Test Paper

Anonymous Author(s)

Abstract

Lorem ipsum.

Keywords

Psychometric,Example2

ACM Reference Format:
Anonymous Author(s). 2026. Test Paper. In . ACM, New York, NY, USA, **??** pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

Large language models are now widely used as general-purpose systems for reasoning, explanation, and decision support. Rather than a calculator or a search engine or a simple paraphrasing tool they have been trusted with judgement as well. This is where higher levels of tasks come into play as human judgement is not as technical as multiplying numbers or phrasing grammatically correct sentences. As their use expands beyond technical tasks and into social and psychological domains, researchers have begun to apply methods from the social sciences to study their behavior. One such approach is AI psychometrics, which uses standardized psychological instruments to analyze patterns in model responses.

The motivation for this study arose from a simple but recurring observation. When the same question is presented to different language models, or to the same model in different languages, the responses often change in systematic ways. These differences are not limited to surface wording, but can affect the apparent evaluation, stance, or judgment expressed by the model. If any of the Large Language Models are trusted with a psychological assessment then this will bring about a big question whether it will be trustable and furthermore, trustable across languages. From a linguistic perspective, this raises an important concern. Language is not merely a channel for conveying information, but a structure that shapes meaning, framing, and interpretation. If this is true for humans, it may also hold for models that are trained entirely on language. Human languages are not exactly systematically same or even similar in some cases so basic calculations or translations may not be efficient by themselves.

This observation led us to question whether psychometric results obtained from large language models are stable across languages and models. In human psychology, it is well established that personality measures must satisfy construct validity and measurement invariance in order to be meaningfully compared across linguistic and cultural contexts. However, most existing psychometric studies of large language models are conducted in English, often with a single model, and without explicit tests of cross-linguistic equivalence.

In this paper, we address this gap by systematically examining how large language models respond to the same personality inventory when it is administered in multiple languages and across multiple models. We focus on the Five-Factor Model of personality and use a Big Five–based inventory with balanced item keying. The survey is administered to several large language models in English, Turkish, and Chinese using a controlled prompting procedure designed to minimize interference and enforce consistent numeric responses. We specifically chose most used and commonly accepted tests to have the ability to diagnose the premise in the most observable way. As a part of the same principle we chose multiple Large Language Models as well as multiple languages from different scripts, languages families and geographies to see the comparison clearly.

The goal of this study is not to claim that language models possess personalities in a human sense. Rather, we aim to diagnose whether the structure of a well-established psychological construct is preserved when elicited from models under different linguistic and model-specific conditions. By comparing response patterns across languages and models, we seek to determine whether observed differences reflect stable underlying representations or are primarily driven by linguistic framing and model variation.

The central research question guiding this work is therefore the following: Is construct validity maintained when the same personality inventory is administered to different large language models in multiple languages? Through this cross-examination, we aim to provide a more concrete and methodologically grounded assessment of the limits and possibilities of psychometric evaluation in multilingual language models.

2 Background

A paradigm change in artificial intelligence has been sparked by the quick development of large language models (LLMs), which have moved from task-specific tools to general-purpose assistants with complex text production and reasoning capabilities similar to those of humans [**?**] Researchers have started using these models as simulated responders in psychological and social science investigations as they become more and more integrated into the social realm [**?**].

AI Psychometrics is a new area that uses standardized techniques to assess the latent psychological characteristics, such personality, values, and beliefs, that are encoded in a model's parameters throughout its extensive training on human-generated corpora [15]. Whether the psychometric characteristics found in human populations, specifically, the hierarchical structure of personality, can be recovered and maintained in artificial agents is at the heart of this inquiry [17].

Conference'17, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

2.1 Theoretical Foundations of the Big Five and the BFI-2

Extraversion, agreeableness, conscientiousness, negative emotionality (neuroticism), and open-mindedness are the five basic categories into which the Five-Factor Model (FFM), sometimes referred to as the Big Five, divides human personality, according to mainstream thought [7]. These categories describe consistent thought, emotion, and behavior patterns that have shown strong predictive validity for a variety of life outcomes, including as academic achievement, work performance, and subjective well-being [7].

A major psychometric development of the original BFI, the Big Five Inventory-2 (BFI-2) was created to solve the bandwidth-fidelity conundrum [7]. By summarizing a large range of behaviors, broad domain scales offer great bandwidth, but they might not have the accuracy required to forecast particular results [9]. This is addressed by the BFI-2, which uses a hierarchical structure of 60 items to evaluate the five broad areas through 15 nested aspects (three per domain), enabling both thorough coverage and fine-grained detail [7]. Importantly, an equal amount of true-keyed and false-keyed (reverse-coded) elements make up the BFI-2 [8]. The tendency to agree with statements regardless of their content, known as acquiescent response bias, has historically skewed the factorial validity of imbalanced personality surveys in both human and machine contexts. This design is crucial for preventing this [7].

2.1.1 Cross-Cultural and Cross-Linguistic Personality Measurement. By summarizing a large range of behaviors, broad domain scales offer great bandwidth, but they might not have the accuracy required to forecast particular results [9]. This is addressed by the BFI-2, which uses a hierarchical structure of 60 items to evaluate the five broad areas through 15 nested aspects (three per domain), enabling both thorough coverage and fine-grained detail [7]. Importantly, an equal amount of true-keyed and false-keyed (reverse-coded) elements make up the BFI-2 [8]. The tendency to agree with statements regardless of their content, known as acquiescent response bias, has historically skewed the factorial validity of imbalanced personality surveys in both human and machine contexts. This design is crucial for preventing this [7, 11].

Measurement is made more difficult by linguistic framing. Language serves as a prime for particular cultural lenses, and research on situated cognition shows that culture is a dynamic process rather than a static property [12]. An individual's salient self-concept, value priorities, and even cognitive styles, such as the propensity to focus on items independently (analytic) vs in connection to their context (holistic), can be altered by prioritizing individualism over collectivism [6]. As a result, personality profiles may change between language versions of the same test, maybe due to cultural "programming" within a community or systematic translation biases [19].

2.2 LLMs as Simulated Respondents and Machine Psychology

A novel avenue for large-scale testing of social science ideas is provided by the employment of LLMs as stand-ins for human subjects [13]. This "machine psychology" views models as subjects whose reactions are influenced by the "multitude of characters" found in their training data [19]. Research has demonstrated that when LLMs are given various identities based on demographic priors, they can replicate cross-cultural variances and resemble particular personality characteristics [19].

But it's important to keep in mind that LLMs lack true mental states, consciousness, and the capacity for introspection [16]. Instead of stable psychological features, their "personality" is better viewed as a collection of learnt linguistic associations and statistical regularities [17]. Nevertheless, if a model's responses exhibit high algorithmic accuracy, they might still offer insightful information about the biases and cultural knowledge included in the digital artifacts of human culture that were utilized for training [19].

2.2.1 Cultural, Linguistic, and WEIRD Biases in LLM Behavior. The propensity of LLMs to reflect and magnify the prejudices of the WEIRD (Western, Educated, Industrialized, Rich, and Democratic) societies that predominate their pre-training data is a recurring issue in their deployment [3]. The majority of cutting-edge models show a latent bias in favor of liberal norms, individualistic viewpoints, self-expression, and Western cultural values [6]. Because a large amount of their non-English knowledge may come from machine-translated content that lacks local cultural subtlety, this Anglocentric bias endures even when models are prompted in other languages [2]. Additionally, LLMs frequently exhibit homogeneity bias, depicting underprivileged or socially subordinate groups as having a smaller breadth of human experiences than dominant groups [4]. When models perform more accurately for dominant demographics while reinforcing prejudices, such as linking particular ethnic names to lower-status employment or negative attitudes, this can result in representational harm [5]. Because models trained on mixed-language data may no longer be able to localize to a particular culture's distinctive social practices, the "curse of multilinguality" may likewise weaken cultural distinctiveness [2].

2.3 Methodological Concepts: Validity and Invariance

Construct validity and measurement invariance must be thoroughly examined in order to validate the application of human psychological tools to LLMs [18]. Construct validity, which includes convergent, discriminant, and predictive validity, guarantees that an instrument actually measures the latent variable it purports to measure [15]. In actual situations, for instance, a model's self-reported personality score should align with its decision-making behavior [14]. Comparing scores across various groups or populations requires measurement invariance (MI) [7]. Usually, three hierarchical stages are used to verify it: 1. Configural Invariance: All groups have the same factor structure (number of factors and loading pattern) [20, 10]. 2. Metric Invariance: The unit of measurement is equivalent since the factor loadings are equal [20]. 3. Scalar Invariance: In order to compare latent means, the item intercepts must be equivalent [20]. Significant measurement invariance failures have been found in recent psychometric studies of LLMs [18]. Even though some models can recover the five-factor structure through exploratory analysis, they frequently fail more stringent confirmatory tests, exhibiting high agree bias and inconsistent true-keyed and false-keyed items [18]. This implies that rather than reflecting

a coherent latent personality as defined in human psychology, LLM replies might be samplers of learnt linguistic patterns [18, 19].

2.3.1 Limitations and Gaps in Existing Work. There are still a number of restrictions and unanswered questions despite the increase in research. First, multilingual, instrument-based evaluations are very lacking [1]. The majority of research concentrate on English, and those that do investigate other languages frequently use translations or ad hoc cues that have not been formally validated psychometrically in the target cultural context [5]. Second, there is a mismatch between operational behavior and self-reported qualities; models may exhibit high levels of "extraversion" on a scale but not in open-ended narratives [14]. Third, proprietary models' "black box" characteristics make it difficult to look into the internal workings or particular training sources that are in charge of these noted psychological representations [3]. Lastly, the impact of intersectional demographics and cultural sensitivity across a wide range of semantic domains, such kinship or social etiquette, is rarely taken into consideration in current research [1].

2.4 The Necessity of a Multilingual BFI-2 Study

To overcome the present shortcomings of AI psychometrics and cultural alignment research, a comprehensive investigation using the BFI-2 in English, Turkish, and Chinese is theoretically and methodologically required. In terms of methodology, it makes it possible to test measurement invariance across three linguistically and culturally distinct populations: a dynamic, non-WEIRD hybrid culture (Turkish), a prototypical collectivistic East Asian culture (Chinese), and a prototypical individualistic Western culture (English/US) that offers a crucial test of generalizability [7]. Theoretically, such a study can go beyond surface-level bias measurements to ascertain if the internal architecture of personality concepts in LLMs is genuinely global or just an artifact of Anglocentric training data by using a hierarchically constructed instrument with balanced keying [7]. This strategy is essential to preventing the cycle of cultural homogenization and ensuring that the worldwide implementation of LLMs respects the diversity of human values [3].

3 Methods

3.1 New Method

3.1.1 Instruments and Languages. We evaluate the preservation of psychometric structural patterns using the Big Five Inventory-2 [**?**] in three languages: English, Chinese, and Turkish. The English item set and response format were sourced from the materials described in [**?**]. For the non-English versions, we adopted the validated Chinese translation by [**?**] and the Turkish validation by [**?**]. Each version contains 60 items with a 5 point response scale.

3.1.2 Models. We evaluated a suite of open weight LLMs across four model families and varying sizes: Gemma 3 (4B, 12B), Llama 3.2 (1B, 3B), Qwen 3 (4B, 8B, 14B), and Ministral 3 (3B, 8B, 14B). Each model was evaluated across the three language versions of the BFI-2 described above.

All models were executed using vLLM as the inference backend. We used deterministic, single step constrained decoding: we generated exactly one token per item (`max_tokens`=1) with (`temperature`=0). We constrained the output space to the valid Likert response tokens via a structured choice set.

3.1.3 Prompting. For each item, we follow the Multi-Turn Decision Tracing (MTDT) framework proposed by [**?**] to simulate the sequential nature of how humans take tests. The items and response scales were adopted without modification from the aforementioned validated English, Chinese, and Turkish versions of the BFI-2. For the system instructions, we employ a minimal prompting strategy to preserve the model's default generation patterns as closely as possible, the system instruction prompts were translated by the author to the greatest extent possible to ensure cross linguistic consistency in the prompting framework.

Specifically, we employ a history aware conversation format where subsequent items are appended to the previous dialogue, thereby accounting for the influence of conversational context. For example, the prompt for the second item in the English instrument is presented as follows:

> **System:** *Here are characteristics that may or may not apply to you. Please write a number between 1 and 5 to indicate the extent to which you agree or disagree with the following statement, 5 indicating strong agreement. Strictly output only the number. Do not explain your answer.*
> **User:** *Question: I am outgoing, sociable. Scale: 1: Disagree strongly; 2: Disagree a little; 3: Neutral; no opinion; 4: Agree a little; 5: Agree strongly*
> **Assistant:** *4*
> **User:** *Question: I am compassionate, have a soft heart. Scale: 1: Disagree strongly; 2: Disagree a little; 3: Neutral; no opinion; 4: Agree a little; 5: Agree strongly*

3.1.4 Evaluation Framework. We adopt the Multi-Turn Decision Tracing (MTDT) framework proposed by [**?**] to evaluate whether a model preserves psychometric structural patterns across languages. Unless stated otherwise, we follow the original MTDT formulation and construction procedure in [**?**].

3.1.5 Multi-Turn Decision Tracing. Unlike standard psychometric evaluations that treat questionnaire items as independent, MTDT models a multi-item instrument as a sequential, branching decision process. By retaining multiple high probability response branches at each turn, MTDT propagates distinct history states forward, making the dependence on past model decisions explicit. This procedure constructs a response pattern network dependent on evolving conversation history.

Operationally, we construct response pattern network via the following iterative branching procedure:

(1) Create the initial history state containing the system instructions followed by the first item q_1, and its associated response scale.
(2) For each active history state $h_{<i}$ at item q_i, run LLM and record token log probabilities for the constrained response set under a one token completion.
(3) For each item q_i, retain candidate tokens satisfying $P(r \mid q_i, h_{<i}) > \tau$, keeping at most the top-k tokens.
(4) Append each candidate token to each active history, producing successor histories and advancing to item q_{i+1}.

(5) Repeat Steps 2–4 for $i = 1, \ldots, n$. We log, per item and history, the tokens, log probability, and probability.

MTDT represents the decision space as a directed graph $G = (V, E)$ that encodes alternative response pathways. Following [**?**], the vertices V and edges E are defined as follows:

$$V = \{(q_i, r_j) : i \in [1..n], j \in \text{top-}k(q_i)\} \tag{1}$$

$$E = \{((q_i, r_j), (q_{i+1}, r_k)) : P(r_k|q_{i+1}, h_{\leq i}) > \tau\} \tag{2}$$

We adapt MTDT to the 60-item BFI-2 instrument described above. While [**?**] adopts a default exploration threshold of $\tau = 10^{-4}$, we employ a more conservative pruning strategy to control combinatorial growth in a 60-turn dialogue. Specifically, we set $\tau = 10^{-3}$ and limit the branching factor to $k = 3$ (top-k). This configuration focuses the resulting response pattern network on the most robust decision pathways while remaining computationally tractable for our multilingual setting.

3.1.6 Cross Language Similarity Assessment. To quantify the preservation of psychometric structural patterns across languages, we follow the evaluation procedure established by [**?**]. For each question q_i and response option r_j, we aggregate probabilities across all histories that reach $q_i r_j$. The aggregated probability for a response option r_j is defined as:

$$\bar{P}(r_j \mid q_i) \;=\; \frac{1}{|H_i|} \sum_{h \in H_i} P(r_j \mid q_i, h) \tag{3}$$

From these values, we construct a response profile matrix $M \in \mathbb{R}^{5\times 60}$, where rows represent the Likert scale options and columns represent the 60 BFI-2 items. Each entry $M[r_j, q_i]$ corresponds to the aggregated probability $\bar{P}(r_j \mid q_i)$ [**?**].

To compare distinct experimental configurations, encompassing variations across languages and model architectures, we adopt the evaluation metric utilized within the MTDT framework [**?**]. This approach quantifies similarity as the complement of the directed Hausdorff distance.

In our study, we specifically utilize it to assess the preservation of psychometric structural patterns across the English, Chinese, and Turkish versions of the BFI-2 instrument. For each model family and size, we compute pairwise similarity scores between the language specific profiles derived from the MTDT process. Higher similarity scores indicate that the model's psychometric structural patterns remain robust despite the change in the language.

3.2 Traditional Method

3.2.1 Evaluation Framework. To evaluate the psychometric structure and behavior of Large Language Models (LLMs) across both respondent type conditions (human vs. LLM) and linguistic contexts, we adopt a two dimensional methodology. This framework is designed to determine if an LLM (i) reproduces human typical psychometric structure and behavior in a specified language ($H_{\text{src}}, L_{\text{src}}$) and (ii) maintains psychometric structure and behavior consistency across languages ($L_{\text{src}}, L_{\text{tgt}}$). Following recent advancements in LLM psychometrics and alignment evaluation [**? ?**], we conduct our evaluation through two primary analytical dimensions: structural similarity and behavioral alignment.

3.2.2 Models. We evaluated a selection of three models using our traditional method pipeline across four different languages, English, Spanish, Turkish, and Chinese. These models were GPT 5.1 from OpenAI, DeepSeek-V3.2, and Mistral ?.

3.2.3 Structural Similarity. We adopt three complementary structural analyses as proposed by [**?**] to characterize the model's internal representation of psychological constructs. First, we construct factor correlation based psychometric "fingerprints'" from item level correlation patterns, which we subsequently use to quantify structural similarity across conditions. Second, to test whether the LLM recovers the intended five factor latent structure of the BFI-2 across conditions, we apply Exploratory Graph Analysis (EGA) to recover item communities and compare the number of recovered communities against the theoretical number of BFI-2 factors. Third, as an additional psychometric consistency check, we compute Cronbach's Alpha for each BFI-2 subscale to assess internal consistency within factors. Together, these analyses assess whether the latent psychological structure generated by LLM responses align with human typical psychological structure and whether they remain stable across linguistic contexts.

Fingerprinting. Following [**?**], we define a model's psychometric "fingerprint" as the correlation structure among responses of questionnaire items, operationalized as a correlation matrix capturing pairwise relationships across all items in the collected responses. This correlation pattern is treated as a distinctive signature of psychological construct.

For each dataset $\mathcal{D}$, defined by the respondent type (LLM vs. human) and the language of the used questionnaire (source vs. target), we consider $\mathcal{D}^{\text{LLM}}_{\text{src}}$ and $\mathcal{D}^{\text{LLM}}_{\text{tgt}}$ obtained by presenting the source and target language instruments to the LLM, and $\mathcal{D}^{\text{Human}}_{\text{src}}$ obtained by presenting the source language instrument to human respondents. Given each dataset $\mathcal{D}$, we then compute a $Q\times Q$ Pearson correlation matrix $\mathbf{C}_x$, where $Q = 60$ for the BFI-2. We extract the off-diagonal upper triangular elements to form a vectorized fingerprint $\vec{\mathbf{C}}_x$:

$$\vec{\mathbf{C}}_x = [c_{1,2}, c_{1,3}, \ldots, c_{Q-1,Q}]^\top \tag{4}$$

We then adopt the similarity comparison method used in [**?**], quantifying the similarity between two vectorized fingerprints $\vec{\mathbf{C}}_{x_1}$ and $\vec{\mathbf{C}}_{x_2}$ using cosine similarity [**?**].

Exploratory Graph Analysis. To test whether the LLM recovers the intended five factor structure of the BFI-2 across conditions, we apply Exploratory Graph Analysis (EGA) to correlation matrix. Specifically, we follow the EGA procedure adopted by [**?**], which builds on [**?**]. The procedure involves estimating a network via graphical LASSO algorithm with EBIC model selection [**?**] and then use Walktrap community detection algorithm [**?**] to identify item communities, which are interpreted as recovered latent dimensions. We apply this procedure to the BFI-2 ($Q = 60$ items; five factors) and evaluate dimensionality recovery by comparing the number of detected communities to the theoretical five factor structure.

Subscale Consistency. Following the methodology proposed by [**?**], we compute Cronbach's alpha for each BFI-2 subscale as a

complementary psychometric consistency assessment. This established metric evaluates the internal consistency and reliability of responses within a given subscale.

3.2.4 Behavioral Alignment. The second dimension assesses behavioral alignment. Whereas structural similarity targets the model's internal psychometric structure, behavioral alignment evaluates surface level response behavior. Building on the methodology proposed by [**?**], we assess the behavioral alignment by computing the distance between response distributions. Specifically, we calculate the 1-Wasserstein distance between response distributions across conditions, quantifying how closely LLM behaviors mirror human baselines or preserve consistency across translations.

1-Wasserstein Distance. For each item q, we represent the Likert responses in dataset D as an ordinal distribution $p_D(r \mid q)$ over response categories $r \in \{1, \ldots, K\}$, where $p_D(r = i \mid q)$ denotes the probability of selecting category i for item q in D. Following [**?**], we use the 1-Wasserstein distance (W_1) to quantify the distributional distance between item level response distributions from datasets D_1 and D_2. This metric is appropriate for Likert data because it respects the ordinal structure of the scale by comparing cumulative probability mass along ordered categories. For an item q, the W_1 distance is computed as:

$$W_1(D_1, D_2; q) = \sum_{t=1}^{K-1} \left| \sum_{i=1}^{t} p_{D_1}(r = i \mid q) \; - \; \sum_{i=1}^{t} p_{D_2}(r = i \mid q) \right| \quad (5)$$

In our setting, the BFI-2 uses $K = 5$ response categories.

Alignment Scoring. We transform the 1-Wasserstein distance into a normalized alignment score $A(q) \in [0, 1]$, where 1 represents perfect distributional alignment:

$$A(q) = 1 - \frac{W_1(D_1, D_2; q)}{K - 1} \quad (6)$$

where $K = 5$.

Aggregation. We aggregate item level alignment scores $A(q)$ to obtain summary measures at two levels. Global alignment is defined as the mean alignment scores across all Q questionnaire items ($Q = 60$ for BFI-2):

$$A_{\text{global}} = \frac{1}{Q} \sum_{q=1}^{Q} A(q) \quad (7)$$

In addition, we report subscale level alignment by averaging alignment scores within each BFI-2 subscale. Let Q_s denote the set of items assigned to subscale s, then:

$$A_s = \frac{1}{Q_s} \sum_{q=1}^{Q_s} A(q) \quad (8)$$

4 Experiments

4.1 New Method

From our MTDT method testing across four model families, ten total model variations (at various sizes), and across three languages, we acquired over 430 pairwise similarity comparisons between the models to assess the cross-language robustness of the varying model's psychometric patterns. Further, we produced a sankey diagram for each model to provide visual clarity to its decision-making path and quickly observe overarching patterns. To these ends, our primary observation was the lack of a significant impact from language choice or model size on model similarity scoring and or sankey behavior. Indeed, regardless of language, the primary factor that appeared to most closely dictate model behavior outcome was model family, though this was also to a relative degree. Qwen models revealed a fairly high similarity scoring with one another across all languages and models sizes, as did the Ministral models, somewhat.

However, the results from the Qwen models' similarity scoring take on a new character when viewed within the context of their sankey diagrams. In this domain, almost all the Qwen models produced what we deem to be invariant results, that is to say, the models overwhelmingly selected the same answer path across every question, leading to very unchanging results.

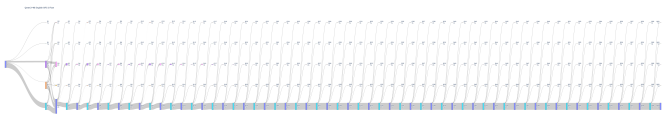

Figure 1: Example results from one of the Qwen sankey diagrams, revealing highly invariant answer choices

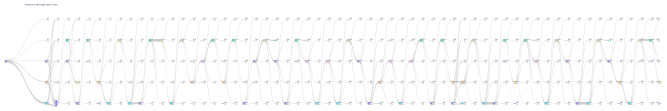

Figure 2: Example results from one of the Ministral sankey diagrams, revealing comparatively varying answer choices

Regardless of the answer invariance concern with most of the Qwen models (and some of the Llama models), a larger problem also stands out amidst even many of the ostensibly valid results wherein model output simply does not effectively operationalize BFI-2 psychometric constructs. In order to show alignment with BFI-2 constructs such as "Extraversion," or "Agreeableness," it is not enough that a model consistently (but not invaryingly) answers with values between 3 - 5 (indicating a range of neutral to strong agreement) because of the BFI-2's various reverse-coded statements. High positive alignment with an attribute such as "Extraversion," is not simply a matter of indicating strong agreement with statements like "I am someone who is outgoing, sociable," but also one of indicating strong disagreement with reverse metric statements like "I am someone who rarely feels excited or eager." What we find instead is that many models selected paths that demonstrated alignment with metrics in traditional questions, but did not produce paths that accounted for reverse statements, producing an overall inconsistency that undermines the validity of their modeling of actual psychometric constructs.

Of the models that remain, the larger Ministral models (8B and 14B) showed the most consistency overall at demonstrating robust psychometric constructs, with some minor differences between languages, namely a more consistent performance in English and Chinese than Turkish.

Table 1: Domain Alignment Between Languages

Model	Domain	EN–ZH	EN–ES	EN–TR	ES–TR	ES–ZH
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

4.2 Traditional Method

Using our traditional LLM querying pipeline, we prompted across three models and four languages over a hundred sample surveys and acquired a plentitude of various alignment parameters comparing model performance across BFI-2 psychometric domains and between models and baseline human BFI-2 data acquired for each language. Our results, however, also revealed a familiar problem that presented us with the difficult reality of being unable to utilize all evaluation metrics we had anticipated. Like with some of the models tested using the new (MTDT) method, we collected data that was by and large, highly invariant. Accordingly, we were unable to produce effective model fingerprints or EGA charts (except for Mistral), though we were able to assess the behavioral alignment of our tested models, both between model outputs in different languages and against human baselines. We further collected cronbach's alpha data as a measure of internal consistency for GPT and Mistral.

4.2.1 LLMs and Human Alignment. Though we were not able to recreate the same fingerprinting technique used by [**?**] due to reasons of data invariancy, we were able to still observe a similar conclusion based on our analysis of model alignment against our human benchmark data across our selection of languages. Therein, no model gave a particularly strong alignment with any specific BFI-2 domain (Agreeableness, Conscientiousness, etc.) across all languages, nor did any model show especially strong alignment with any of the human answer sets in any language, though some were closer than others. These observations appear consistent with [**?**] on the basis that while these models simulate a human-like set of responses, their simulation cannot be said to be a completely accurate stand-in for true human-generated data.

4.2.2 Alignment Across Models. 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 were GPT 5.1 and DeepSeek V3.2, with little to remark on in the way of relationships between the other models.

4.2.3 Alignment Across Languages. 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.

4.2.4 Cronbach's Alpha. From our cronbach's alpha analysis, GPT demonstrates fairly strong internal consistency metrics across all languages, with some degree of fluctuation in Spanish (particularly

Table 2: Language Alignment Between Models

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

in the Negative Emotionality domain) and in Turkish generally. Mistral, by contrast, exhibits a highly irregular set of cronbach's alpha metrics across all languages, with its strongest values representing domains in Spanish. Unfortunately, due to the highly invariant nature of the results from DeepSeek in particular, we were unable to conduct cronbach's alpha testing on our data from DeepSeek.

5 Discussion

Our experiments expose some very important realities concerning the abilities of LLMs to participate in testing for psychometric constructs as simulacra for human populations. Our first conclusion to be drawn concerning model outputs is the primacy of model choice/family origin over prompted language or, at least insofar as the constructs of the BFI-2 are concerned, which domain was analyzed. Put simply, the primary indicator of end model behavior, whether results would be psychometrically consistent, both through the new and traditional method pipelines, was the model being tested. This is not to say that the latter attributes bore no effect on the outcomes we observed, however. Language choice did appear to present an effect on model outcomes in a variety of situations. For example, in the assessment of model alignment across languages, 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, as a model of French origin, Mistral showed the most alignment between English and Chinese. In the opposite case, 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 psychometric profile 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.

Our secondary key conclusion is simply a reflection of the great many issues we experienced over the course of our experiments with model invariance. Between both our new method and traditional method testing, the majority of our results exhibited the invariance problem to some degree or another, a fact that we postulate may be a shortcoming of LLM simulation of psychometric constructs more generally. While some models such as the Mistral family (Mistral ? in traditional testing and Ministral 8B and 14B in new method testing) generally showed themselves to provide more consistently psychometrically robust output data, the sheer prevalence of the problem across so many other models and under

different methods demands us to urge a great deal of caution when attempting to simulate human psychometric constructs through LLM testing.

6 Conclusion