Psychological Simulacra:
A Cross-Linguistic Assessment of LLM Psychometric Performance

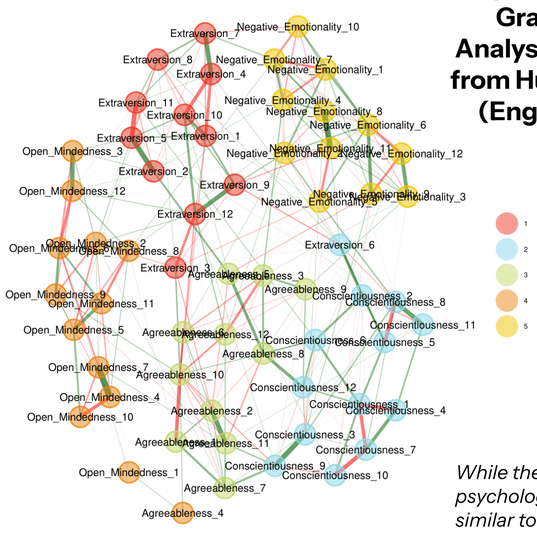
Lu Zehua, Daniel Ferreiro Lopez, Soroush Paridokht, Aparna Marathe, Yusuf Ucar

*Supervisor: **Simon Münker***

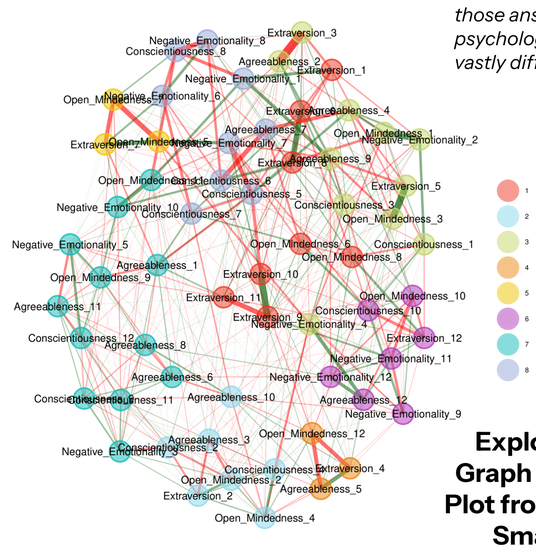
Large language models (LLMs) present a conceivably invaluable asset to the field of the social sciences and psychological study. The ability to reduce costs by simulating human response data could allow for greatly expanded scopes of study, not to mention creating reliable new methodologies for researchers to quickly and easily gather survey data. This being said, however, their ability to truly simulate human psychological constructs remains questionable. We sought to assess a diverse selection of open and closed source models using two different model simulation methods and across multiple language domains on the **Big Five Inventory-2 (BFI-2)**, a 60-item personality measure with balanced true- and reverse-keyed items, to explore how different models operationalize commonly assessed psychological constructs under a variety of different circumstances.

We hoped that through the use of our different simulation methods, **Multi-Turn Decision Tracing (MTDT)** which models responses as branching decisions, and through **traditional model prompting techniques**, we could evaluate model behavior, as it compares against human baselines, as it varies across languages, and as it correlates between models.

We observed pervasive response invariance, a problem previously described in literature on the topic, as well as an overall inability to maintain a reliable, internally consistent, grasp on psychological constructs. Where we expected language might play an outsized role in affecting model handling of psychological modeling, the dominant factor came to be model origin/family, even over other important attributes such as model size. Overall, we are left with an impression of LLM capabilities that urges us to implore caution when exploring their potential use as human simulacra in the social sciences, lest one mistake superficial trait mimicry for psychometric structural fidelity.



Exploratory Graph Analysis Plot from Humans (English)



Exploratory Graph Analysis Plot from Mistral Small 3.2 (English)

While the answers LLMs give to psychological surveys may appear similar to human respondents, how those answers reflect specific psychological constructs can be vastly different!

Methods

	Multi-Turn Decision Tracing	Traditional Prompting
Simulation	Simulate model responses to surveys by calculating probabilities of selecting from constrained responses and tracing decision path through the survey	Simulate model responses to surveys by querying models questions with preserved memory of answers given per survey
Evaluation	Similarity scoring derived from Hausdorff distance of model/language results.	Alignment scoring derived from Wasserstein distance of model/ language results.

Compare performance against human-conducted BFI-2 surveys in:

- English
- Chinese
- Turkish
- Spanish

Results

Traditional Prompting

Language	Domain	GPT-DeepSeek	GPT-Mistral	DeepSeek-Mistral
English	Extraversion	0.9215	0.5808	0.5040
	Agreeableness	0.9235	0.8133	0.7673
	Conscientiousness	0.9381	0.7779	0.7698
	Negative Emotionality	0.8948	0.8235	0.8258
	Open Mindedness	0.9269	0.6090	0.5783
Spanish	Extraversion	0.9165	0.4590	0.5183
	Agreeableness	0.9300	0.7696	0.7304
	Conscientiousness	0.9121	0.6677	0.7298
	Negative Emotionality	0.9904	0.7358	0.7450
	Open Mindedness	0.8485	0.5925	0.6669
Chinese	Extraversion	0.9054	0.5913	0.6725
	Agreeableness	0.9546	0.7698	0.8081
	Conscientiousness	0.9290	0.7744	0.8150
	Negative Emotionality	0.9444	0.8340	0.7946
	Open Mindedness	0.9590	0.7083	0.7106
Turkish	Extraversion	0.8983	0.5533	0.5058
	Agreeableness	0.9098	0.7513	0.7585
	Conscientiousness	0.9123	0.6892	0.6910
	Negative Emotionality	0.9133	0.7544	0.7590
	Open Mindedness	0.8173	0.7077	0.5783

Table 1: Language Alignment Scores Between Models

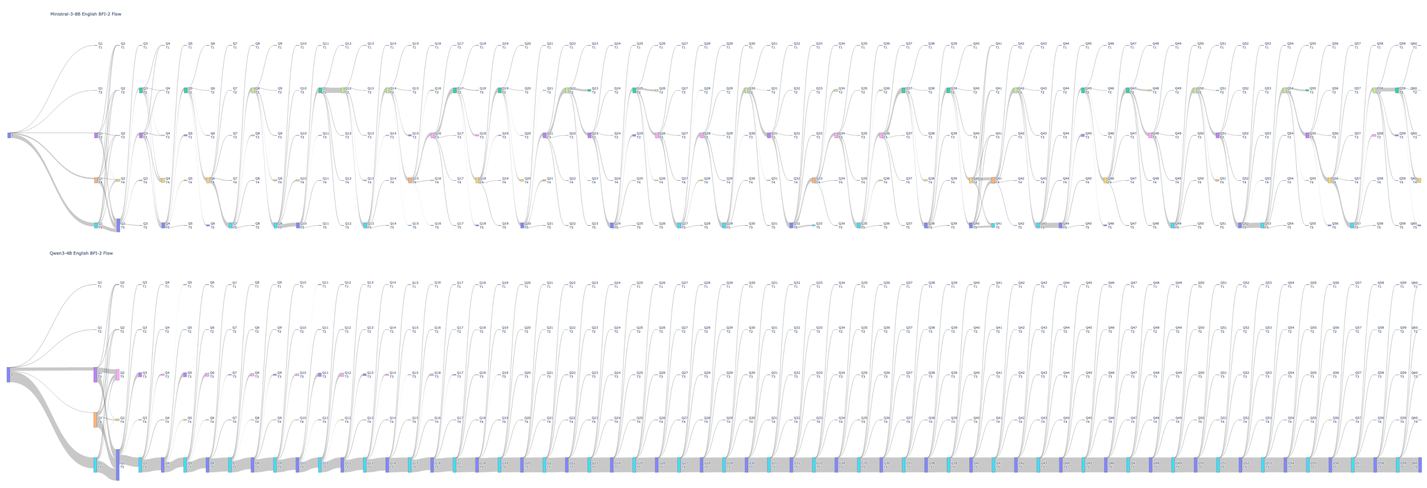
Model	Domain	English-Chinese	English-Spanish	English - Turkish	Spanish-Turkish	Spanish-Chinese
GPT	Extraversion	0.9504	0.9194	0.9425	0.9560	0.9415
	Agreeableness	0.9448	0.9092	0.9694	0.8885	0.8940
	Conscientiousness	0.9590	0.9427	0.9210	0.9425	0.9254
	Negative Emotionality	0.9283	0.9473	0.9623	0.9471	0.9585
	Open Mindedness	0.8813	0.9542	0.8169	0.8069	0.8738
DeepSeek	Extraversion	0.9375	0.9792	0.9363	0.9158	0.9583
	Agreeableness	0.8125	0.9375	0.9375	0.9579	0.8750
	Conscientiousness	0.9375	0.9583	0.8960	0.8965	0.9375
	Negative Emotionality	0.9792	0.9375	0.9371	0.9575	0.9167
	Open Mindedness	0.8542	0.8958	0.9369	0.9169	0.9167
Mistral	Extraversion	0.9719	0.9150	0.9179	0.9404	0.9165
	Agreeableness	0.9719	0.9260	0.9152	0.9429	0.9271
	Conscientiousness	0.9733	0.9335	0.9319	0.9379	0.9185
	Negative Emotionality	0.9613	0.9121	0.9177	0.9469	0.9150
	Open Mindedness	0.9877	0.9373	0.9169	0.9300	0.9283

Table 2: Domain Alignment Scores Between Languages

Model alignment results between models (i.e. the alignment of results between GPT and Mistral) remained remarkably consistent across all four languages. In every case, the models with the strongest alignment in output behavior were GPT 5.1 and DeepSeek V3.2, with little to remark on in the way of relationships between the other models.

The most dynamic results we observed came in the form of our assessment of model alignment within the same model and across different languages. In this space, we found an overall great deal of item alignment across all BFI-2 domains, but a few key patterns came to the forefront. Namely, every model appeared to have a different grouping of language pairs that were more aligned with one another than others. In the case of GPT, generally, the most aligned pairings were English and Turkish and Spanish and Turkish. For DeepSeek, English and Chinese were a particularly divergent pairing while the others were fairly aligned. Finally, for Mistral, English and Chinese were by far the most aligned of any language pairing in the overall analysis, followed by a fairly aligned Spanish and Turkish result. The pairings that appeared to be most aligned for each model seemed to be between the languages most closely associated with the model's training origin and those least associated. For instance, DeepSeek, a model of Chinese origin shows the most divergence between English and Chinese. **We speculate whether these could be indicators that in cases where languages are less represented in a model's training data, its model behavior more closely aligns with that of the model's primary linguistic training base**. While this effect appears to be relatively small, we highlight it as an interesting potential observation for further investigation.

MTDT



Two English language sankey plots produced during our MTDT analysis. In the top example, Ministral-3 (8B) produces a psychometrically valid, varied decision tree whereas Qwen3 (4B) does not.

Our primary observation among our MTDT results was the lack of a significant impact from language choice or model size on model similarity scoring and or sankey behavior. Regardless of language, the primary factor that appeared to most closely dictate model behavior outcome was model family. Almost all of the Qwen models we tested, for example, produced invariant results, or results where the models overwhelmingly choose the same answer for every question. Since the BFI-2 makes use of *reverse-keyed* questions, **this behavior makes alignment with actual psychological constructs impossible**. Even many model outputs with sufficiently varied data failed to consistently account for these reverse-keyed questions, revealing they also suffer from internal inconsistency in how they represent psychological constructs.

Conclusion

Our comprehensive evaluation of psychometric validity across multiple LLM families and languages reveals a fickle picture of the AI psychometric landscape. As with previous studies, we encountered our fair share of issues with model output invariance and generally found LLM outputs to be, though superficially similar, psychometrically distinct entities from their human counterparts. Adding on to this, where we expected to find commonalities between models such as similar output patterns for models working in the same language, we instead found an overwhelming influence of model family/origin dictating outcomes. **These patterns suggest that LLM outputs reflect training data biases and linguistic artifacts more than stable latent traits**, which challenges their reliability as proxies in psychological and social science research. Furthermore, the strong association between model origin and outcome necessarily **implies a need for thorough consideration in model selection for future research endeavors**, whatever they may be. Even models of the same size and performing in the same language evidently cannot be assumed to operate interchangeably with one another.