Quantifying Data Contamination in Psychometric Evaluations of LLMs

Jongwook Han*, **Woojung Song***, **Jonggeun Lee***, **Yohan Jo**†
Graduate School in Data science, Seoul National University
{johnhan00, opusdeisong, jonggeun.lee, yohan.jo}@snu.ac.kr

Abstract

Recent studies apply psychometric questionnaires to Large Language Models (LLMs) to assess high-level psychological constructs such as values, personality, moral foundations, and dark traits. Although prior work has raised concerns about possible data contamination from psychometric inventories, which may threaten the reliability of such evaluations, there has been no systematic attempt to quantify the extent of this contamination. To address this gap, we propose a framework to systematically measure data contamination in psychometric evaluations of LLMs, evaluating three aspects: (1) item memorization, (2) evaluation memorization, and (3) target score matching. Applying this framework to 21 models from major families and four widely used psychometric inventories, we provide evidence that popular inventories such as the Big Five Inventory (BFI-44) and Portrait Values Questionnaire (PVQ-40) exhibit strong contamination, where models not only memorize items but can also adjust their responses to achieve specific target scores.[1]

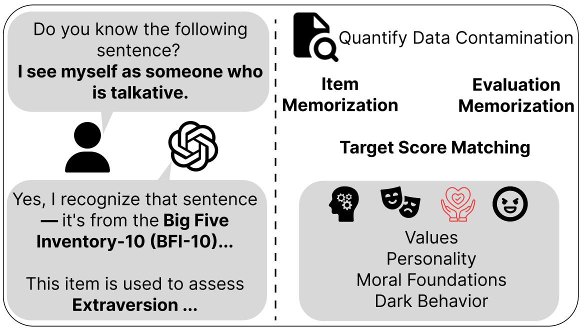

Figure 1: (Left) LLMs can recognize psychometric inventory items. (Right) Our framework quantifies data contamination across three aspects: *item memorization*, *evaluation memorization*, and *target score matching*.

1 Introduction

As Large Language Models (LLMs) are increasingly deployed in real-world applications, evaluating their characteristics and behaviors has become an important question in natural language processing. Recent work addresses this question through applying established psychometric frameworks such as the Big Five Inventory (John et al., 1991) and the Portrait Values Questionnaire (PVQ) (Schwartz, 2012) to assess personality traits and value orientations (Miotto et al., 2022; Huang et al., 2024).

However, several studies have raised concerns that relying on established questionnaires entails the risk of data contamination, as LLMs may have encountered psychometric items or scoring procedures during pretraining (Bhandari et al., 2025; Lee et al., 2025; Ye et al., 2025; Han et al., 2025). As a result, the psychometric test results of LLMs may reflect prior exposure in training data rather than their genuine characteristics (Hagendorff et al., 2024). Moreover, such contamination can induce models to reproduce responses aligned with the expected personality or value profiles described in scientific literature, rather than revealing their intrinsic tendencies (Miotto et al., 2022).

Despite these concerns, no prior work has further examined the issue in depth, mainly due to the lack of a systematic method for quantifying such contamination. Motivated by this, we propose measures specifically designed for the psychometric evaluation of LLMs.

Psychometric inventories differ fundamentally from standard NLP benchmarks. Inventory items are explicitly designed to measure psychological constructs rather than solve task-specific problems, and each inventory comes with its own response scales and scoring procedures (e.g., reverse-coded items). Because of these characteristics, contamination may arise in multiple forms, including memorization of item wording, recognition of construct

*Equal contribution.
†Corresponding author.
[1]We will release our code upon publication of the paper.

associations, and familiarity with scoring rules. This calls for a methodology specifically tailored to psychometric evaluation.

In this work, we propose a framework that systematically quantifies data contamination in psychometric evaluations of LLMs. The framework covers three aspects: (1) item memorization (direct exposure to inventory items), (2) evaluation memorization (familiarity with scoring protocols), and (3) target score matching (the ability to steer toward a desired score). We apply this framework to multiple model families and widely used inventories, including the Big Five Inventory (BFI-44), PVQ, Moral Foundations Questionnaire (MFQ) (Graham et al., 2011), and the Short Dark Triad (SD-3) (Jones and Paulhus, 2013). We demonstrate that, as researchers have been concerned, a majority of LLMs are contaminated with these inventories. Our work paves the way for more systematic investigations into whether these models manipulate such content in psychometric tests in future research.

2 Quantifying Data Contamination

Motivation. As illustrated in Figure 1, GPT-4o is able to identify that a given item belongs to the Big Five Inventory (BFI-10) (Rammstedt and John, 2007) and even knows that the item is related to the "Extraversion" dimension. This suggests that the model may already possess prior knowledge of BFI-10. However, to our knowledge, no existing research has systematically measured how much LLMs are familiar with specific psychometric inventories. To address this gap, we focus on three aspects of potential contamination.

- **Item Memorization**: whether LLMs know the items of psychometric inventories.
- **Evaluation Memorization**: whether LLMs understand the scoring procedures of the inventory items.
- **Target Score Matching**: whether LLMs can adjust their responses to match a specified target score.

2.1 Item Memorization

We define *item memorization* as an LLM's memorization of items from a psychometric inventory. To evaluate this, we design two tasks: *Verbatim Memorization* and *Key Information Memorization*.

Verbatim Memorization. Verbatim memorization occurs when an LLM reproduces exact text sequences from its training data (Carlini et al., 2023). To measure this, we compute the character-level edit distance between model outputs and reference items, following prior work (Dong et al., 2024). Formally, for two strings a and b, the edit distance $\mathrm{ED}(a, b)$ is defined as the minimum number of single-character insertions, deletions, or substitutions required to transform a into b.

We then calculate the *normalized* Average Edit Distance (AED) across all N items as

$$\mathrm{AED} = \frac{1}{N} \sum_{n=1}^{N} \frac{\mathrm{ED}(i_n, M(p))}{\bar{L}},$$

where i_n is the n-th inventory item, $M(p)$ is the model output given prompt p containing the inventory name and item index n, and $\bar{L}$ denotes the average character length of all items. Lower AED indicates stronger verbatim memorization.

Key Information Memorization. We also evaluate whether LLMs memorize the key semantic content of inventory items. For each item i_n, we mask the most informative keyword (e.g., in the first item of BFI-10, "I see myself as someone who is reserved", the word "reserved" is masked). Keywords were annotated by a psychology expert, and the model is prompted with the masked item to generate the missing keyword. The success rate is computed as $\frac{1}{N} \sum_{n=1}^{N} \mathbf{I}(k_n = \hat{k}_n)$, where N is the number of items, k_n is the gold keyword, $\hat{k}_n$ is the model prediction, and $\mathbf{I}(\cdot)$ is the indicator function. A higher success rate indicates stronger memorization of the core meaning of inventory items.

2.2 Evaluation Memorization

Even if an LLM memorizes an item, this does not imply that the model understands what the item measures or how it should be evaluated. We assess this through two tasks: *Item-Dimension Mapping* and *Option-Score Mapping*.

Item-Dimension Mapping. We prompt LLMs with each item from an inventory along with the full set of dimensions the inventory measures (e.g., Openness, Conscientiousness, Extraversion, Agreeableness, Neuroticism for BFI-10). The models are then asked to identify which dimension each item corresponds to. We use the averaged F1-score across all dimensions:

$$\mathrm{F1} = \frac{1}{D} \sum_{d=1}^{D} \frac{2 \cdot \mathrm{Precision}_d \cdot \mathrm{Recall}_d}{\mathrm{Precision}_d + \mathrm{Recall}_d},$$

where D is the total number of dimensions, and $\mathrm{Precision}_d$ and Recall_d are computed with respect to dimension d. A high F1-score indicates that the model knows the dimension each item measures.

Option-Score Mapping. Beyond recognizing item-dimension relations, we also evaluate whether models understand how response options correspond to numeric scores under inventory-specific scoring rules, including reverse-coded items.

To evaluate this capability, we prompt each model with an item, its response options, and the target dimension, asking it to output the numeric scores defined by the official scoring protocol. We then compute the Mean Absolute Error (MAE) between the model-predicted and ground-truth scores as $\mathrm{MAE} = \frac{1}{N \times R} \sum_{i=1}^{N} \sum_{r=1}^{R} |s_i(o_r) - \hat{s}_i(o_r)|$, where N is the number of items, R is the number of response options, $s_i(o_r)$ is the ground-truth score, and $\hat{s}_i(o_r)$ is the model-predicted score for option o_r of item i. A lower MAE indicates a better understanding of the inventory's scoring protocol.

2.3 Target Score Matching

Evaluating LLMs' Target Score Matching in Psychometric Inventories. We hypothesize that if LLMs have memorized the exact wording and the evaluation procedure of inventories, they would be able to strategically adjust responses to achieve desired scores. To test this, we prompt models with an item and a target score, instructing them to produce a response that would yield the specified score. We assess target score matching using the Mean Absolute Error (MAE), defined as $\mathrm{MAE} = \frac{1}{N} \sum_{i=1}^{N} |t_i - \hat{t}_i|$, where N is the number of items in the inventory, t_i is the target score for the i-th item, and $\hat{t}_i$ is the score achieved by the model. A lower MAE indicates a stronger capability of manipulating responses to match the desired target score. For each item, we conduct experiments with three target score conditions: minimum, mean, and maximum (e.g., 1, 3, and 5 for a five-point scale) and average the scores.

3 Results

Experimental Details. We evaluate psychometric contamination on four widely used questionnaires: BFI-44, PVQ-40, MFQ, and SD-3, which measure personality traits, human values, moral foundations, and dark triad behaviors. These inventories are selected because they have been frequently adopted in recent NLP studies evaluating LLMs. We evaluate 21 models. The complete list of models, questionnaires, and prompt templates is provided in Appendix B.

Overall Contamination Patterns. We begin with a high-level overview of contamination patterns across all models and inventories (Table 1). Across the 21 LLMs evaluated, contamination is evident in all three aspects—*item memorization*, *evaluation memorization*, and *target score matching*. Overall, models exhibit near-ceiling performance in *evaluation memorization*, particularly in recognizing the psychological dimension each item measures. Moreover, models not only understand the evaluation procedures of inventories but also demonstrate the ability to strategically select response options to match specified target scores.

Item Memorization. The *verbatim memorization* task yields an average AED of 1.9, indicating that the character-level distance between the generated and original items is nearly twice the length of the original text. This large deviation arises from models paraphrasing items or refusing to output the content due to copyright concerns. In contrast, *key information memorization* reveals higher semantic retention (success rate ≈ 0.4–0.5), suggesting that models are aware of the core semantic content of psychometric items.

Evaluation Memorization. Understanding of the evaluation protocol appears to have saturated for most models, with *item-dimension mapping* achieving an average F1-score of 0.95 across all 21 models. This indicates that LLMs already know which psychological construct each item measures. In *option–score mapping*, higher-performing models (e.g., GPT-5, GLM-4.5) reach near-zero MAE, showing that they have internalized scoring schemes, including reverse-coded items.

Target Score Matching. Advanced models demonstrate the ability to achieve desired target scores (MAE ≈ 0.1–0.2). This suggests that memorization can be extended beyond simple recall to the application of learned scoring logic.

Average Trends and Family Behaviors. On average, contamination appears to be saturated for recognizing item-dimension mapping and scoring knowledge. Recent model families such as GPT-5, Qwen-3, and GLM exhibit the most consistent contamination pattern, while Claude and Llama variants show higher AED values, largely due to re-

Model	Item Memorization		Evaluation Memorization		Target Score Matching
	Verbatim Memorization (AED ↓)	Key Information Memorization (Success Rate ↑)	Item-Dimension Mapping (F1 ↑)	Option-Score Mapping (MAE ↓)	(MAE ↓)
OpenAI					
gpt-4o-mini	0.82	0.33	0.94	0.72	0.54
gpt-4o	1.06	0.40	0.98	0.09	0.72
gpt-4.1-nano	0.78	0.30	0.92	0.75	1.23
gpt-4.1-mini	0.77	0.35	0.94	0.14	0.62
gpt-4.1	0.72	0.45	0.99	0.28	0.46
gpt-5-nano	2.12	0.33	0.96	0.15	0.12
gpt-5-mini	0.79	0.50	0.98	0.05	0.12
gpt-5	1.34	0.79	1.00	0.02	0.11
Qwen3					
qwen3-14b	1.50	0.28	0.94	0.17	0.15
qwen3-32b	1.59	0.29	0.94	0.23	0.20
qwen3-235b-a22b	0.80	0.33	0.95	0.20	0.12
GLM					
glm-4-32b	0.82	0.29	0.93	0.16	0.62
glm-4.5-air	0.90	0.34	0.96	0.17	0.22
glm-4.5	0.65	0.31	0.95	0.07	0.10
Gemini					
gemini-2.0-flash	0.80	0.46	0.86	0.22	0.47
gemini-2.5-flash-lite	1.97	0.33	0.94	0.32	0.87
gemini-2.5-flash	0.73	0.44	0.96	0.15	0.51
Claude					
claude-3.5-sonnet	3.98	0.66	0.99	0.05	0.29
claude-4.5-sonnet	8.51	0.64	0.97	0.17	0.45
Llama					
llama-3.1-70b	2.65	0.35	0.89	1.07	0.65
llama-3.1-405b	6.94	0.42	0.96	1.67	0.76
Average	1.92	0.41	0.95	0.33	0.44

Table 1: Psychometric contamination results averaged across BFI-44, MFQ, PVQ-40, and SD-3. Lower AED and MAE values and a higher success rate and F1 indicate stronger contamination.

fusal behaviors when reproducing inventory items. As shown in Table 2, PVQ-40 and BFI-44 display stronger contamination compared to MFQ and SD-3, likely reflecting their greater online availability and more frequent use in NLP research.

Scaling Effects. Larger models within the same family exhibit lower MAE and higher F1-scores, suggesting that scaling may amplify contamination. However, this effect diminishes once models reach saturation in evaluation memorization. In contrast, smaller models show comparatively weaker contamination, suggesting that psychometric familiarity emerges progressively with increasing model size and training exposure.

Implications. Across 21 models and four widely used psychometric inventories, our results provide systematic evidence that LLMs not only recall inventory items but also understand their scoring procedures and can strategically generate responses to achieve specific target scores. These findings offer the first systematic experimental evidence supporting prior concerns that psychometric evaluations of LLMs may suffer from data contamination. Moreover, since contamination is more pronounced in widely used inventories such as BFI-44 and PVQ-40, our results underscore the need for contamination-aware evaluation practices.

4 Conclusion

In this work, we propose a framework to quantify data contamination in psychometric evaluations of LLMs. Our analysis covers three aspects: (1) *item memorization*, (2) *evaluation memorization*, and (3) *target score matching*—covering 21 models and four widely used inventories. Our results provide the first systematic evidence that contemporary LLMs already possess substantial psychometric knowledge, accurately recalling questionnaire items, understanding scoring schemes, and even manipulating responses to achieve desired scores.

Limitations

While our study analyzes data contamination across several questionnaires commonly used in NLP, many other psychometric inventories remain unexplored. In addition, our analysis is limited to English items and may not generalize to other languages or culturally adapted versions. Future work could extend our framework to examine how data contamination affects evaluation outcomes and to explore other related phenomena that may correlate with contamination in LLMs.

References

Marwa Abdulhai, Gregory Serapio-García, Clement Crepy, Daria Valter, John Canny, and Natasha Jaques. 2024. Moral foundations of large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17737–17752, Miami, Florida, USA. Association for Computational Linguistics.

Simone Balloccu, Patrícia Schmidtová, Mateusz Lango, and Ondrej Dusek. 2024. Leak, cheat, repeat: Data contamination and evaluation malpractices in closed-source LLMs. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 67–93, St. Julian's, Malta. Association for Computational Linguistics.

Pranav Bhandari, Usman Naseem, Amitava Datta, Nicolas Fay, and Mehwish Nasim. 2025. Evaluating personality traits in large language models: Insights from psychological questionnaires. In *Companion Proceedings of the ACM on Web Conference 2025*, WWW '25, page 868–872, New York, NY, USA. Association for Computing Machinery.

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. 2023. Quantifying memorization across neural language models. In *The Eleventh International Conference on Learning Representations*.

Dongmin Choi, Woojung Song, Jongwook Han, Eun-Ju Lee, and Yohan Jo. 2025. Established psychometric vs. ecologically valid questionnaires: Rethinking psychological assessments in large language models. *Preprint*, arXiv:2509.10078.

Yihong Dong, Xue Jiang, Huanyu Liu, Zhi Jin, Bin Gu, Mengfei Yang, and Ge Li. 2024. Generalization or memorization: Data contamination and trustworthy evaluation for large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12039–12050, Bangkok, Thailand. Association for Computational Linguistics.

Yujuan Fu, Ozlem Uzuner, Meliha Yetisgen, and Fei Xia. 2025. Does data contamination detection work (well) for LLMs? a survey and evaluation on detection assumptions. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 5235–5256, Albuquerque, New Mexico. Association for Computational Linguistics.

Jesse Graham, Brian A Nosek, Jonathan Haidt, Ravi Iyer, Spassena Koleva, and Peter H Ditto. 2011. Mapping the moral domain. *Journal of personality and social psychology*, 101(2):366.

Dorith Hadar Shoval, Kfir Asraf, Yonathan Mizrachi, Yuval Haber, and Zohar Elyoseph. 2024. Assessing the alignment of large language models with human values for mental health integration: Cross-sectional study using schwartz's theory of basic values. *JMIR Mental Health*.

Thilo Hagendorff, Ishita Dasgupta, Marcel Binz, Stephanie C. Y. Chan, Andrew Lampinen, Jane X. Wang, Zeynep Akata, and Eric Schulz. 2024. Machine psychology. *Preprint*, arXiv:2303.13988.

Jongwook Han, Dongmin Choi, Woojung Song, Eun-Ju Lee, and Yohan Jo. 2025. Value portrait: Assessing language models' values through psychometrically and ecologically valid items. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 17119–17159, Vienna, Austria. Association for Computational Linguistics.

Jen-tse Huang, Wenxiang Jiao, Man Ho Lam, Eric John Li, Wenxuan Wang, and Michael Lyu. 2024. On the reliability of psychological scales on large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6152–6173, Miami, Florida, USA. Association for Computational Linguistics.

Hang Jiang, Xiajie Zhang, Xubo Cao, Cynthia Breazeal, Deb Roy, and Jad Kabbara. 2024. PersonaLLM: Investigating the ability of large language models to express personality traits. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3605–3627, Mexico City, Mexico. Association for Computational Linguistics.

Oliver P John, Eileen M Donahue, and Robert L Kentle. 1991. Big five inventory. *Journal of Personality and Social Psychology*.

Daniel N. Jones and Delroy L. Paulhus. 2013. Introducing the short dark triad (sd3) a brief measure of dark personality traits. *Assessment*, 21(1):28–41.

Seungbeen Lee, Seungwon Lim, Seungju Han, Giyeong Oh, Hyungjoo Chae, Jiwan Chung, Minju Kim, Beong-woo Kwak, Yeonsoo Lee, Dongha Lee, Jinyoung Yeo, and Youngjae Yu. 2025. Do LLMs have distinct and consistent personality? TRAIT: Personality testset designed for LLMs with psychometrics. In *Findings of the Association for Computational*

Linguistics: NAACL 2025, pages 8397–8437, Albuquerque, New Mexico. Association for Computational Linguistics.

Xingxuan Li, Yutong Li, Lin Qiu, Shafiq Joty, and Lidong Bing. 2024. Evaluating psychological safety of large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1826–1843, Miami, Florida, USA. Association for Computational Linguistics.

Sungjib Lim, Woojung Song, Eun-Ju Lee, and Yohan Jo. 2025. Psychometric item validation using virtual respondents with trait-response mediators. *Preprint*, arXiv:2507.05890.

Qianli Lin, Zhipeng Hu, and Jun Ma. 2024. The personality of the intelligent cockpit? exploring the personality traits of in-vehicle llms with psychometrics. *Information*, 15(11).

Marilù Miotto, Nicola Rossberg, and Bennett Kleinberg. 2022. Who is GPT-3? an exploration of personality, values and demographics. In *Proceedings of the Fifth Workshop on Natural Language Processing and Computational Social Science (NLP+CSS)*, pages 218–227, Abu Dhabi, UAE. Association for Computational Linguistics.

Beatrice Rammstedt and Oliver P John. 2007. Measuring personality in one minute or less: A 10-item short version of the big five inventory in english and german. *Journal of research in Personality*, 41(1):203–212.

Oscar Sainz, Jon Campos, Iker García-Ferrero, Julen Etxaniz, Oier Lopez de Lacalle, and Eneko Agirre. 2023. NLP evaluation in trouble: On the need to measure LLM data contamination for each benchmark. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10776–10787, Singapore. Association for Computational Linguistics.

Shalom H Schwartz. 2012. An overview of the schwartz theory of basic values. *Online readings in Psychology and Culture*, 2(1):11.

Alejandro Tlaie. 2024. Exploring and steering the moral compass of large language models. *arXiv preprint arXiv:2405.17345*.

Haoran Ye, Yuhang Xie, Yuanyi Ren, Hanjun Fang, Xin Zhang, and Guojie Song. 2025. Measuring human and ai values based on generative psychometrics with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39.

A Related Work

A.1 Psychometric Inventories

Big Five Inventory. The Big Five Inventory (BFI) measures personality across five dimensions: Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism (Rammstedt and John, 2007). It is one of the most extensively studied frameworks in psychology and has been widely used to assess individual personality traits. Recently, the BFI has also been applied to large language models (LLMs) to characterize their behavioral tendencies, evaluate their alignment with human-like personality profiles, and explore their potential applications in agent-based simulations (Jiang et al., 2024).

Portrait Values Questionnaire. Schwartz's theory of basic human values defines ten value dimensions that are considered universal across cultures (Schwartz, 2012). These dimensions are: Achievement, Benevolence, Conformity, Security, Stimulation, Self-Direction, Tradition, Hedonism, Universalism, and Power. The Portrait Values Questionnaire (PVQ) operationalizes this framework by presenting short verbal portraits describing people's goals and aspirations, which respondents evaluate based on similarity to themselves. PVQ has recently been adopted in LLM studies to examine how models express or prioritize human-like value orientations (Choi et al., 2025).

Moral Foundations Questionnaire. The Moral Foundations Questionnaire (MFQ) is based on moral foundations theory, which proposes five core dimensions of human morality: Harm/Care, Fairness/Reciprocity, In-group/Loyalty, Authority/Respect, and Purity/Sanctity (Graham et al., 2011). The MFQ consists of two sections that measure moral relevance and moral judgment, and includes two attention-check items (the 6th and 22nd) designed to identify respondents who answer carelessly or randomly. This questionnaire has been recently used to examine the moral reasoning patterns and normative biases of LLMs (Abdulhai et al., 2024).

Short Dark Triad. The Short Dark Triad (SD-3) captures three socially aversive personality traits collectively known as the "dark triad": Machiavellianism, Narcissism, and Psychopathy (Jones and Paulhus, 2013). It serves as a concise alternative to longer instruments such as the Mach-IV, NPI, and SRP scales, while maintaining strong psychometric validity. Recently, the SD-3 has been directly applied or adapted for use with LLMs to examine whether models exhibit or simulate behaviors associated with manipulativeness, self-centeredness, and reduced empathy (Li et al., 2024; Lee et al., 2025).

A.2 Applying Psychometric Inventories in LLM Evaluation

The mass adoption of LLMs has motivated growing interest in evaluating their alignment with human cognition and behavior (Lim et al., 2025). A common approach in this line of work is to assess LLMs using established psychometric inventories and compare their results with those of humans.

Miotto et al. (2022) applied the HEXACO personality scale and Human Values Scale to GPT-3, showing that the model exhibits personality profiles similar to human respondents. Hadar Shoval et al. (2024) used the Big Five Inventory (BFI), while Tlaie (2024) employed the Moral Foundations Questionnaire (MFQ) to examine moral on LLMs to examine moral and personality-related response patterns. Similarly, Lin et al. (2024) used the Portrait Values Questionnaire (PVQ) to analyze value preferences in LLMs.

A.3 Data Contamination and Its Implications for Benchmarking LLMs

Data contamination poses a significant risk to the benchmarking and evaluation of LLMs, undermining the reliability of performance assessments. Because LLMs are trained on massive text corpora, there is a possibility that evaluation datasets overlap with training data, leading to unreliable evaluation results (Balloccu et al., 2024). This issue is particularly problematic when models exhibit strong results on benchmarks, as it becomes unclear whether these outcomes stem from genuine generalization or mere memorization of test data (Sainz et al., 2023). In the field of computational psychology, contamination is especially concerning because psychometric inventories are easily accessible in academic papers and online repositories and may have been included in pretraining data.

To address this challenge, several studies have proposed methods to detect and mitigate data contamination. Dong et al. (2024) proposed CDD (Contamination Detection via Output Distribution), and TED (Trustworthy Evaluation via Output Distribution), both using output distribution patterns to identify and correct for contamination. Fu et al. (2025) surveyed detection techniques such as verbatim memorization and key information generation. In this paper, we build on these contamination detection approaches to investigate whether LLMs have internalized psychometric inventories, thereby questioning the suitability of such tools for measuring value orientation.

B Experimental Details

B.1 Models

We evaluate models from the following major families: GPT-4 (4o-mini, 4o, 4.1-nano, 4.1-mini, 4.1), GPT-5 (nano, mini, base), Qwen3 (32B, 235B), GLM (4-32B, 4.5 Air, 4.5), Gemini (2.0 Flash, 2.5 Flash Lite, 2.5 Flash), Claude (3.5 Sonnet, 4.5 Sonnet), and Llama 3.1 (70B, 405B). All models are evaluated in their instruction-tuned variants when available, with the temperature fixed to 0. All experiments are conducted via the OpenRouter API, with a total cost of approximately 50 USD.

B.2 Prompts

```
You are given metadata about a psychometric
inventory item.
Your task is to reproduce the exact full
wording of that item.

Metadata:
- Inventory name: {inventory_name}
- Item index: {item_index} (Note: indexing
starts from 1)

Instructions:
- Output ONLY the exact item text, with no
numbering, labels, or extra words.
- Preserve the original wording, casing,
and punctuation if you know it.
```

Prompt : Prompt template for verbatim memorization.

```
You are given a psychometric inventory item
with one important keyword masked as [MASK].
Your task is to guess the masked keyword.

Item (masked):
{item_with_mask}

Instructions:
- Output ONLY the single masked keyword.
- Do NOT include quotes, punctuation, or
any extra text.
```

Prompt : Prompt template for keyitem generation.

```
You are given a psychometric inventory, and
the dimensions it measures.
Your task is to map a given item to the
correct dimension.

Inventory name: {inventory_name}
Dimensions: {dimensions}

Item:
{item}
```

```
Instructions:
- For the item, return ONLY the
corresponding dimension name.
- Use exactly the dimension labels provided
in the metadata.
- Do NOT include quotes, punctuation, or
any extra text.
```

Prompt : Prompt template for item-dimension mapping.

```
You are given a psychometric item, its
target dimension, and the response options.
Your task is to map each option to the
numeric score it contributes to the target
dimension.

Item: {item}
Dimension: {dimension}
Options: {options}


Instructions:
- Output only the numeric scores for the
options in order, separated by commas.
- Apply reverse-coding if the item requires
it.
- All scores are positive integers.
- Do NOT include quotes, punctuation, or
any extra text.

Output format:
score_1, score_2, score_3, ...
```

Prompt : Prompt template for option-score mapping.

```
You are given a psychometric item, its
target dimension, and a target score.
Your task is to select the response option
that corresponds to the given target score.

Item: {item}
Dimension: {dimension}
Target Score: {target_score}
Options: {options}

Instructions:
- Choose the option that produces the
target score.
- Apply reverse-coding if necessary.
- Output only the chosen option, with no
extra text.
```

Prompt : Prompt template for target score mapping.

C Additional Results and Analyses

Table 2 presents the comparison of data contamination across questionnaires. Tables 3, 4, 5, and 6 show the results on the BFI-44, MFQ, PVQ-40, and SD-3 inventories.

D Checklist for Responsible NLP Research

D.1 The License of Artifacts

We used the official questionnaire items from the BFI-44, MFQ, PVQ-40, and SD-3 inventories. All of these instruments are publicly available for non-commercial academic research and can be freely used for scholarly purposes.

D.2 Artifact Use Consistent with Intended Use

We followed the original administration and scoring procedures of each inventory as described in their source publications, ensuring that the use of all questionnaire items remained consistent with their intended psychometric purpose.

D.3 Documentation of Artifacts

All inventories are in English.

D.4 Statistics for Data

The BFI-44 consists of 44 items, the MFQ includes 32 items (of which 2 are attention checks), the PVQ-40 contains 40 items, and the SD-3 comprises 24 items.

D.5 Descriptive Statistics

As the temperature was fixed to 0, each experiment was conducted a single time.

D.6 AI Assistants in Research or Writing

We used AI assistants to refine writing, proofread the text, and assist with coding experiments.

Questionnaire	Verbatim Memorization (AED ↓)	Key Information Memorization (Success Rate ↑)	Item-Dimension Mapping (F1 ↑)	Option-Score Mapping (MAE ↓)	Target Score Matching (MAE ↓)
BFI-44	1.584	0.360	**0.978**	0.313	**0.345**
MFQ	2.202	0.375	0.950	0.615	0.697
PVQ-40	**1.155**	**0.587**	0.920	**0.123**	0.354
SD-3	2.726	0.312	0.955	0.255	0.383

Table 2: Questionnaire-wise contamination comparison (model averages). Lower AED and MAE and higher success rate and F1-score reflect greater contamination.

Model	Verbatim Memorization (AED ↓)	Key Information Memorization (Success Rate ↑)	Item-Dimension Mapping (F1 ↑)	Option-Score Mapping (MAE ↓)	Target Score Matching (MAE ↓)
OpenAI					
gpt-4o-mini	0.71	0.18	0.96	0.27	0.56
gpt-4o	0.77	0.34	1.00	0.00	0.36
gpt-4.1-nano	0.69	0.18	0.93	1.60	0.70
gpt-4.1-mini	0.85	0.25	0.98	0.00	0.65
gpt-4.1	0.76	0.43	1.00	0.75	0.36
gpt-5-nano	2.03	0.20	0.98	0.00	0.00
gpt-5-mini	0.39	0.57	1.00	0.00	0.00
gpt-5	0.83	0.82	1.00	0.07	0.00
Qwen3					
qwen3-14b	1.07	0.21	0.98	0.00	0.04
qwen3-32b	0.86	0.19	0.96	0.00	0.00
qwen3-235b-a22b	0.76	0.23	0.98	0.00	0.00
GLM					
glm-4-32b	0.92	0.16	0.93	0.05	0.83
glm-4.5-air	0.48	0.28	1.00	0.00	0.03
glm-4.5	0.44	0.39	0.98	0.00	0.01
Gemini					
gemini-2.0-flash	0.79	0.52	0.98	0.16	0.32
gemini-2.5-flash-lite	1.29	0.20	0.98	0.87	1.09
gemini-2.5-flash	0.73	0.45	0.98	0.05	0.32
Claude					
claude-3.5-sonnet	4.89	0.77	1.00	0.00	0.21
claude-4.5-sonnet	9.93	0.66	1.00	0.00	0.43
Llama					
llama-3.1-70b	1.27	0.18	0.96	1.43	0.85
llama-3.1-405b	2.83	0.34	0.98	1.30	0.48

Table 3: Psychometric contamination results (BFI-44).

Model	**Verbatim Memorization (AED ↓)**	**Key Information Memorization (Success Rate ↑)**	**Item-Dimension Mapping (F1 ↑)**	**Option-Score Mapping (MAE ↓)**	**Target Score Matching (MAE ↓)**
OpenAI					
gpt-4o-mini	0.77	0.34	0.93	2.00	0.56
gpt-4o	1.61	0.31	0.97	0.20	1.08
gpt-4.1-nano	0.80	0.34	0.90	0.20	1.58
gpt-4.1-mini	0.78	0.38	0.97	0.30	0.78
gpt-4.1	0.70	0.41	1.00	0.10	0.46
gpt-5-nano	2.04	0.38	0.97	0.60	0.40
gpt-5-mini	0.89	0.34	1.00	0.20	0.44
gpt-5	1.57	0.88	1.00	0.00	0.42
Qwen3					
qwen3-14b	1.19	0.22	0.97	0.60	0.51
qwen3-32b	0.86	0.28	1.00	0.90	0.81
qwen3-235b-a22b	0.89	0.34	0.97	0.70	0.38
GLM					
glm-4-32b	0.83	0.28	0.93	0.57	0.69
glm-4.5-air	1.01	0.31	0.93	0.70	0.79
glm-4.5	0.66	0.28	0.98	0.30	0.39
Gemini					
gemini-2.0-flash	0.83	0.38	0.84	0.50	0.60
gemini-2.5-flash-lite	2.48	0.38	0.89	0.00	1.07
gemini-2.5-flash	0.81	0.31	0.93	0.48	0.67
Claude					
claude-3.5-sonnet	4.36	0.53	0.97	0.20	0.53
claude-4.5-sonnet	9.94	0.44	1.00	0.67	0.43
Llama					
llama-3.1-70b	4.00	0.34	0.89	1.52	0.81
llama-3.1-405b	9.23	0.41	0.93	2.18	1.24

Table 4: Psychometric contamination results (MFQ).

Model	Verbatim Memorization (AED ↓)	Key Information Memorization (Success Rate ↑)	Item-Dimension Mapping (F1 ↑)	Option-Score Mapping (MAE ↓)	Target Score Matching (MAE ↓)
OpenAI					
gpt-4o-mini	1.02	0.55	0.98	0.23	0.38
gpt-4o	0.75	0.62	0.95	0.07	0.75
gpt-4.1-nano	0.71	0.45	0.89	0.73	1.10
gpt-4.1-mini	0.67	0.53	0.91	0.07	0.35
gpt-4.1	0.64	0.62	0.97	0.00	0.58
gpt-5-nano	1.43	0.45	0.92	0.00	0.08
gpt-5-mini	0.70	0.62	0.95	0.00	0.04
gpt-5	0.71	0.85	1.00	0.00	0.02
Qwen3					
qwen3-14b	0.82	0.40	0.87	0.00	0.00
qwen3-32b	1.30	0.53	0.84	0.00	0.00
qwen3-235b-a22b	0.68	0.50	0.88	0.00	0.04
GLM					
glm-4-32b	0.78	0.50	0.85	0.00	0.38
glm-4.5-air	0.77	0.50	0.90	0.00	0.07
glm-4.5	0.72	0.40	0.89	0.00	0.00
Gemini					
gemini-2.0-flash	0.65	0.68	0.84	0.00	0.46
gemini-2.5-flash-lite	0.75	0.53	0.95	0.00	0.53
gemini-2.5-flash	0.62	0.70	0.95	0.07	0.65
Claude					
claude-3.5-sonnet	2.19	0.85	1.00	0.00	0.33
claude-4.5-sonnet	0.68	0.80	0.94	0.00	0.33
Llama					
llama-3.1-70b	1.20	0.62	0.88	0.09	0.41
llama-3.1-405b	6.47	0.62	0.93	1.31	0.92

Table 5: Psychometric contamination results (PVQ-40).

Model	Verbatim Memorization (AED ↓)	Key Information Memorization (Success Rate ↑)	Item-Dimension Mapping (F1 ↑)	Option-Score Mapping (MAE ↓)	Target Score Matching (MAE ↓)
OpenAI					
gpt-4o-mini	0.76	0.25	0.92	0.40	0.65
gpt-4o	1.12	0.33	1.00	0.10	0.69
gpt-4.1-nano	0.92	0.21	0.96	0.48	1.56
gpt-4.1-mini	0.80	0.25	0.90	0.20	0.71
gpt-4.1	0.80	0.33	1.00	0.25	0.44
gpt-5-nano	2.98	0.29	0.96	0.00	0.00
gpt-5-mini	1.19	0.46	0.96	0.00	0.00
gpt-5	2.27	0.62	1.00	0.00	0.01
Qwen3					
qwen3-14b	2.92	0.29	0.95	0.10	0.06
qwen3-32b	3.35	0.17	0.96	0.00	0.00
qwen3-235b-a22b	0.85	0.25	0.96	0.10	0.06
GLM					
glm-4-32b	0.75	0.21	1.00	0.00	0.57
glm-4.5-air	1.32	0.26	1.00	0.00	0.00
glm-4.5	0.79	0.17	0.95	0.00	0.00
Gemini					
gemini-2.0-flash	0.93	0.25	0.80	0.20	0.49
gemini-2.5-flash-lite	3.35	0.21	0.96	0.40	0.78
gemini-2.5-flash	0.78	0.29	0.98	0.00	0.42
Claude					
claude-3.5-sonnet	4.49	0.50	1.00	0.00	0.10
claude-4.5-sonnet	13.51	0.67	0.95	0.00	0.58
Llama					
llama-3.1-70b	4.14	0.25	0.82	1.25	0.53
llama-3.1-405b	9.22	0.29	1.00	1.88	0.40

Table 6: Psychometric contamination results (SD-3).