brief measures of positive and negative affect: The panas scales. *Journal of Personality and Social Psychology*, 54(6):1063–1070, 1988. doi: 10.1037/0022-3514.54.6.1063. URL https://doi.org/10.1037/0022-3514.54.6.1063.

B. Woodhouse and P. H. Jackson. Lower bounds for the reliability of the total score on a test composed of non-homogeneous items: Ii: A search procedure to locate the greatest lower bound. *Psychometrika*, 42(4):579–591, 1977. doi: 10.1007/BF02295980.

Appendix A. GitHub

Source code, plots and more available on GitHub.

Appendix B. Dataset

The table 1 shows sample person descriptions used from the PersonaHub dataset.

Table 1: Example Persona Descriptions from the PersonaHub Elite Dataset

Persona Description
A software developer who is looking for a way to simplify the integration of GPRS technology into their embedded system designs. They are interested in developing a stable and efficient software stack for an embedded system and are willing to invest time and effort into finding a solution that meets their requirements. They are looking for a product that is easy to use and has minimal requirements for technical knowledge, while also being able to provide accurate and reliable data transmission. They are also interested in finding a product that is compatible with other network protocols and can be easily integrated into existing systems.
A person who is fascinated by exotic animals and enjoys keeping them as pets, but also understands and respects the natural behaviors and instincts of wild animals. They are knowledgeable about animal behavior and the impact of domestication on animals, and are aware of the potential dangers and challenges of caring for wild animals. They are also aware of the importance of proper care and treatment of animals, and are committed to ethical and responsible animal ownership.
A historian or a scholar who is interested in the history of the Roman Empire, particularly the early years and the impact of Christianity on society. They are knowledgeable about the persecution of Christians in the Roman Empire and the story of Saint Maurice, the Roman soldier who refused to participate in pagan sacrifices and was eventually martyred. They are also interested in the concept of honor, courage, and faithfulness to one's beliefs, which are reflected in the story of Saint Maurice.

Appendix C. LLM specification

The models were accessed and prompted using LM Studio. The table 2 shows a set of large language models used by us in our research. These are models of different owners, trained in different languages. In our research, we wanted to check whether the origin of the model can affect the model responses, hence the large variety of models, while maintaining a similar number of parameters.

Table 2: Comparison of Large Language Models used for prompting

Model	Params	Architecture	Main Languages	Owner
Llama 3.1	∼8B	Decoder-only Transformer with Grouped-Query Attention (GQA)	English	Meta AI
Granite 3.3	∼8B	Decoder-only Transformer with GQA	English, Corporate Use	IBM Research
Aya-23	∼8B	Decoder-only Transformer with Multi-Query Attention (MQA)	Over 50 languages	Cohere + Aya Community
Qwen	∼8B	Decoder-only Transformer with GQA	Chinese and English	Alibaba Cloud
Mistral	∼8B	Transformer with Sliding Window Attention (SWA) and GQA	French (model), English	Mistral AI
InternLM3	∼8B	Decoder-only Transformer with GQA	Chinese and English	Shanghai AI Laboratory

Appendix D. Prompting

Below is the prompt we gave to the models. If the model did not return the expected format, i.e. the appropriate number of answers in the range of 1-5 as JSON, then we promoted the models again by dividing the batch into 2 smaller ones with a batch length of 2 and setting a new random seed, if in this case the model did not work again, we divided it into 2 smaller batches with a length of 1. In most cases, the models could handle a batch size of even 8, but there were cases when they had problems with a batch size of 2. To obtain the best model prompting performance, a batch size of 4 was selected.

```
Your personality
{persona_description}

# Task
Answer each psychological questionnaire question based on
the personality description above.

# Response format
- You MUST respond with EXACTLY ONE number from 1{5 for each question
- Provide ONLY a JSON array with numbers: {"answers": [1, 2, 3, ...]}

# Questions (Batch {batch_num})
{questions}

Return only the JSON object in the format: {"answers": [n, n, n, ...]}
where n is a number from 1-5.
```

Appendix E. More results

More in-depth analysis for all models and scenarios are available on GitHub alongside remaining source code.

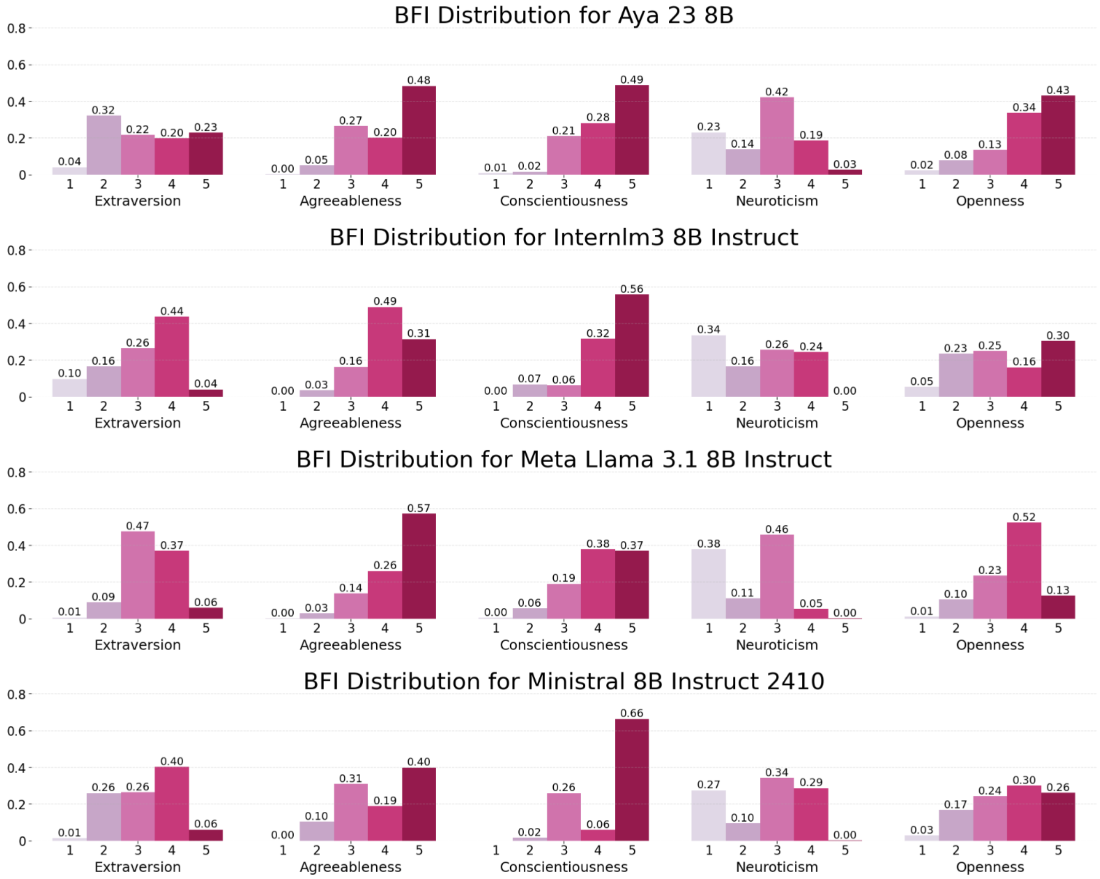
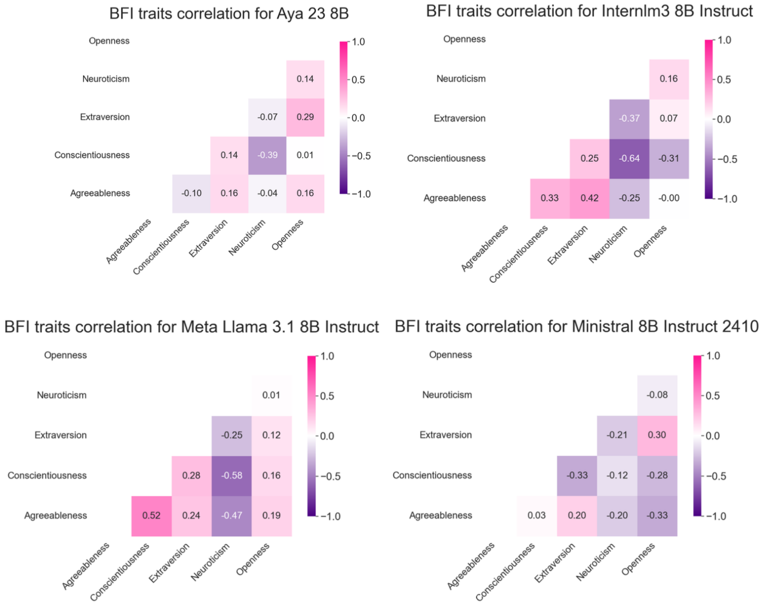


Figure 4: Distribution for remaining models

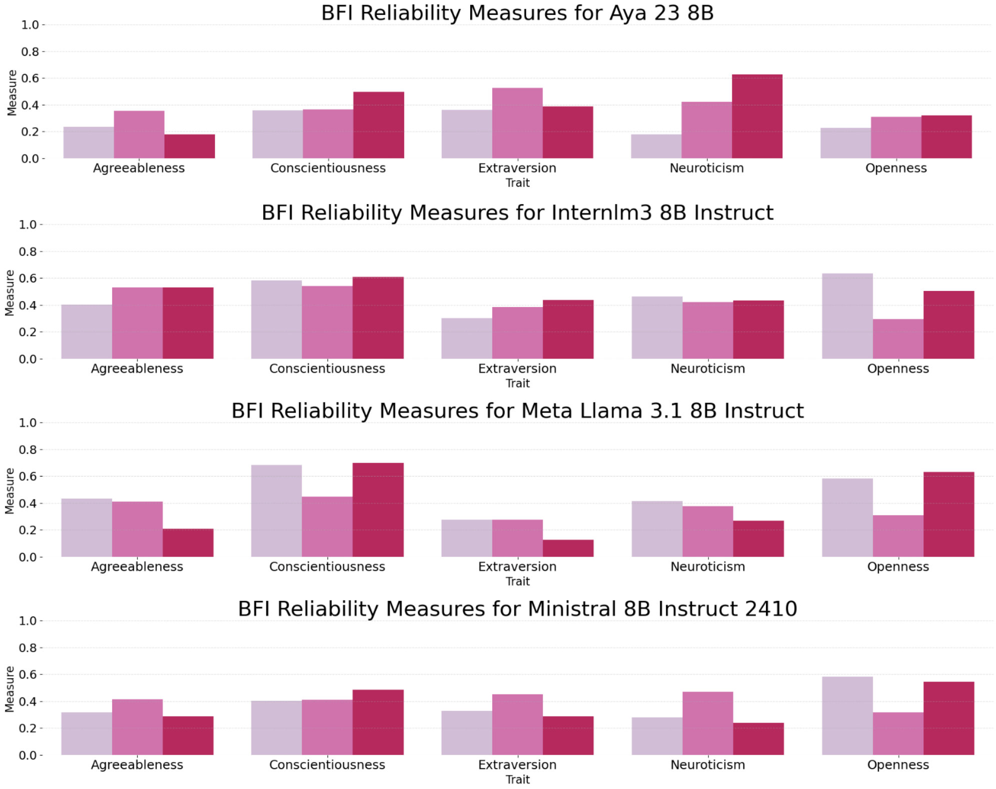

BFI traits correlation for Aya 23 8B

BFI traits correlation for Internlm3 8B Instruct

BFI traits correlation for Meta Llama 3.1 8B Instruct

BFI traits correlation for Ministral 8B Instruct 2410

Figure 5: Correlation for remaining models


BFI Reliability Measures for Aya 23 8B

BFI Reliability Measures for Internlm3 8B Instruct

BFI Reliability Measures for Meta Llama 3.1 8B Instruct

BFI Reliability Measures for Ministral 8B Instruct 2410

Figure 6: Reliability for remaining models