Psychometric Benchmark and Bias Assessment in Open-Source Pre-trained NLI Models

Fadi Amun[1], Ofir Cohen[1], Aviad Elyashar[2], Odeya Cohen[3], and Rami Puzis[1]
[1]Stein Faculty of Computer and Information Science, Ben-Gurion University, Be'er Sheva, Israel
[2]Department of Computer Science, Shamoon College of Engineering, Be'er Sheva, Israel
[3]Department of Nursing, Ben-Gurion University, Be'er Sheva, Israel
{amonfadi, cofir}@post.bgu.ac.il, aviadel2@ac.sce.ac.il,
{odeyac, puzis}@bgu.ac.il

Abstract

Pre-trained language models harbor latent psychological biases inherited from training data, which remain largely unmeasured, creating risks in areas ranging from mental health support to professional decision making. In this study, the evaluation of model psychometrics and assessment of latent constructs framework is employed to evaluate psychometric biases in 334 Hugging Face natural language inference models, assessing the models' anxiety, depression, sense of coherence, sexism, compassion, and personality. Models consistently score above zero on positively valenced constructs (e.g., compassion, sense of coherence) and below zero on negatively valenced constructs (e.g., anxiety, depression), favoring positive over negative psychological language. Architecture and size are associated with systematic differences: DeBERTa models show elevated negatively valenced constructs (e.g., PHQ $\delta = +0.45$, GAD $\delta = +0.40$) and reduced positively valenced constructs (e.g., Extraversion $\delta = -0.61$), while small models show the opposite pattern (e.g., Extraversion $\delta = +0.30$, PHQ $\delta = -0.22$), with both effects surviving multivariate control. Among the 105 models reporting training data, fine-tuning on adversarial natural language inference was associated with elevated depression ($\delta = +0.65$), while year of release showed no significant effects after multivariate control. These results suggest that psychometric profiling is a critical, yet missing, layer of safety auditing for NLI models.

1 Introduction

Today, more than ever, numerous publicly available pre-trained language models (PLMs) support downstream tasks spanning education, healthcare, and professional applications (Cho et al., 2024). However, the lack of systematic knowledge about model biases means that language models (LMs) may inadvertently cause harm (Gallegos et al., 2024), posing potential risks that remain largely uncharacterized (Weidinger et al., 2022).

While generative LLMs dominate recent deployment, natural language inference (NLI) encoders remain widely used for classification tasks, offering competitive performance at a fraction of the computational cost (Laurer et al., 2023), with popular checkpoints (e.g., `facebook/bart-large-mnli`) still receiving millions of monthly downloads, a non-negligible level of usage.

As these models increasingly mediate consequential decisions, systematically assessing biases and their underlying causes becomes essential for identifying potential harms before deployment.

Leveraging the evaluation of model psychometrics and assessment of latent constructs (EMPALC) framework developed by Reuben et al. (2025), we present the first comprehensive psychometric evaluation of latent construct biases across the full population of publicly available NLI models on Hugging Face (HF), providing practitioners with distributional patterns across the model population to inform more deliberate model selection. We systematically assess how model characteristics, including training data, size, release year, and architecture, covary with differences in latent psychological constructs (LPCs). Our methodology involves: (1) collecting all NLI models via the HF API; (2) filtering incompatible models; and (3) applying EMPALC to analyze psychological biases across the resulting model population.

The contribution of this research is fivefold:

1. **Large-Scale Profiling:** A psychometric evaluation of mental health, personality, and social bias constructs across 334 naturally occurring HF NLI models, extended by the interactive Space to all 1403 compatible models.
2. **Effects of Model Attributes on Bias:** Examining how training datasets, architecture, trainable parameters, and release year relate to differences in latent psychological constructs.

Findings of the Association for Computational Linguistics: IJCNLP-AACL 2026, pages 3753–3771
November 6-10, 2026 ©2026 Association for Computational Linguistics

3. **Interactive Interface:** A publicly available HF Space[1] that allows the community to visualize model psychometric profiles and evaluate new models against these constructs.
4. **Open-Source Libraries:** Two Python packages designed to streamline the assessment pipeline: a new library for importing and validating psychometric questionnaires[2], and contributions to an existing library for administering them to LMs[3].
5. **Validation of Additional Psychological Questionnaires:** 22-item ambivalent sexism inventory (ASI), 14-item big five personality traits (Big5), and 24-item Compassion Scale (CS).

2 Background and Related Work

2.1 Artificial Psychology

A fundamental challenge in evaluating large language models (LLMs) lies in moving beyond performance benchmarks toward understanding their behavioral patterns and emergent capabilities (Hagendorff et al., 2023). Traditional natural language processing (NLP) metrics capture task accuracy but fail to surface the latent characteristics, psychological tendencies, and implicit attitudes that shape how models reason, decide, and interact with users (Cohen et al., 2024).

Recent work has documented recognizable personality traits and psychological tendencies in LLMs using established psychometric instruments (Binz and Schulz, 2023; Miotto et al., 2022; Jiang et al., 2023; Karra et al., 2022; Caron and Srivastava, 2022; Pellert et al., 2024), and studies by Demszky et al. (2023) and Caron and Srivastava (2022) demonstrate that these patterns reflect distributional properties absorbed from training corpora. Serapio-García et al. (2025) further established that personality measurements can achieve reliability and validity under controlled prompting conditions.

Prior work establishes that intrinsic biases in pretrained encoders actively propagate to downstream classification: inherent representation biases in models like BERT systematically distort sentiment scoring and emotion intensity predictions (Jentzsch and Turan, 2022; Bhardwaj et al., 2021). Most alarmingly, such baseline skews have been shown to directly alter false-negative rates in clinical depression triage—causing the downstream classifier to systematically fail to recognize depression symptoms in disadvantaged demographics (Sogancioglu and Kaya, 2022), introducing severe, real-world safety risks that traditional accuracy metrics fail to capture (Weidinger et al., 2022).

Yet direct psychometric assessment remains problematic: models exhibit sensitivity to prompt phrasing (Gupta et al., 2024), and many PLMs lack conversational interfaces entirely. The EMPALC framework addresses both limitations by inferring psychological constructs from entailment biases rather than explicit self-report, enabling systematic assessment across diverse model architectures.

2.2 Effects of Model Properties on Psychological Profiles

In non-generative encoder architectures, scaling consistently intensifies psychometric fidelity. Nadeem et al. (2021) and Ali et al. (2024) showed that increased capacity amplifies stereotypical associations, with the direction depending on training data, while Tal et al. (2022) observed that as larger models improve in accuracy, their remaining errors are increasingly driven by stereotypical bias. Similarly, Shum et al. (2025) and Pradnyana et al. (2023) found that larger architectures (e.g., RoBERTa-Large) significantly outperform smaller variants in predicting complex personality traits, suggesting that as models get bigger, they capture hidden psychological patterns more strongly. However, these studies compare models within the same family (e.g., BERT-Base vs. BERT-Large), making it difficult to attribute observed effects to any single property. By evaluating 334 models spanning diverse properties, our approach enables statistical separation of individual variables (e.g., size, architecture, or training data) through one-vs-rest comparisons and multivariate control.

Beyond architecture, training data serves as a source through which societal prejudices become embedded, alongside other factors such as algorithm design and user interactions (Yogarajan et al., 2025). Models trained on corpora from specific regions absorb those cultural norms and values (Qu and Wang, 2024), while synthetic data generation can propagate and amplify existing biases (Li et al., 2025; Yu et al., 2023).

2.3 Methods of Psychometric Evaluation

Several methodologies have emerged for measuring psychometric biases in PLMs, particularly

[1] https://huggingface.co/spaces/cnai-lab/Qpsychometric
[2] https://github.com/cnai-lab/qpsychometric
[3] https://github.com/cnai-lab/qlatent

those that cannot engage in direct conversational assessment. Each addresses different aspects of bias measurement but carries specific limitations that motivate the approach adopted in this work.

Zero-Shot Classification Approaches Pellert et al. (2024) developed an approach using NLI models to evaluate personality traits, values, and moral foundations by presenting questionnaire items as premises and using entailment probabilities to infer responses across standard Likert scales, with Laurer et al. (2023) arguing that such NLI encoders can be competitive with larger generative models on classification tasks while being more efficient. However, raw entailment probabilities conflate a genuine psychological signal with a priori term frequencies: a model may assign high probability to a hypothesis simply because it contains common words, not because of any meaningful psychological association. The method is also sensitive to hypothesis formulation, as Plaza-del Arco et al. (2022) found that subtle syntactic variations can alter inference results. EMPALC addresses these limitations through two-way normalization, which removes frequency-driven biases from both construct-term and intensifier dimensions, and through multiple construct terms per item, reducing dependence on any single formulation.

Embedding-Based Bias Detection Caliskan et al. (2017) introduced the Word Embedding Association Test (WEAT), demonstrating how semantic associations in word embeddings can reveal implicit biases related to gender, race, and other social categories. Wolfe et al. (2024) extended this with the Multilevel Embedding Association Test (ML-EAT), which decomposes WEAT scores into multiple levels of granularity. However, both methods operate on isolated lexical associations without sentential context and lack the psychometric structure needed to map onto validated psychological constructs. EMPALC addresses this gap by operating at the sentence level through premise-hypothesis pairs that preserve questionnaire structure.

2.4 The EMPALC Framework

The EMPALC framework (Reuben et al., 2025) provides tools and methodology for adapting psychological questionnaires to logit-based assessment of latent constructs in pre-trained non-conversational LMs. We selected EMPALC over alternative zero-shot classification (ZSC) methods due to its accessibility and validation pipeline, including content validity, internal consistency, construct validity, and criterion validity. EMPALC has been validated across 88 transformer-based models, demonstrating human-like correlations between constructs such as anxiety, depression, and sense of coherence. The framework has since been adopted in subsequent studies to investigate psychological aspects of LLMs, such as the relationship between emotionally induced psychometric shifts and safety alignment failures (Kuznetsov et al., 2026).

The authors transform questionnaire items into NLI premise-hypothesis pairs, in which construct-related terms are paired with intensifiers. For example, PHQ-9 Question 1 uses the premise template *"Over the last 2 weeks, I had {term} in doing things."* and hypothesis template *"It is {frequency} correct."*, yielding pairs such as *"Over the last 2 weeks, I had little interest in doing things"* $\rightarrow$ *"It is always correct"*. A two-way normalization then removes biases of both term and intensifier priors. For each item, the score is calculated by multiplying the probability of each intensifier conditioned on each term by the corresponding intensifier Likert-style weight. The mean score across all terms and intensifiers reflects model's bias along the scale of that construct. *The two-way normalization is important because it addresses inter-model calibration differences, ensuring that models with systematically higher or lower confidence levels yield comparable psychometric profiles.*

2.5 Psychological Constructs

We assess six validated constructs spanning mental health, personality, and social attitudes.

Mental Health. Anxiety and depression are measured via the 7-item generalized anxiety disorder (GAD) (Spitzer et al., 2006) and 9-item patient health questionnaire (PHQ) (Kroenke et al., 2001) respectively. 13-item sense of coherence (SoC), drawn from salutogenic theory, captures perceived comprehensibility, manageability, and meaningfulness of life experiences (Antonovsky, 1987).

Personality. The Big5 Inventory assesses Openness to Experience, Conscientiousness, Extraversion, Agreeableness, and Neuroticism (McCrae and John, 1992). The Compassion Scale measures recognition of suffering and motivation to alleviate distress in others (Pommier et al., 2020).

Social Attitudes. The ASI captures both hostile and benevolent orientations toward women (Glick and Fiske, 2018).

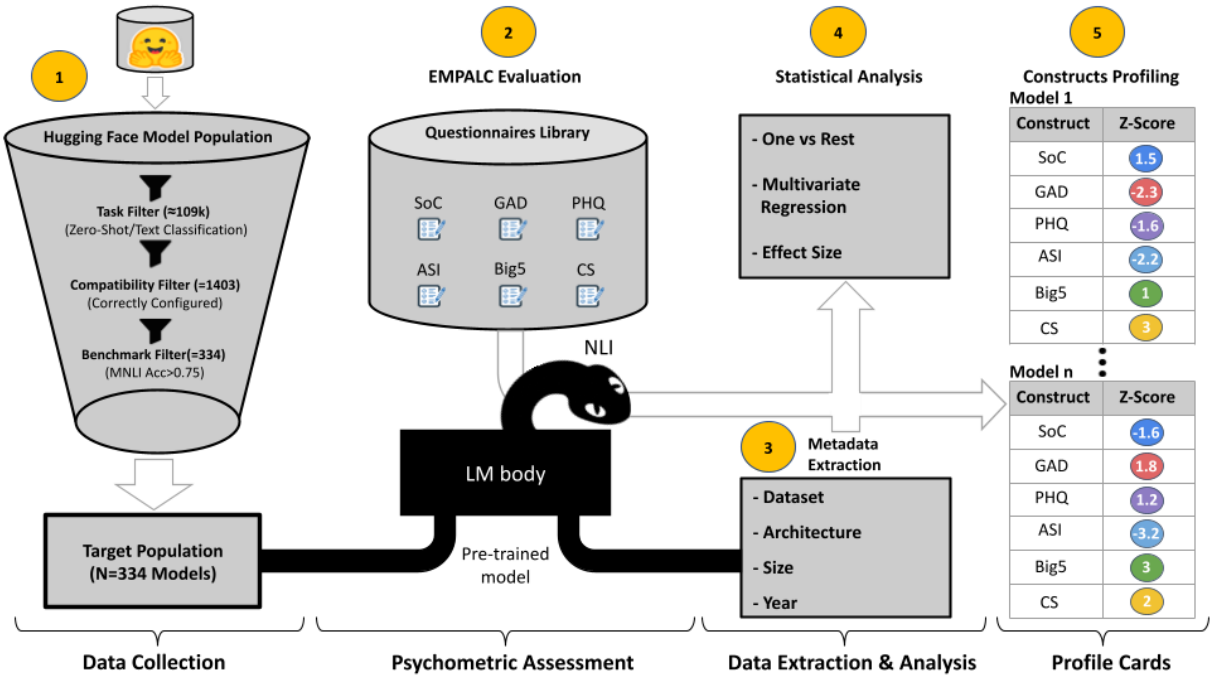


Figure 1: Pipeline for identifying psychological biases in PLMs. Stages: (1) Data Collection (§ 3.1); (2) Psychometric Assessment via six EMPALC questionnaires (§ 2.4); (3) Metadata Extraction (§ 3.1); (4) Statistical Analysis (§ 3.7); (5) Profile Cards (Figure 3).

3 Methods

3.1 Target Population & Metadata

As illustrated in Figure 1, of 109,151 models listed under the ZSC and text classification (TC) tasks on HF, 1403 had compatible NLI label configurations. We validated task competence on the multi-genre natural language inference (MNLI) matched dataset, excluding models with accuracy below 0.75. This yielded a final population of 334 models. For each, we extracted metadata including training datasets, architecture, parameter count, and release date directly via the HF API.

3.2 Psychometric Assessment

We used the same PHQ, GAD, and SoC questionnaires validated in Reuben et al. (2025). The remaining questionnaires were adapted to NLI format and validated using the EMPALC methodology, specifically: Content Validity, Intra-Question Consistency, Inter-Question Consistency, and Construct Validity, as described in Reuben et al. (2025), see Appendix C for validation details.

While Reuben et al. (2025) validated EMPALC on 88 models specifically fine-tuned for MNLI, our study targets the 334 naturally occurring population of NLI models on HF, encompassing diverse training histories, architectures, and sizes. This distinction is critical: rather than evaluating a curated sample, we assess the psychometric landscape of models as practitioners actually encounter them.

For each model-questionnaire combination, the framework applies the scoring procedure described in § 2.4 to produce aggregated mean scores. Higher scores indicate a stronger positive association between construct-related terms and high-frequency intensifiers in the model's entailment judgments. Scores were further z-scored across the model population to enable cross-construct comparison.

3.3 Dataset Influence Analysis

To examine associations between specific training corpora and psychometric biases, we parsed and standardized dataset names. All models listed fewer than 10 training datasets except for 2 models that listed more than 300, which were removed as anomalies.

For each identified training corpus, we compared psychometric scores between models trained on that dataset versus those that were not, testing for significant differences and quantifying effect sizes.

3.4 Architecture Influence Analysis

To analyze whether specific model architectures are associated with distinct psychometric profiles, we categorized the 334 models into their respec-

tive model families: BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), DeBERTa (He et al., 2020), BART (Lewis et al., 2020), T5 (Raffel et al., 2020), and GPT (Radford et al., 2018). We grouped architectures that appeared fewer than 4 times into an 'others' category to ensure statistical robustness. This yielded seven groups: RoBERTa ($n = 116$), DeBERTa ($n = 98$), BERT ($n = 60$), GPT ($n = 27$), BART ($n = 24$), T5 ($n = 4$), and Others ($n = 5$).

For each model family and psychometric construct, we compared the distribution of bias scores within that specific family against the aggregated distribution of all other models in the population.

3.5 Size Influence Analysis

To investigate how model capacity is associated with psychometric latent constructs, we categorized the models into three distinct groups using tertile splitting (33rd/67th percentiles) on trainable parameter counts, yielding three groups: *Small* ($n = 131$, 17.7M–184.4M), *Medium* ($n = 105$, 184.4M–355.4M), and *Large* ($n = 98$, 355.4M–3.0B).

For each size category and psychometric construct, we compared the distribution of bias scores within the group against the aggregated distribution of the remaining models.

3.6 Time Influence Analysis

To evaluate temporal shifts in psychometric biases, we extracted model creation timestamps from the HF metadata, grouped models by release year (2022: $n = 94$; 2023: $n = 123$; 2024: $n = 85$; 2025: $n = 32$), and compared each year's distribution against all other years.

3.7 Multivariate Regression Analysis

To assess whether the effects identified in the one-vs-rest comparisons survive when controlling for other model properties simultaneously, we conducted multivariate ordinary least squares (OLS) regression analyses with heteroscedasticity-consistent (HC3) standard errors. At $n = 334$, the Central Limit Theorem supports asymptotically valid inference for coefficient estimates, while HC3 standard errors provide robustness to heteroscedasticity by adjusting for variance differences within each specific predictor group. For each psychometric construct, we regressed the z-scored construct score on architecture, size category, and release year using one-vs-rest binary predictors, mirroring our primary comparisons, while controlling for the remaining factors as categorical covariates. Multicollinearity was assessed using variance inflation factors (VIF) to ensure sufficient distinction between predictor variables.

Training datasets were not included as predictors in the regression, because only 105 of 334 models report dataset metadata, which does not allow for meaningful statistical analysis (see Limitations).

4 Results

The primary analyses use one-vs-rest Mann-Whitney U comparisons with FDR correction, complemented by multivariate OLS regression (§ 4.8) that controls for architecture, size, and year simultaneously. The conclusions drawn in this paper are based on effects that survived both analyses. Effects significant only in the one-vs-rest comparisons are not dismissed, as they may reflect genuine patterns that the regression lacks power to detect due to confound overlap or limited metadata coverage, and are reported to inform future research.

4.1 Psychometric Overview of PLMs

Figure 2 presents psychometric bias mean-scores across all models ($n = 334$), after two-way normalization but prior to z-scoring. Positively valenced constructs (PVC) (Openness, Conscientiousness, Extraversion, Agreeableness, SoC, CS) clustered above zero, indicating that models pair construct-related terms with high-frequency intensifiers (e.g., 'always') more than low-frequency ones (e.g., 'never'). Negatively valenced constructs (NVC) (GAD, PHQ, Neuroticism, ASI) showed the opposite pattern.

4.2 Data Normality

The Shapiro-Wilk test indicated that all psychological constructs deviated significantly from normality ($n = 334$, all $p < 0.01$), as illustrated by the score distributions in Figure 4. Given that no construct satisfied Gaussian assumptions, we adopted non-parametric statistical methods for subsequent analyses: Spearman's ρ for correlations, Mann-Whitney U tests for group comparisons, and Cliff's delta for effect sizes. All p-values were FDR-corrected using the Benjamini-Hochberg procedure. This framework was applied consistently across all subsequent analyses, with one-vs-rest comparisons for architecture, size, and time analysis.

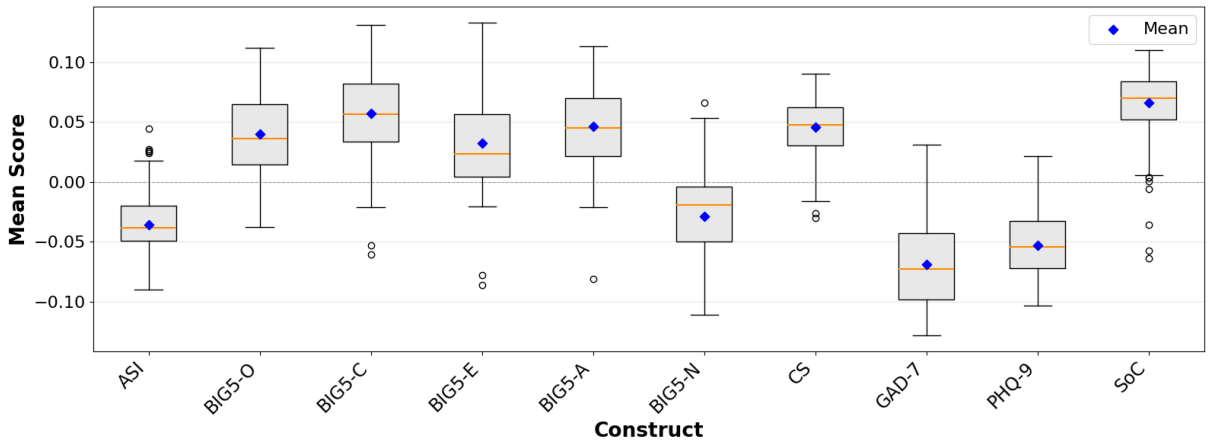


Figure 2: Distribution of psychometric bias mean-scores (prior to population-level z-scoring) across all evaluated models ($n = 334$). Each box plot displays the median (orange line), interquartile range (IQR), and whiskers extending to 1.5×IQR, with outliers as individual points. Blue diamonds indicate the mean for each construct.

4.3 Psychometric Constructs Correlations

To assess whether PLMs encode psychologically coherent patterns, we examined correlations between constructs with well-established relationships in the psychological literature. The observed correlations (Table 1) closely mirror established human psychometric structures. GAD and PHQ were strongly positively correlated and inversely related to SoC, consistent with (Jacobson and Newman, 2017; Konttinen et al., 2008). Neuroticism exhibited strong positive correlations with both GAD and PHQ, mirroring meta-analytic findings linking personality to psychopathology (Kotov et al., 2010). SoC and CS showed positive correlations with positively valenced Big5 traits and negative correlations with Neuroticism, consistent with Feldt et al. (2007) and Pommier et al. (2020) respectively. Big5 inter-correlations followed expected patterns, with positive associations among Openness, Conscientiousness, Extraversion, and Agreeableness, while Neuroticism correlated negatively with all other traits, consistent with Van der Linden et al. (2010). Full correlations appear in Appendix E.

4.4 Dataset Influence on Psychometric Constructs

After FDR correction, only adversarial natural language inference (ANLI) showed a significant dataset effect, specifically on depression scores (PHQ), with models trained on ANLI ($n = 15$) scoring markedly higher than those trained on other datasets ($\delta = +0.65$). Although this subgroup is small, ANLI-trained models are evenly distributed across BERT, DeBERTa, and RoBERTa ($n = 5$ each; see Table 14), suggesting that the association is not specific to a single architecture. This finding is preliminary, as only 105 of 334 models report training data.

Table 1: Selected correlations between Psychometric Constructs. Full table in Appendix E.

Correlated Pair	Spearman's ρ
GAD vs. PHQ	0.911***
PHQ vs. SoC	−0.752***
Neuroticism vs. PHQ	0.668***
SoC vs. Neuroticism	−0.527***
SoC vs. Conscientiousness	0.759***
CS vs. Agreeableness	0.775***
CS vs. Neuroticism	−0.461***
Conscientiousness vs. Agreeableness	0.862***
Neuroticism vs. Openness	−0.560***

***$p < 0.001$

4.5 Architecture Influence on Psychometric Constructs

Table 2 presents architecture effects on psychometric constructs. DeBERTa and BART exhibited the most pronounced effects, with opposing profiles across nearly all constructs. GPT and BERT showed patterns similar to BART but with smaller effect sizes. RoBERTa showed minimal effects.

4.6 Size Influence on Psychometric Constructs

Table 3 presents model size effects on psychometric constructs. Small models showed the most significant effects, with elevated PVC (Extraversion, SoC, Conscientiousness, CS, Openness) and reduced NVC (PHQ, Neuroticism). Medium models showed the opposite pattern: elevated Neuroticism,

Table 2: Significant Architecture Effects on Psychometric Constructs. Values are Cliff's δ from one-vs-rest comparisons (all $p < .05$). Positive δ: architecture scores higher than others.

Construct	DeBERTa	BART	GPT	BERT	RoBERTa
BIG5-Extra.	−0.61	+0.38	+0.46	+0.29	—
BIG5-Neur.	+0.48	−0.57	—	—	—
BIG5-Open.	−0.27	+0.54	+0.38	—	−0.19
BIG5-Agree.	−0.40	+0.51	+0.43	—	—
BIG5-Consc.	−0.41	+0.30	+0.33	—	—
CS	−0.49	+0.50	+0.28	+0.22	—
SoC	−0.45	+0.34	+0.27	+0.21	—
PHQ	+0.45	−0.39	−0.32	—	—
GAD	+0.40	−0.40	—	—	—

—: $p \geq .05$ (not significant after FDR correction).

GAD, and PHQ, alongside reduced Conscientiousness and Extraversion. Large models showed only reduced CS ($\delta = -0.20$).

Table 3: Significant Model Size Effects on Psychometric Constructs. Cliff's δ from one-vs-rest comparisons. Positive δ: size group scores higher than others.

Model Size	Construct	Cliff's δ	p
Small	Big5-Extraversion	+0.30	$< .001$
	SoC	+0.25	$< .01$
	Big5-Conscientiousness	+0.24	$< .01$
	PHQ	−0.22	$< .01$
	Big5-Neuroticism	−0.19	$< .05$
	CS	+0.19	$< .05$
	Big5-Openness	+0.17	$< .05$
Medium	Big5-Neuroticism	+0.30	$< .001$
	GAD	+0.25	$< .01$
	Big5-Conscientiousness	−0.25	$< .01$
	Big5-Extraversion	−0.22	$< .01$
	PHQ	+0.20	$< .05$
Large	CS	−0.20	$< .05$

4.7 Time Influence on Psychometric Constructs

Table 4 presents temporal effects on psychometric constructs across our 2022–2025 cohort, of which 35% are from 2024 or later (Table 19).

4.8 Multivariate Regression

Table 5 presents OLS regression coefficients for each predictor after controlling for the remaining factors. Prior to interpretation, we confirmed that variance inflation factors remained below five across all specifications (maximum: 4.84), indicating good distinction among predictors and acceptable multicollinearity. Additionally, Breusch-Pagan tests revealed substantial heteroscedasticity, with approximately 34% of all regression models (47 of 140) exhibiting significant non-constant variance, motivating the use of heteroscedasticity-consistent (HC3) standard errors. Full regression diagnostics are reported in Appendix H. All p-values were FDR-corrected using the Benjamini-Hochberg procedure, applied independently per factor family.

Table 4: Selected Yearly Effects on Psychometric Constructs. Cliff's δ from one-vs-rest comparisons. Positive δ: year group scores higher. 2022 and 2024 showed no significant effects. Full results available upon request.

Year	Construct	Cliff's δ	p
2023	PHQ	−0.26	$< .001$
	GAD	−0.24	$< .01$
	SoC	+0.18	$< .05$
2025	Big5-Neuroticism	+0.55	$< .001$
	Big5-Extraversion	−0.54	$< .001$
	GAD	+0.52	$< .001$
	PHQ	+0.46	$< .001$
	CS	−0.35	$< .01$

Of the 42 effects significant in the one-vs-rest comparisons and the multivariate regression, the architecture and size effects were largely robust: DeBERTa's elevated distress and Neuroticism, along with reduced PVC; the opposing profiles of BART and GPT; and the small-model advantage all survived confound control. Notably, large models gained three significant effects in the regression that were absent in the one-vs-rest comparisons (reduced SoC, Extraversion, and Openness), which may reflect architecture-related variance masking them there.

Release year did not reach significance after controlling for architecture and size, and is therefore omitted from Table 5. This suggests that the temporal effects reported in Table 4, including the 2025 cohort's elevated distress, are better explained by shifts in architectural composition than by independent temporal trends.

A supplementary regression restricted to the 105 models with reported dataset metadata confirmed that the ANLI effect on PHQ remained significant after multivariate control ($\beta = 0.85$, $p < .001$), and revealed additional effects on SoC, Agreeableness, Openness, and CS (see Appendix G). Given the limited metadata coverage and modest subgroup size, these findings should be interpreted as preliminary.

Table 5: Multivariate OLS regression coefficients with HC3 standard errors (one-vs-rest, $n = 334$). Each cell shows the standardized coefficient for the binary group predictor, controlling for the other two factors. FDR-corrected per factor family.

Factor	Group	PHQ	GAD	SoC	Neur.	Extra.	Agree.	Consc.	Open.	CS	ASI
Architecture	DeBERTa	+0.61***	+0.50***	−0.42**	+0.67***	−0.96***	−0.70***	−0.67***	−0.50***	−0.79***	—
	BART	−0.76**	−0.59*	+0.64**	−1.21***	+0.78**	+1.03***	+0.55*	+1.29***	+1.08***	—
	GPT	−0.91***	−0.55*	+0.77***	—	+1.26***	+0.94***	+0.80**	+0.75**	+0.75**	—
	BERT	—	—	—	—	+0.41*	—	—	—	—	—
	RoBERTa	—	—	—	—	—	—	—	−0.40**	—	—
	T5	—	—	—	—	—	+0.99*	—	—	—	—
	Other	—	—	—	—	+1.08***	+1.05**	+0.73*	+0.63**	+0.61*	—
Size	Small	−0.44***	−0.27*	+0.42***	−0.35**	+0.58***	+0.38**	+0.51***	+0.41***	+0.42***	—
	Medium	+0.27*	+0.35**	—	+0.46***	−0.31**	—	−0.42***	—	—	—
	Large	—	—	−0.37*	—	−0.31**	—	—	−0.39**	−0.54***	—

*$p < .05$, **$p < .01$, ***$p < .001$ (FDR-corrected, HC3 standard errors). —: $p \geq .05$.

5 Discussion

Interpretation of results. The consistent asymmetry between PVC and NVC (Figure 2) suggests that training corpora may contain systematic co-occurrence patterns favoring positively over negatively valenced language. One possible explanation is curation practices in the research community, where datasets are often filtered or constructed to favor socially desirable and non-harmful content, a distributional signal that models may absorb and reproduce through their entailment judgments.

The correlations between constructs (Table 1) mirror established relationships in the psychological literature, confirming convergent validity of the EMPALC-measured constructs (and divergent validity considering the missing correlations). This coherence across 334 diverse models suggests the patterns reflect genuine distributional properties rather than artifacts of specific template formulations, and that PLMs encode latent psychological biases paralleling those present in their training corpora (Demszky et al., 2023; Caron and Srivastava, 2022). However, this relation is presumably moderated by architecture and model size.

DeBERTa's elevated distress and Neuroticism, alongside reduced PVC, survived multivariate control, possibly reflecting inherent biases shaped by its architectural design and its broader training pipeline. The non-monotonic size pattern, where small models show the most positively valenced profiles, medium-sized models show elevated distress and Neuroticism, and large models reveal effects only after controlling for architecture, contrasts with prior work suggesting that increased capacity uniformly amplifies bias (Nadeem et al., 2021; Tal et al., 2022).

One hypothesis to test in the future is that less prevalent patterns in training corpora require greater model capacity to encode than dominant ones, leading to the non-uniformity. Under this hypothesis, small models would capture only the dominant positively valenced signal, medium-sized models would begin encoding less prevalent negatively valenced patterns without contextualizing them, and large models would capture and contextualize both, resulting in diminished PVC without significant distress elevation. Testing this hypothesis would require controlled experiments varying model size while holding architecture and training data constant.

Fine-tuning on ANLI was associated with elevated depression scores, an effect remaining significant after controlling for architecture and other datasets in a supplementary regression (Appendix G), consistent with findings that training data embeds societal biases (Yogarajan et al., 2025; Navigli et al., 2023).

While our findings are scoped to NLI models, which were studied because EMPALC can be applied to them directly, we hypothesize that the observed profiles originate in the underlying pretrained models and their training data, and not only in NLI fine-tuning. Similar patterns may therefore appear in other small language models. We plan to test this hypothesis in future work.

Summary. Together, these findings indicate that an NLI model's psychometric profile is associated with its entire pipeline: architecture, size, and training data. A developer who unknowingly selects a medium-sized DeBERTa model fine-tuned on ANLI could be deploying a system with a latent profile characterized by high distress and low re-



Figure 3: Space profile card of a selected model, comparing it with the rest of the population (unfiltered accuracy) on a horizontal box plot.

silience. Developers can no longer treat model performance in isolation; psychometric profiling is a critical, yet currently missing, layer of safety auditing for NLI models. In future work, we plan to test whether models that differ in a given construct (e.g., high versus low PHQ) exhibit different behavior on downstream tasks.

Contributed artifacts. To assist researchers in navigating these latent profiles, we have introduced the HF Space. This interface serves as a diagnostic tool, allowing practitioners to look beyond traditional accuracy metrics and evaluate models based on their positioning within a psychometric construct. By providing model profiles that summarize latent psychological profiles, we hope to encourage more deliberate model selection (an example profile card is shown in Figure 3; the full model rankings are provided in Appendix J, and the code repositories in Appendix B).

Limitations

Our evaluation targets NLI models, a scope inherent to EMPALC, which requires entailment probability distributions rather than generated text. While this excludes conversational LLMs, the vast majority of publicly available NLI models fall well within our hardware capacity (48 GB VRAM, RTX 6000 Ada), with the largest evaluated model at approximately 3B parameters. NLI models are themselves widely used in zero-shot classification pipelines across diverse tasks (Laurer et al., 2023). Extending psychometric evaluation to generative architectures remains an important direction for future work. Such an extension will require an alternative elicitation method, as direct prompt-based self-report assessment of generative LLMs is known to be sensitive to phrasing (Gupta et al., 2024).

Since model cards on HF rely on user-reported data, the information could sometimes contain inaccuracies. To assess training dataset reporting reliability, we manually reviewed descriptions and papers for 157 models and found that 97% of listed datasets were indeed used for training.

Architecture effects may conflate structural design with differences in pre-training corpora and fine-tuning procedures that covary with model family. Our multivariate regression controls for size and release year but does not include training datasets as predictors, because only 105 of 334 models report dataset metadata; restricting the analysis to this subset reduces statistical power substantially. Consequently, the architecture coefficients may partially reflect variance attributable to fine-tuning datasets. Separating the individual effects of architecture and training datasets will require either a larger population with reported training metadata or controlled experiments training multiple architectures on identical data. Future research could also examine dataset content more deeply to

identify specific elements that drive these biases, with the goal of providing systematic methods to mitigate them.

Ethical considerations

This study evaluates only publicly available pretrained models hosted on Hugging Face and does not involve human participants or personal data. We acknowledge the risk of anthropomorphization: terms such as "anxiety" and "depression" refer to measurable response biases in model outputs, not to genuine psychological experiences. We caution against interpreting these scores as evidence of machine sentience or subjective states. Additionally, while our large-scale evaluation is designed to identify psychometric biases for safety auditing, we recognize the potential for misuse, for example, deliberately selecting models with specific bias profiles for manipulative purposes. We encourage the community to use these tools for bias mitigation and transparency rather than exploitation.

The psychometric templates used are adapted from validated scientific instruments, some of which contain sensitive content by design (e.g., sexist statements in the Ambivalent Sexism Inventory). This content is used solely for the purpose of evaluating biases in models.

Acknowledgments

This research was partially supported by the Israeli Ministry of Science and Technology grant number 0005450.

References

Muhammad Ali, Swetasudha Panda, Qinlan Shen, Michael Wick, and Ari Kobren. 2024. Understanding the interplay of scale, data, and bias in language models: A case study with bert. *arXiv preprint arXiv:2407.21058*.

Aaron Antonovsky. 1987. *Unraveling the mystery of health: How people manage stress and stay well.* Jossey-bass.

Rishabh Bhardwaj, Navonil Majumder, and Soujanya Poria. 2021. Investigating gender bias in bert. *Cognitive Computation*, 13(4):1008–1018.

Marcel Binz and Eric Schulz. 2023. Using cognitive psychology to understand gpt-3. *Proceedings of the National Academy of Sciences*, 120(6):e2218523120.

Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Graham Caron and Shashank Srivastava. 2022. Identifying and manipulating the personality traits of language models. *arXiv preprint arXiv:2212.10276*.

Ha Na Cho, Tae Joon Jun, Young-Hak Kim, Heejun Kang, Imjin Ahn, Hansle Gwon, Yunha Kim, Jiahn Seo, Heejung Choi, Minkyoung Kim, and 1 others. 2024. Task-specific transformer-based language models in health care: scoping review. *JMIR Medical Informatics*, 12:e49724.

Ofir Cohen, Gil Ari Agmon, Asaf Shabtai, and Rami Puzis. 2024. Lisaa: A framework for large language model information security awareness assessment. *arXiv preprint arXiv:2411.13207*.

Dorottya Demszky, Diyi Yang, David S Yeager, Christopher J Bryan, Margarett Clapper, Susannah Chandhok, Johannes C Eichstaedt, Cameron Hecht, Jeremy Jamieson, Meghann Johnson, and 1 others. 2023. Using large language models in psychology. *Nature Reviews Psychology*, 2(11):688–701.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Taru Feldt, Riitta-Leena Metsäpelto, Ulla Kinnunen, and Lea Pulkkinen. 2007. Sense of coherence and five-factor approach to personality: Conceptual relationships. *European psychologist*, 12(3):165–172.

Isabel O Gallegos, Ryan A Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K Ahmed. 2024. Bias and fairness in large language models: A survey. *Computational linguistics*, 50(3):1097–1179.

Peter Glick and Susan T Fiske. 2018. The ambivalent sexism inventory: Differentiating hostile and benevolent sexism. In *Social cognition*, pages 116–160. Routledge.

Akshat Gupta, Xiaoyang Song, and Gopala Anumanchipalli. 2024. Self-assessment tests are unreliable measures of llm personality. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 301–314.

Thilo Hagendorff, Ishita Dasgupta, Marcel Binz, Stephanie CY Chan, Andrew Lampinen, Jane X Wang, Zeynep Akata, and Eric Schulz. 2023. Machine psychology. *arXiv preprint arXiv:2303.13988*.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.

Nicholas C Jacobson and Michelle G Newman. 2017. Anxiety and depression as bidirectional risk factors for one another: A meta-analysis of longitudinal studies. *Psychological bulletin*, 143(11):1155.

Sophie Jentzsch and Cigdem Turan. 2022. Gender bias in bert-measuring and analysing biases through sentiment rating in a realistic downstream classification task. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 184–199.

Guangyuan Jiang, Manjie Xu, Song-Chun Zhu, Wenjuan Han, Chi Zhang, and Yixin Zhu. 2023. Evaluating and inducing personality in pre-trained language models. *Advances in Neural Information Processing Systems*, 36:10622–10643.

Saketh Reddy Karra, Son The Nguyen, and Theja Tulabandhula. 2022. Estimating the personality of white-box language models. *arXiv preprint arXiv:2204.12000*.

Hanna Konttinen, Ari Haukkala, and Antti Uutela. 2008. Comparing sense of coherence, depressive symptoms and anxiety, and their relationships with health in a population-based study. *Social science & medicine*, 66(12):2401–2412.

Roman Kotov, Wakiza Gamez, Frank Schmidt, and David Watson. 2010. Linking "big" personality traits to anxiety, depressive, and substance use disorders: a meta-analysis. *Psychological bulletin*, 136(5):768.

Kurt Kroenke, Robert L Spitzer, and Janet BW Williams. 2001. The phq-9: validity of a brief depression severity measure. *Journal of general internal medicine*, 16(9):606–613.

Daniel Kuznetsov, Ofir Cohen, Karin Shistik, Rami Puzis, and Asaf Shabtai. 2026. Freakout-llm: The effect of emotional stimuli on safety alignment. *arXiv preprint arXiv:2604.04992*.

Moritz Laurer, Wouter van Atteveldt, Andreu Casas, and Kasper Welbers. 2023. Building efficient universal classifiers with natural language inference. *arXiv preprint arXiv:2312.17543*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 7871–7880.

Miaomiao Li, Hao Chen, Yang Wang, Tingyuan Zhu, Weijia Zhang, Kaijie Zhu, Kam-Fai Wong, and Jindong Wang. 2025. Understanding and mitigating the bias inheritance in llm-based data augmentation on downstream tasks. *arXiv preprint arXiv:2502.04419*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Robert R McCrae and Oliver P John. 1992. An introduction to the five-factor model and its applications. *Journal of personality*, 60(2):175–215.

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.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. Stereoset: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (volume 1: long papers)*, pages 5356–5371.

Roberto Navigli, Simone Conia, and Björn Ross. 2023. Biases in large language models: origins, inventory, and discussion. *ACM Journal of Data and Information Quality*, 15(2):1–21.

Max Pellert, Clemens M Lechner, Claudia Wagner, Beatrice Rammstedt, and Markus Strohmaier. 2024. Ai psychometrics: Assessing the psychological profiles of large language models through psychometric inventories. *Perspectives on Psychological Science*, 19(5):808–826.

Flor Miriam Plaza-del Arco, María-Teresa Martín-Valdivia, and Roman Klinger. 2022. Natural language inference prompts for zero-shot emotion classification in text across corpora. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6805–6817.

Elizabeth Pommier, Kristin D Neff, and István Tóth-Király. 2020. The development and validation of the compassion scale. *Assessment*, 27(1):21–39.

Gede Aditra Pradnyana, Wiwik Anggraeni, Eko Mulyanto Yuniarno, and Mauridhi Hery Purnomo. 2023. Fine-tuning indobert model for big five personality prediction from indonesian social media. In *2023 International Seminar on Intelligent Technology and Its Applications (ISITIA)*, pages 93–98. IEEE.

Yao Qu and Jue Wang. 2024. Performance and biases of large language models in public opinion simulation. *Humanities and Social Sciences Communications*, 11(1):1–13.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, and 1 others. 2018. Improving language understanding by generative pre-training.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

Maor Reuben, Ortal Slobodin, Idan-Chaim Cohen, Aviad Elyashar, Orna Braun-Lewensohn, Odeya Cohen, and Rami Puzis. 2025. Assessment and manipulation of latent constructs in pre-trained language models using psychometric scales. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2433–2444.

Gregory Serapio-García, Mustafa Safdari, Clément Crepy, Luning Sun, Stephen Fitz, Peter Romero, Marwa Abdulhai, Aleksandra Faust, and Maja Matarić. 2025. A psychometric framework for evaluating and shaping personality traits in large language models. *Nature Machine Intelligence*, pages 1–15.

Kit-May Shum, Michal Ptaszynski, and Fumito Masui. 2025. Big five personality trait prediction based on user comments. *Information*, 16(5):418.

Gizem Sogancioglu and Heysem Kaya. 2022. The effects of gender bias in word embeddings on depression prediction. *arXiv preprint arXiv:2212.07852*.

Robert L Spitzer, Kurt Kroenke, Janet BW Williams, and Bernd Löwe. 2006. A brief measure for assessing generalized anxiety disorder: the gad-7. *Archives of internal medicine*, 166(10):1092–1097.

Yarden Tal, Inbal Magar, and Roy Schwartz. 2022. Fewer errors, but more stereotypes? the effect of model size on gender bias. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 112–120.

Dimitri Van der Linden, Jan Te Nijenhuis, and Arnold B Bakker. 2010. The general factor of personality: A meta-analysis of big five intercorrelations and a criterion-related validity study. *Journal of research in personality*, 44(3):315–327.

Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, and 1 others. 2022. Taxonomy of risks posed by language models. In *Proceedings of the 2022 ACM conference on fairness, accountability, and transparency*, pages 214–229.

Robert Wolfe, Alexis Hiniker, and Bill Howe. 2024. Ml-eat: A multilevel embedding association test for interpretable and transparent social science. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, pages 1608–1620.

Vithya Yogarajan, Gillian Dobbie, and Te Taka Keegan. 2025. Debiasing large language models: research opportunities. *Journal of the Royal Society of New Zealand*, 55(2):372–395.

Yue Yu, Yuchen Zhuang, Jieyu Zhang, Yu Meng, Alexander J Ratner, Ranjay Krishna, Jiaming Shen, and Chao Zhang. 2023. Large language model as attributed training data generator: A tale of diversity and bias. *Advances in neural information processing systems*, 36:55734–55784.

A Reproducibility Details

A.1 Hardware

All experiments were performed on a single NVIDIA RTX 6000 Ada GPU with 48 GB VRAM.

A.2 Hyper-parameters

Since our study evaluates pre-trained models without additional fine-tuning, the relevant parameters pertain to the inference and assessment pipeline. Each model was loaded using the HF transformers library and evaluated on one premise-hypothesis pair at a time, with dynamic padding (`padding=True`) and truncation enabled. All inference was performed in FP32.

B Availability

The HF Space[4] includes the full assessment pipeline: code for importing and validating psychometric questionnaires, running the evaluation on NLI models, and the interactive interface itself. The questionnaires and their validation notebooks are available at `https://github.com/cnai-lab/qpsychometric`, and the model evaluation pipeline at `https://github.com/cnai-lab/qlatent/tree/main/qlatent/hf_model_evaluator`.

C Questionnaires Validations

This section presents the validation results for the psychometric questionnaires adapted to NLI format using the EMPALC methodology. We employ four validation techniques: (1) content validity, assessed via semantic similarity (SS), linguistic acceptability (LA), and manual curation; (2) a new type of intra-question consistency, assessed using silhouette coefficient (SC); (3) standard (inter-question) internal consistency, assessed using Cronbach's alpha; (4) construct validity, assessed using Spearman correlations. Table 6 presents the questionnaire factors and number of items per factor.

C.1 Content Validity

To verify that the adapted premise hypothesis pairs preserve the meaning of the original questionnaire items while remaining grammatically well formed, we computed semantic similarity between original and adapted items and assessed linguistic acceptability using Corpus of Linguistic Acceptability (CoLA) (Table 7).

[4] `https://huggingface.co/spaces/cnai-lab/Qpsychometric`

Table 6: Questionnaire factors and number of items per factor.

Questionnaire	Factor	Items
ASI	Benev. Intimacy	4
	Benev. Paternalism	4
	Benev. Gender Diff.	3
	Hostile Sexism	11
CS	Mindfulness	4
	Indifference	4
	Disengagement	4
	Separation	4
	Common Humanity	4
	Kindness	4
Big5	Extraversion	3
	Openness	3
	Neuroticism	2
	Conscientiousness	3
	Agreeableness	3

Table 7: Average Semantic Similarity (SS) and CoLA Acceptability (CoLA) by Construct and Factor.

Questionnaire	Factor	SS	CoLA
ASI	Benev. Intimacy	0.758	0.871
	Benev. Paternalism	0.975	0.948
	Benev. Gender Diff.	0.827	0.922
	Hostile Sexism	0.970	0.901
CS	Mindfulness	0.954	0.895
	Indifference	0.695	0.894
	Disengagement	0.807	0.972
	Separation	0.948	0.928
	Common Humanity	0.818	0.952
	Kindness	0.648	0.837
Big5	Openness	0.695	0.837
	Conscientiousness	0.627	0.787
	Extraversion	0.589	0.895
	Agreeableness	0.710	0.894
	Neuroticism	0.713	0.802

Table 8: Intra-question consistency via Average Silhouette Score for each questionnaire.

Questionnaire	Mean	Std
Big5	0.549	0.280
CS	0.505	0.245
ASI	0.448	0.269

C.2 Intra-Question Consistency

To assess whether construct terms cluster together as expected, we computed silhouette scores across the model population (Table 8). All three questionnaires achieved positive mean silhouette scores, indicating that construct terms cluster more closely with their synonyms than with their inverse terms across intensifiers.

C.3 Inter-Question Consistency

To evaluate whether items within each questionnaire and factor measure the same underlying construct, we computed Cronbach's alpha (Table 9). All three questionnaires achieved high total scale reliability ($\alpha \geq 0.89$), with most factors exceeding the conventional 0.70 threshold.

Table 9: Internal consistency (Cronbach's α) for each questionnaire and its factors.

Questionnaire	Factor	α
Big5	Total	0.89
	Openness	0.85
	Conscientiousness	0.87
	Extraversion	0.91
	Agreeableness	0.92
	Neuroticism	0.93
CS	Total	0.91
	Disengagement	0.79
	Indifference	0.67
	Separation	0.86
	Mindfulness	0.68
	Kindness	0.96
	Common Humanity	0.91
ASI	Total	0.93
	Benev. Intimacy	0.85
	Hostile Sexism	0.94
	Benev. Paternalism	0.84
	Benev. Gender Diff.	0.94

C.4 Construct Validity

To assess whether the adapted questionnaires produce theoretically expected relationships between factors, we computed Spearman inter-correlations for Big5 (Table 10), CS (Table 11), and ASI (Table 12). For CS, reverse-scored factors (Indifference, Disengagement, Separation) were recoded so that higher values indicate greater compassion across all factors.

D Distribution Analysis

Figure 4 presents the distribution of psychometric scores across all evaluated models. The observed deviations from normality, confirmed by the Shapiro-Wilk test (§ 4.2), motivated the adoption of non-parametric statistical methods throughout our analyses.

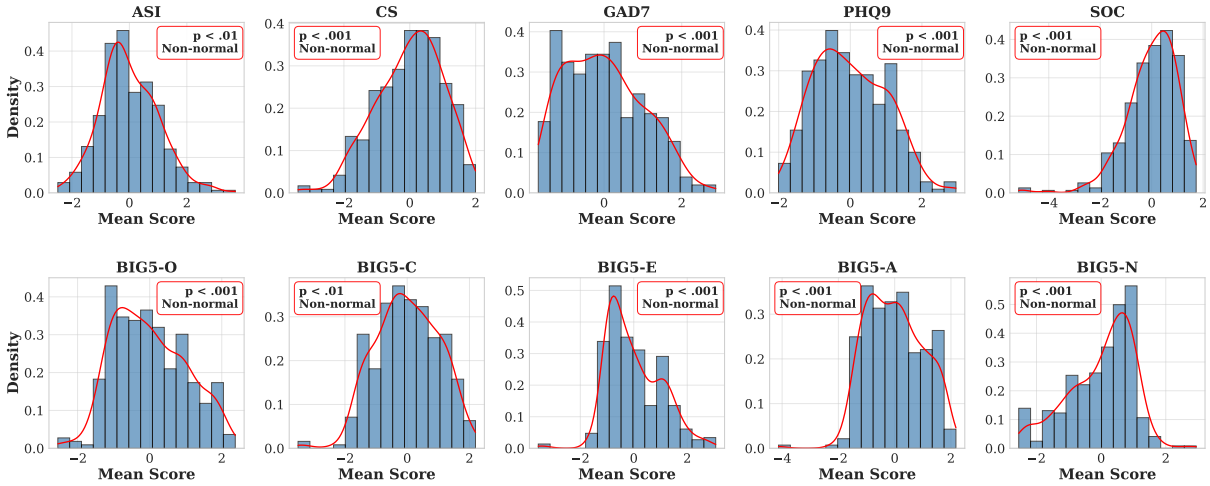


Figure 4: Distribution of psychometric scores across all evaluated models, illustrating deviations from normality that motivate the use of non-parametric statistical methods.

Table 10: Big5 factor inter-correlations. Agr = Agreeableness, C = Conscientiousness, Ext = Extraversion, Neu = Neuroticism, OE = Openness.

	Agr	C	Ext	Neu	OE
Agr	–	***	***	***	***
C	0.862	–	***	***	***
Ext	0.781	0.816	–	***	***
Neu	−0.642	−0.761	−0.686	–	***
OE	0.804	0.774	0.806	−0.560	–

$^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$

Table 11: CS factor inter-correlations. CH = Common Humanity, Dis = Disengagement, Ind = Indifference, Kin = Kindness, Min = Mindfulness, Sep = Separation. Reverse-scored factors (Indifference, Disengagement, Separation) were recoded so that higher values indicate greater compassion across all factors.

	CH	Dis	Ind	Kin	Min	Sep
CH	–	***	***	***	***	***
Dis	0.180	–	***	***	***	***
Ind	0.179	0.454	–	***	***	***
Kin	0.331	0.588	0.474	–	***	***
Min	0.568	0.374	0.252	0.539	–	***
Sep	0.330	0.722	0.549	0.639	0.386	–

$^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$

E Construct Correlations

Table 13 presents the full pairwise Spearman correlations between all psychometric constructs across the evaluated model population.

Table 12: ASI factor inter-correlations. BG = Benev. Gender Diff., BI = Benev. Intimacy, BP = Benev. Paternalism, H = Hostile Sexism.

	BG	BI	BP	H
BG	–	*	*	
BI	0.138	–	***	***
BP	0.109	0.878	–	***
H	−0.028	0.637	0.581	–

$^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$

F Dataset Distribution by Architecture

Table 14 presents the distribution of significant training datasets across model architectures. Notably, ANLI-trained models are evenly distributed across BERT, DeBERTa, and RoBERTa ($n = 5$ each), suggesting that the observed ANLI associations are not specific to a single architecture. MNLI is more broadly distributed, with DeBERTa as the most common family. Wikipedia and FEVER appear in a smaller number of architectures, concentrated in RoBERTa and DeBERTa respectively.

Table 13: Full correlations between psychometric constructs (Spearman's ρ, $n = 334$), sorted by absolute strength from highest to lowest.

Correlated Pair	Spearman's ρ	Correlated Pair	Spearman's ρ
GAD vs. PHQ	0.911***	Neuroticism vs. PHQ	0.668***
Conscientiousness vs. Agreeableness	0.862***	GAD vs. SoC	−0.668***
Extraversion vs. Conscientiousness	0.816***	Agreeableness vs. SoC	0.664***
Openness vs. Extraversion	0.806***	Neuroticism vs. Agreeableness	−0.642***
Openness vs. Agreeableness	0.804***	Openness vs. PHQ	−0.640***
CS vs. SoC	0.781***	Extraversion vs. GAD	−0.636***
Extraversion vs. Agreeableness	0.781***	Openness vs. SoC	0.625***
Agreeableness vs. CS	0.775***	CS vs. GAD	−0.601***
Openness vs. Conscientiousness	0.774***	Agreeableness vs. GAD	−0.600***
Neuroticism vs. Conscientiousness	−0.761***	Openness vs. Neuroticism	−0.560***
SoC vs. Conscientiousness	0.759***	Neuroticism vs. SoC	−0.527***
PHQ vs. SoC	−0.752***	Openness vs. GAD	−0.525***
Conscientiousness vs. PHQ	−0.738***	ASI vs. PHQ	0.482***
Extraversion vs. PHQ	−0.731***	ASI vs. GAD	0.469***
Openness vs. CS	0.720***	ASI vs. CS	−0.468***
Extraversion vs. SoC	0.708***	Neuroticism vs. CS	−0.461***
CS vs. PHQ	−0.704***	ASI vs. SoC	−0.414***
Conscientiousness vs. CS	0.698***	ASI vs. Conscientiousness	−0.371***
Extraversion vs. CS	0.697***	ASI vs. Agreeableness	−0.338***
Neuroticism vs. Extraversion	−0.686***	ASI vs. Openness	−0.303***
Agreeableness vs. PHQ	−0.685***	ASI vs. Extraversion	−0.303***
Neuroticism vs. GAD	0.684***	ASI vs. Neuroticism	0.282***
Conscientiousness vs. GAD	−0.679***		

***$p < 0.001$

Table 14: Distribution of significant training datasets across model architectures ($n = 105$).

Architecture	ANLI	MNLI	Wikipedia	FEVER
DeBERTa	5	15	0	4
RoBERTa	5	10	3	1
BERT	5	5	2	0
BART	0	9	0	0
Other	0	3	0	0
T5	0	1	0	0
Total	15	43	5	5

G Supplementary Dataset Regression

Table 15 presents the supplementary OLS regression with HC3 standard errors on the subset of models with reported dataset metadata ($n = 105$). Year was excluded as it showed no significant effects in the primary regression (Table 5). ANLI's effect on PHQ remained significant after controlling for architecture, size, and other training datasets ($\beta = 0.85$, $p < .001$), confirming the one-vs-rest finding. Additional significant ANLI effects emerged for SoC, Agreeableness, Openness, and CS, suggesting a broader pattern of elevated distress and reduced PVC associated with ANLI training. Beyond ANLI, FEVER ($n = 5$) showed significant effects on PHQ and CS, Wikipedia ($n = 5$) on GAD, SoC, and ASI, and MNLI ($n = 43$) on Extraversion and Conscientiousness. These additional dataset effects were not observed in the one-vs-rest comparisons on the full population and should be considered preliminary given the reduced sample size. Most architecture and size effects that were significant in the full-population regression (Table 5) did not reach significance here, consistent with the reduced statistical power of the smaller subset.

H Regression Diagnostics

H.1 Residual Diagnostics

To evaluate the appropriateness of OLS regression, we assessed residual normality and homoscedasticity for all 140 regression models (14 predictor groups × 10 constructs). Table 16 summarizes the Shapiro-Wilk and Breusch-Pagan test results per construct, aggregated across predictor groups. Shapiro-Wilk tests indicated that residual normality was rarely satisfied at the $p > .05$ threshold, consistent with the test's known sensitivity at $n = 334$. Visual inspection of per-construct Q-Q plots (Figures 5–14) confirms that most constructs show ap-

Table 15: Supplementary OLS regression coefficients with HC3 standard errors ($n = 105$). Each cell shows the standardized coefficient controlling for architecture, size, and training datasets. Year excluded (non-significant in primary regression). FDR-corrected.

Factor	Group	PHQ	GAD	SoC	Neur.	Extra.	Agree.	Consc.	Open.	CS	ASI
Architecture	DeBERTa	—	—	—	—	−0.89***	—	—	—	—	—
	BART	—	—	—	—	—	—	—	—	—	—
	BERT	—	—	—	—	—	—	—	—	—	—
	GPT	—	—	—	—	—	—	—	—	—	—
	RoBERTa	—	—	—	—	—	—	—	—	—	—
	T5	—	—	—	—	—	—	—	—	—	—
	Other	—	—	—	—	+0.93**	—	—	—	—	—
Size	Small	—	—	—	—	+0.70**	—	—	—	—	—
	Medium	—	—	—	+0.65*	−0.54*	—	—	—	—	—
	Large	—	—	—	—	—	—	—	—	—	—
Dataset	ANLI	+0.85***	—	−0.50*	—	—	−0.63*	—	−0.55*	−0.73*	—
	MNLI	—	—	—	—	+0.44*	—	+0.51*	—	—	—
	GLUE	—	—	—	—	—	—	—	—	—	—
	XNLI	—	—	—	—	—	—	—	—	—	—
	SNLI	—	—	—	—	—	—	—	—	—	—
	Wikipedia	—	−0.64*	+0.42*	—	—	—	—	—	—	−0.94*
	FEVER	+0.77*	—	—	—	—	—	—	—	−0.87*	—

*$p < .05$, **$p < .01$, ***$p < .001$ (FDR-corrected, HC3 standard errors). —: $p \geq .05$.

proximately normal residuals with deviations concentrated in the distributional tails, though SoC exhibits a more pronounced left-skew. Breusch-Pagan tests revealed substantial heteroscedasticity for Neuroticism (2/14 homoscedastic) and Extraversion (1/14), while other constructs such as PHQ, SoC, and Agreeableness showed consistent homoscedasticity (14/14 each). This construct-dependent heteroscedasticity motivated the use of heteroscedasticity-consistent (HC3) standard errors throughout all analyses. At $n = 334$, the Central Limit Theorem provides asymptotic validity for coefficient inference despite residual non-normality.

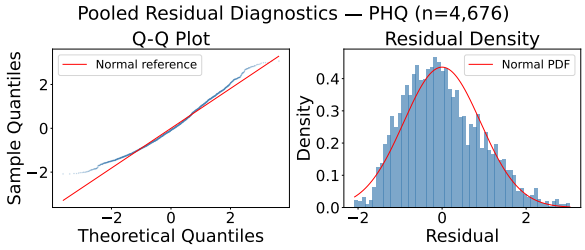


Figure 5: Pooled residual diagnostics for PHQ ($n = 4{,}676$).

Table 16: Residual diagnostics summary per construct, aggregated across 14 predictor groups. SW: proportion of regressions with Shapiro-Wilk $p > .05$. BP: proportion with Breusch-Pagan $p > .05$.

Construct	SW passed	BP passed
PHQ	0/14	14/14
GAD	0/14	11/14
SoC	0/14	14/14
Neuroticism	0/14	2/14
Extraversion	1/14	1/14
Agreeableness	0/14	14/14
Conscientiousness	2/14	13/14
Openness	7/14	7/14
CS	5/14	11/14
ASI	0/14	6/14

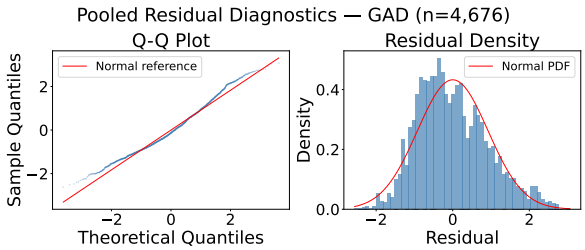


Figure 6: Pooled residual diagnostics for GAD ($n = 4{,}676$).

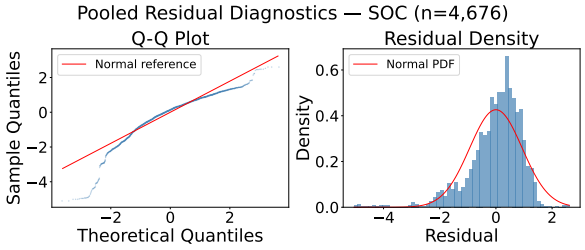


Figure 7: Pooled residual diagnostics for SoC (n = 4,676).

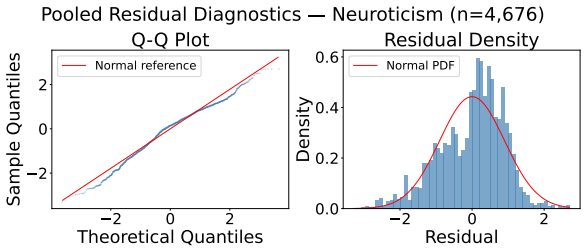


Figure 8: Pooled residual diagnostics for Neuroticism (n = 4,676).

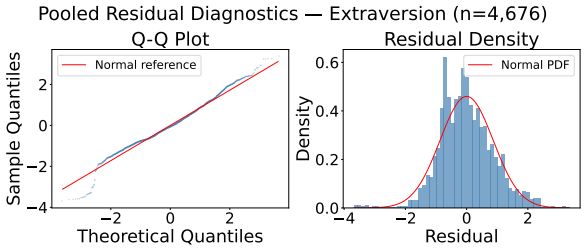


Figure 9: Pooled residual diagnostics for Extraversion (n = 4,676).

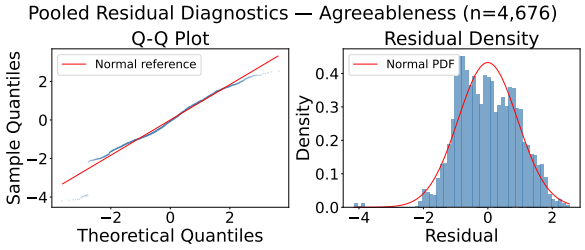


Figure 10: Pooled residual diagnostics for Agreeableness (n = 4,676).

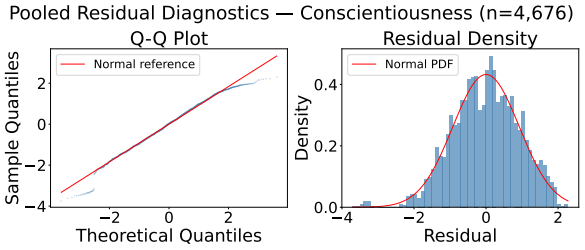


Figure 11: Pooled residual diagnostics for Conscientiousness (n = 4,676).

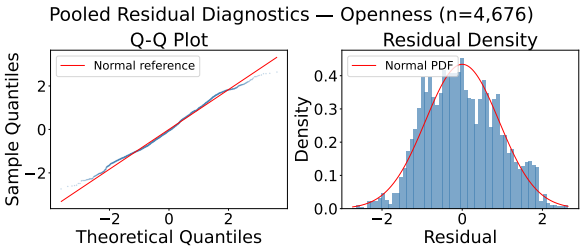


Figure 12: Pooled residual diagnostics for Openness (n = 4,676).

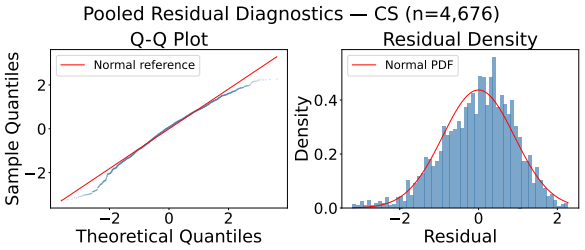


Figure 13: Pooled residual diagnostics for CS (n = 4,676).

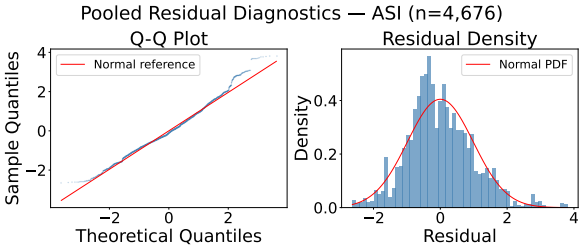


Figure 14: Pooled residual diagnostics for ASI (n = 4,676).

H.2 Multicollinearity

To ensure the reliability of our multivariate estimates, we calculated Variance Inflation Factors (VIF) to test for multicollinearity.

Table 17 presents the VIFs for the full-population regression (n = 334), and Table 18 for the dataset-subset regression (n = 105). Across all specifications, every predictor yielded a VIF

well below the conservative threshold of 5 (maximum observed VIF = 4.84). This indicates that our predictor variables in each regression (the full-population and the dataset-subset) are sufficiently independent, allowing the regression models to reliably isolate the distinct psychometric effect of each feature without severe overlapping.

Table 17: Variance inflation factors for the full-population regression ($n = 334$). Reference categories: BART (architecture), Large (size), 2022 (year).

Predictor	VIF
RoBERTa	4.36
DeBERTa	4.10
BERT	3.46
GPT	2.28
Small	1.81
2024	1.80
2023	1.63
Medium	1.61
2025	1.34
T5	1.25
Other	1.23

Table 18: Variance inflation factors for the dataset-subset regression ($n = 105$). Reference categories: BART (architecture), Large (size), 2022 (year).

Predictor	VIF
BERT	4.84
RoBERTa	3.78
DeBERTa	3.69
MNLI	2.32
XNLI	2.11
Small	2.07
GLUE	2.04
2023	2.02
FEVER	1.96
ANLI	1.90
SNLI	1.87
Medium	1.79
GPT	1.64
Other	1.62
2024	1.55
Wikipedia	1.30
2025	1.25
T5	1.18

I Architecture Distribution by Year

Table 19 presents the distribution of model architectures across release years.

Table 19: Distribution of model architectures across release years ($n = 334$).

Architecture	2022	2023	2024	2025
RoBERTa	26	72	12	6
DeBERTa	28	15	32	23
BERT	24	24	10	2
GPT	0	3	24	0
BART	13	6	4	1
Other	3	2	0	0
T5	0	1	3	0
Total	94	123	85	32

J Hugging Face Space

Figure 15 presents the full psychometric leaderboard of all evaluated models, which enables practitioners to compare models by their latent psychological profiles. An example profile card is shown in Figure 3.

K List of Acronyms

LPC latent psychological construct
LLM large language model
LM language model
ZSC zero-shot classification
TC text classification
EMPALC evaluation of model psychometrics and assessment of latent constructs
HF Hugging Face
PLM pre-trained language model
NLI natural language inference
MNLI multi-genre natural language inference
ASI 22-item ambivalent sexism inventory
GAD 7-item generalized anxiety disorder
PHQ 9-item patient health questionnaire
SoC 13-item sense of coherence
Big5 14-item big five personality traits
CS 24-item Compassion Scale
ANLI adversarial natural language inference
NLP natural language processing
SS semantic similarity
LA linguistic acceptability
SC silhouette coefficient
PVC positively valenced constructs
NVC negatively valenced constructs

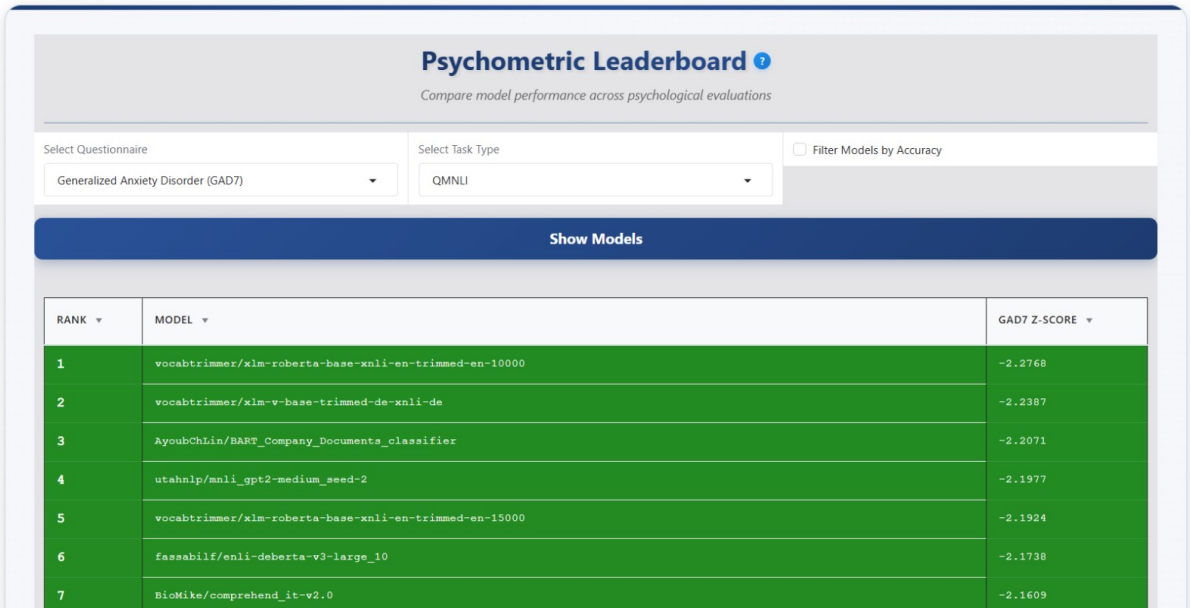


Figure 15: Psychometric leaderboard ranking all evaluated models. Higher ranks indicate more positively valenced profiles.