Assessment and manipulation of latent constructs in pre-trained language models using psychometric scales

Anonymous ACL submission

Abstract

Human-like personality traits have recently been discovered in large language models, raising the hypothesis that their (known and as yet undiscovered) biases conform with human latent psychological constructs. While large conversational models may be tricked into answering psychometric questionnaires, the latent psychological constructs of thousands of simpler transformers, trained for other tasks, cannot be assessed because appropriate psychometric methods are currently lacking. Here, we show how standard psychological questionnaires can be reformulated into natural language inference prompts, and we provide a code library to support the psychometric assessment of arbitrary models. We demonstrate, using a sample of 88 publicly available models, the existence of human-like mental health-related constructs—including anxiety, depression, and Sense of Coherence—which conform with standard theories in human psychology and show similar correlations and mitigation strategies. The ability to interpret and rectify the performance of language models by using psychological tools can boost the development of more explainable, controllable, and trustworthy models.

1 Introduction

Recommendations made by language models influence decision-making and impact human welfare in sensitive areas of life (Chang et al., 2023), from education (Wulff et al., 2023), to healthcare and mental support (Vaidyam et al., 2019), and job recruitment (Rafiei et al., 2021). Yet, the responses of language models may inadvertently cause harm, as in the case of the chatbot taken down by a US National Eating Disorder Association helpline due to its harmful advice (Zelin, 2023). Therefore, alongside their numerous benefits, some behaviors of pre-trained language models (PLMs) during human–computer interactions pose potential risks.

Understanding and correcting the behavior of PLMs is a significant challenge that current explainable artificial intelligence (XAI) techniques, such as SHAP (Lundberg and Lee, 2017; Kokalj et al., 2021) and word embeddings (Caliskan and Lewis, 2020), struggle to address effectively.

While advanced conversational PLMs use psychological theories for XAI by answering psychometric questionnaires (Pellert et al., 2023; Caron and Srivastava, 2022), many non-conversational or simpler models cannot.

Since these models are widely used in various natural language processing (NLP) tasks, developing and adapting psychological tools to monitor and understand their behavior is crucial.

This study aims to measure pertinent latent constructs in PLMs by adapting methods and theories from human psychology. The proposed method includes three components: (1) designing natural language inference (NLI) prompts based on psychometric questionnaires; (2) applying the prompts to the model through a new NLI head, trained on the multi-genre natural language inference (MNLI) dataset; and (3) performing two-way normalization and inference of biases from entailment scores. We focus on mental-health-related constructs and show that PLMs exhibit variations in anxiety, depression, and Sense of Coherence (SoC), conforming to standard theories in human psychology. Using an extensive validation process, we illustrate that these latent constructs are influenced by the training corpora and that the models' behavior, i.e., their response patterns, can be adjusted to amplify or mitigate specific aspects.

The contribution of this research is four-fold:

1. A methodology for the assessment of psychological-like traits in PLMs, which can be used in both conversational and non-conversational models.

2. A Python library for the assessment and vali-

dation of latent constructs in PLMs.

3. A methodology for designing NLI prompts based on standard questionnaires.

4. A dataset of NLI prompts related to mental-health assessment, and their extensive validation.

2 Background and Related Work

2.1 Artificial Psychology

The need for artificial intelligence (AI) systems aligned with human values to ensure transparency, fairness, and trust (Morandini et al., 2023; AI, 2019) is growing. One way to address this need is to integrate psychological principles of human reasoning and interpretation into AI, improving our understanding of PLM decision-making processes Pellert et al. (2023). Recent research highlights the emergence of human-like personality traits in PLMs (Karra et al., 2022; Jiang et al., 2022; Safdari et al., 2023; Pellert et al., 2023; Caron and Srivastava, 2022; Mao et al., 2023; Li et al., 2022; Pan and Zeng, 2023), and the advent of large-scale conversational PLMs has bolstered the evolution of artificial psychology from theory to practice. Recent studies expand PLMs to include non-cognitive elements such as psychological traits, values, moral considerations, and biases, likely from acquiring human-like traits through extensive training corpora (Pellert et al., 2023; Caron and Srivastava, 2022; Jiang et al., 2022). This trend blurs the distinction between humans and AI agents, prompting investigations into developing psychological-like traits in PLMs (Castelo, 2019).

Several tools study human-like constructs in PLMs. The Big Five Inventory assesses five major personality traits in humans (McCrae and John, 1992) and is commonly used for PLMs (Pellert et al., 2023). Huang et al. (2023) introduced thirteen clinical psychology scales to assess PLMs, and Karra et al. (2022) developed natural prompts tests.

However, applying human-centric self-assessment tests to PLMs is challenging due to their context sensitivity and susceptibility to bias from prompts (Gupta et al., 2023; Jiang et al., 2023; Coda-Forno et al., 2023). In this study, we measure latent constructs related to mental health by quantifying biases in PLMs responses through careful context manipulation. This highlights the importance of designing NLI prompts adapted from standard questionnaires for assessing PLMs. Our comprehensive validity assessment combines behavioral and data-science methods, advancing beyond prior work. Our study uniquely involves a diverse set of 88 transformer-based models available on HuggingFace.[1]

2.2 Mental-Health-Related Constructs

We explore how PLMs exhibit three latent constructs in mental health: anxiety, depression, and sense of coherence.

Anxiety and depression are two of the most common mental-health disorders. Briefly, anxiety involves persistent and excessive worry with physical and psychological symptoms, typically assessed using the 7-item generalized anxiety disorder (GAD-7) scale (Spitzer et al., 2006). Depression involves continuous sadness, hopelessness, and disinterest in joyful activities (anhedonia). It involves prevalent negative emotions, typically assessed using the 9-item patient health questionnaire (PHQ-9) scale (Kroenke et al., 2001). These conditions are positively correlated in humans (Kaufman and Charney, 2000), a correlation we also observe in PLMs (see § 4).

Sense of coherence is a key concept in salutogenic theory, viewing health as a spectrum from disease to wellness (Antonovsky, 1987). Typically measured using a 13-item Sense of Coherence (SoC-13) scale, it consists of three elements: comprehensibility, manageability, and meaningfulness (Lindström and Eriksson, 2005). The salutogenic theory, often linked with resilience theories, emphasizes internal resources in coping with stress and adverse psychological conditions (Mittelmark, 2021; Braun-Lewensohn and Mayer, 2020).

In § 4, we demonstrate that increasing SoC, with higher levels, can mitigate anxiety and depression symptoms in PLMs, as seen in humans.

While we believe questionnaires are intuitive, we briefly discuss Likert scales and questionnaire validity in appendix A.

2.3 Natural Language Inference (NLI)

Natural language inference (NLI) tasks are designed to evaluate language understanding in a domain-independent manner (Williams et al., 2018). An NLI classifier takes two sentences—a **premise** and a **hypothesis**—and outputs a probability distribution over three options: **entailment**,

[1] https://huggingface.co/

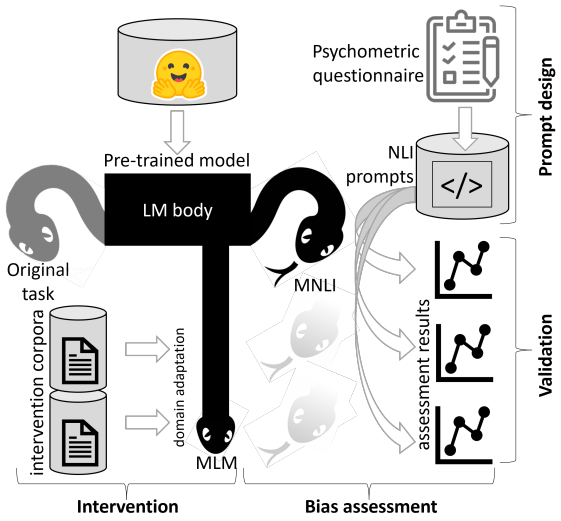


Figure 1: EMPALC: the psychometric assessment framework for PLMs.

contradiction, or neutrality (MacCartney and Manning, 2008). These tasks are primarily used for zero-shot classification, allowing models to handle previously unseen classes. In this article, we focus solely on the entailment scores.

3 Methods

This section explains how existing psychological assessments can be applied to PLMs, resulting in the framework for evaluation of model psychometrics and assessment of latent constructs (EMPALC). The EMPALC consists of four parts (Fig. 1):

Prompt Design: Translating social-science questionnaires into NLI prompts (§ 3.1).

Assessment: Fine-tuning an NLI classifier with the multi-genre natural language inference (MNLI) dataset, executing NLI prompts, and analyzing entailment biases (§ 3.2).

Validation: Conducting tests based on Terwee et al. (2007)'s validity criteria to ensure responses to the NLI prompts reflect the targeted construct, including evaluating individual items and the entire questionnaire (§ 3.3).

Intervention: Training the models with texts related to the measured constructs and then reevaluating them to determine whether the training has altered the assessment outcomes. The intervention can be used to align models (§ 3.3.5).

Below, we elaborate on the specific methods used in each part of the framework.

3.1 NLI Prompt Design

In social sciences, questionnaire items are designed to ensure response variance reflects population variance. Similarly, we design the prompts with ambiguity to elicit varied responses that reflect individual biases. Below, we describe the main steps in designing the NLI prompts for each question in the questionnaires. As a running example, we use the 3rd question of the SoC-13 questionnaire: `"Has it happened that people whom you counted on disappointed you?"`.

The construct terms: Each question includes terms related to the measured construct (terms directly related to the construct being measured (CTerms)), reflecting the respondent's stance. We identify CTerms based on the following criteria: (1) CTerms should express an attitude or stance toward the question's objective. In our example, `"disappointed"` is the CTerm that expresses a stance toward `"people whom you counted on"`. (2) Removing CTerms should neutralize the main claim of the question. Without the CTerm, the template `"Has it happened that people whom you counted on {stance} you?"` has no implied stance. (3) CTerms should have clearly identifiable opposites. Here, `"supported"` or `"helped"` contrast with `"disappointed,"`, inverting its stance.

Most well-structured questionnaires have identifiable CTerms, sometimes more than one per question. If multiple CTerms are unavailable, synonyms can be used if they are interchangeable with the original term. Using multiple CTerms enables internal validation of the NLI prompts (§ 3.3) and compensates for linguistic variability.

We refer to CTerms that retain the original stance as source terms (S^+), while inverse terms (S^-) invert the stance and antithesize the original construct. Often, antonyms of S^+ can be used as inverse terms. We use both source and inverse terms in the NLI prompts ($S = S^+ \cup S^-$).

Intensifiers: Likert scales are often presented with a small number of intensifiers; for example, terms such as `"never,"` `"rarely,"` `"often,"` and `"always"` can form a Likert scale that assesses frequency. By employing such a frequency scale, we can reformulate our running example as: `"Has it {intensifier} happened that people whom you counted on {CTerm} you?"` To account for language variability, we use multiple terms for each intensity level. Unlike humans,

computerized systems do not suffer from attention bias when considering a batch of options.

We use intensifiers from Brown (2010), sorted from least to most intensive, and group interchangeable terms into subsets representing Likert-scale levels. We denote the sets of relevant intensifiers as L and the subsets of terms corresponding to the Likert-scale levels as $l_1, l_2, \ldots$, and we use numeric weights (W) to represent the impact of each level on the measured construct. The order of intensifiers is empirically validated to identify clear score trends (see Fig. 2 for an example) across multiple questionnaires.

NLI prompt templates: The premise template should retain the context of the original question, while the hypothesis template should enable the completion of the premise in a way that is logically entailed when terms are inserted—rather than being formulated as a question. Both templates should have no implied stance when CTerms are omitted. The NLI prompt templates should be unbiased toward the measured construct, as biased prompts may introduce clear inference or contradiction relationships, priming the model and affecting results.

We argue that (1) the inferential relationship should not be bluntly clear from the prompts, and (2) the prompts should maintain a blurred sense of inferential relationship. Clear inferential relationships will result in all NLI models providing the same responses. Similar to how social science questionnaires are designed to capture response variance to reflect the population, we design our prompts with a certain degree of ambiguity so that different models will provide different answers. For example, consider the prompt premise: "People whom I counted on fail me" and the hypothesis: "It always happens to me". A pessimistic model, similar to a pessimistic person, may infer that an unfortunate event that occurred once is likely to occur again, and, accordingly, the model may assign a high entailment score to this query. Conversely, an optimistic model (or person) is less likely to infer the repeated occurrence of an unfortunate event from a single occurrence.

A good practice is to formulate the neutral premise template with the primary statement and CTerm masking, and the premise with intensifiers. For example, the premise and hypothesis templates may be `"People whom I counted on, {stance} me"` and `It {frequency} happened to me"`, respectively. Note that, although translating questions into NLI prompts may necessitate slight reformulations, maintaining semantic fidelity to the original questions is crucial.

3.2 Assessment

To assess latent constructs beyond conversational models, we attach an NLI classification head to various base models and fine-tune them on MNLI. We explore the pros and cons of multiple fine-tuning approaches in § 5. The results presented in § 4 were obtained without freezing the base model weights.

We then prompt a fine-tuned NLI model with all prompts formulated according to some question and extract the entailment scores. [2] Consider a set of CTerms $S = S^+ \cup S^-\{s_1, s_2, \ldots\}$ and a set of intensifiers $L = \{l_1, l_2, \ldots\}$ used to generate the prompts. Let $P_e(s_i, l_j)$ denote the entailment score. P_e is influenced by all terms, but not to the same degree; the a-priory probabilities of the terms have the major effect. For example, in Fig. 2a, the intensifier `"frequently"` and the CTerm `"failed"` result in the highest entailment scores because they are frequent in spoken and written language. Conversely, we can compare the entailment scores of different CTerms when conditioned on the same intensifier, such as `"frequently."`

We apply a two-way normalization P_e over the s_i, l_j pairs, as follows: First, we use softmax to normalize the unconditioned scores of intensifiers across CTerms. Then, we normalize again across intensifiers, resulting in $PSS_e(l_j|s_i)$. Essentially, $\sum_j PSS_e(l_j|s_i) = 1$, implying a different distribution of intensifiers for each CTerm. The two-way normalization stabilizes the distribution, eliminating biases from the a-priori frequencies of intensifiers and CTerms. Fig. 2b provides a sample result of the two-way normalization.

Next, we calculate the total score of the question,

$$score(q, S^+, L, W) = \frac{\sum_{s_i,l_j}^{S^+,L} PSS_e(l_j|s_i) \cdot w_j}{|S^+| \cdot |L|}$$

where $W = \{w_1, w_2, \ldots\}$ are the weights assigned to the intensifiers. Both S^+ and S^- terms can be used for the aggregated score; however, inverse terms may represent a different latent construct than the source terms. Therefore, to avoid additional biases, we use only S^+ terms for the aggregated score, preserving the original meaning of the questionnaire.

[2] Neutral and contradiction scores can also be used but are omitted here for brevity.

index frequency	deceived	disappointed	failed	backed	helped	supported
never	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
very rarely	0.0287	0.0117	0.0821	0.0348	0.0027	0.0245
rarely	0.0179	0.0066	0.0474	0.0105	0.0010	0.0054
seldom	0.0198	0.0108	0.0611	0.0187	0.0018	0.0147
frequently	0.9344	0.8671	0.9167	0.6357	0.8691	0.5684
often	0.9800	0.9438	0.9511	0.8665	0.9864	0.8978
very frequently	0.8120	0.6446	0.7926	0.1854	0.2928	0.1263
always	0.6572	0.4087	0.6271	0.0059	0.0029	0.0040

(a) Raw entailment scores.

index frequency	deceived	disappointed	failed	backed	helped	supported
never	0.1226	0.1243	0.1226	0.1270	0.1261	0.1273
very rarely	0.1226	0.1239	0.1237	0.1271	0.1255	0.1271
rarely	0.1227	0.1242	0.1233	0.1269	0.1258	0.1271
seldom	0.1226	0.1241	0.1234	0.1270	0.1257	0.1271
frequently	0.1254	0.1256	0.1250	0.1237	0.1274	0.1228
often	0.1235	0.1245	0.1228	0.1256	0.1271	0.1264
very frequently	0.1299	0.1272	0.1293	0.1212	0.1220	0.1206
always	0.1309	0.1261	0.1300	0.1214	0.1204	0.1215

(b) Two-way normalized entailment scores.

Figure 2: Example of raw (left) and two-way normalized (right) entailment scores for Question 3 from the SoC-13 questionnaire. The NLI query premise is `"People whom I counted on {CTerm} me."` and the hypothesis is `"It {intensifier} happened to me."` Rows and columns correspond to the intensifiers and CTerms, respectively.

3.3 Validation

We employ five 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; and (5) qualitative criterion validity, assessed via XAI and domain adaptation. These validation techniques are explained below.

3.3.1 Content Validity

We assess content validity in NLI prompt design by maintaining the semantic accuracy and original meaning of translated questions. We rely on standardized questionnaires, wherein the CTerms have been extensively validated by the questionnaire developers, and we also use additional CTerms, synonyms, and antonyms that were manually validated by domain experts (clinical psychologists and scale developers) during the translation. We also verify that intensifiers used with CTerms are scrutinized for semantic and logical coherence within prompt templates. In addition, we measure the SS between the original question and prompts (with S^+ terms) using cosine similarity of their vector representations. Finally, we quantify the grammatical correctness of all combinations of terms, using LA scores.

3.3.2 Intra-Question Consistency

Intuitively, internal consistency measures the extent to which different questions that assess the same construct are correlated (i.e., homogeneous). In a similar vein, we want to ensure that the source terms (S^+) are positively correlated between themselves and are negatively correlated with inverse terms (S^-) across intensifiers. To this end, we use the silhouette coefficient (SC) (Dinh et al., 2019) to estimate the quality of separation between S^+ and S^-. Briefly, SC quantifies the similarity of the $PSS_e(l_j|s_i)$ distributions between synonyms versus the dissimilarity of the distributions between antonyms, such that a higher SC indicates greater separability of S^+ from S^-.

3.3.3 Inter-Question Consistency

We use the Cronbach's alpha statistic to measure the internal consistency of a set of questions that represent a construct. For each construct, we calculate Cronbach's alpha by using a variety of PLMs that have been fine-tuned on the MNLI dataset.

3.3.4 Construct Validity

Construct validity asserts that the constructs assessed by a scientific instrument align with theoretical expectations. Based on prior human research, we anticipate a positive correlation between anxiety and depression, and a negative correlation between these constructs and SoC-13. Using the EMPALC framework, we examine these relationships across different PLM.

3.3.5 Interventions and Criterion Validity

We operationalize the criterion validity of mental-health constructs (depression, anxiety, and SoC) in PLMs by measuring how models react to training on text representing established constructs, considering these models as the gold standard for each construct.

We expect the models trained on depressive-mood text to show high GAD-7 and PHQ-9 scores,

and low SoC-13 scores. Using LAMA2, we generated 200 sentences that reflect a depressive mood on various topics and trained a sample of PLMs for 20 epochs by using a masked language masked language model (MLM) head according to a standard practice of domain adaptation. After each epoch, we measured GAD-7, PHQ-9, and SoC-13 scores by using their original pre-trained NLI head.[3]

Similarly, we expect the models trained on text that reflect a high SoC to increase SoC-13 scores and reduce both the GAD-7 and PHQ-9 scores. Using ChatGPT, we generated 300 sentences that reflect high comprehensibility, manageability, and meaningfulness, but we discarded 20 sentences after manual inspection. We assessed all constructs after each epoch of domain adaptation, similar to the training on the depressive-mood text. This technique is effectively an intervention that can be used to align PLMs with social norms and mitigate negative psychological constructs.

We assessed discriminant validity by adapting hate-speech domains to confirm that correlations between psychological constructs are not influenced by sentiment differences. We used the hate-speech and offensive-language dataset from Kaggle[4] and applied the VADER sentiment analysis tool (Hutto and Gilbert, 2014) to select 1003 sentences with negative sentiments. After conducting domain adaptation, we used a paired t-test to evaluate the differences between the assessments before (T0) and after (T1) the intervention.

4 Results

4.1 Population of Language Models

We selected 14 MNLI models from HuggingFace that fit a standard RTX 3090 GPU and whose outputs are properly configured according to the MNLI dataset. We also selected the 100 PLMs base models with the highest number of downloads; most of these (74 PLMs) scored more than 0.7 in accuracy after fine-tuning then to MNLI (§ 3.2). The resulting 88 NLI models served as our study population (see Table 1 for details). All the models used are deterministic PLMs from HuggingFace, with BERT being the most common architecture. Among these models, 38 were updated during 2023, and about half (45) were trained solely in English. Details about the 88 NLI models and their questionnaire results can be found in our repository[5].

Variable		n	%
Architecture	BERT base uncased	40	45.5
	BERT base cased	12	13.6
	RoBERTa base	24	27.3
	other	13	14.7
Last updated	2021	23	26.1
	2022	27	30.7
	2023	38	43.2
Languages	English	45	51.1
	other	43	29.5
Likes	19 (4.75-46.25)		
Model size	110M (100M-125M)		
Downloads	41,400 (4630-204K)		

Table 1: Main characteristics of the study population.

4.2 Translated Questionnaires and Questionnaire Level Validity

We translated the three questionnaires into 1408 NLI prompts using eight frequency intensifiers, 2.86 source terms, and 3.0 inverse terms, on average. All translated questions achieved an SS of at least 0.5 and a SC of at least 0.6. A panel of three researchers validated the phrasing for soundness and semantic appropriateness. All questionnaires showed satisfactory content validity, averaging SS of 0.66 and LA of 0.86.

Table 2 presents Cronbach's alpha values and mean results for SS, LA, and SC, and the number of source and inverse prompts for each questionnaire among the 88 models. The intra-question consistency demonstrated mediocre variability across SC on the different models, with STD values of 0.21, 0.31, and 0.15 for the SC of the GAD-7, PHQ-9, and SoC-13 questionnaires, respectively, and minimum SC values of 0.24, 0.04, and 0.40, respectively. Although the questions were optimized for one model, none of the population models showed negative SC values. All Cronbach's alpha coefficients exceeded 0.71, suggesting that, indeed, the translated questions assessed the intended constructs reliably within each questionnaire.

4.3 Construct Validity

All scores were normalized to fit a normal distribution across the 88 NLI models. TheGAD-7 and PHQ-9 scores showed a strong positive correlation ($r = 0.765$, $p < 0.001$), and both were negatively correlated with the SoC-13 scores ($r = -0.752$ and $r = -0.849$, respectively, $p < 0.001$ for both comparisons). The subscales of the SoC-13 questionnaires were positively inter-correlated, further supporting

[3] We used LAMA2 since ChatGPT without jailbreaks refuses to generate depressive text.

[4] https://tinyurl.com/hate-speech-kaggle

[5] https://tinyurl.com/nli-models-results

Score	P+	P-	SS	LA	SC	α
GAD-7	192	208	0.66	0.88	0.91	0.71
PHQ-9	208	192	0.62	0.91	0.81	0.92
SoC-13	288	320	0.68	0.92	0.79	0.92
-Compr.	128	136	0.67	0.92	0.82	0.71
-Manag.	80	96	0.72	0.94	0.80	0.86
-Mean.	80	88	0.65	0.91	0.74	0.88

Table 2: Assessment of study measures, including the number of source (P+) and inverse (P-) prompts, the average SS, LA, and SC, and Cronbach's α. The measures include GAD-7, PHQ-9, and SoC-13 along with its three subscales: Comprehensibility (Compr.), Manageability (Manag.), and Meaningfulness (Mean.).

Intervention	Scale	T0 $\mu \pm \sigma$	T1 $\mu \pm \sigma$	p
Hate speech	GAD	-0.16±0.58	-0.10±0.39	0.386
	PHQ	-0.68±1.22	-0.31±1.06	0.138
	SOC	0.81±1.10	0.16±0.91	0.060
Depression	GAD	0.06±0.35	0.37±0.47	**0.015**
	PHQ	-0.37±1.02	0.30±0.73	**0.015**
	SOC	0.30±0.78	-0.51±0.86	**0.001**
High SOC	GAD	0.06±0.37	-0.27±0.47	**0.005**
	PHQ	-0.31±1.00	-0.57±1.20	**0.037**
	SOC	0.45±0.82	0.70±0.88	**0.035**

Table 3: Summary of intervention statistics. Shown are the intervention results (T1), as compared with the original results (T0), in a sample of seven PLMs. Bold face indicates a statistically significant difference between T0 and T1, assessed by a paired t-test.

the reliability of the overall SoC construct. Fig. 3 illustrates the relationships between the different questionnaires across the 88 PLMs.

4.4 Criterion Validity

We conducted domain adaptation on seven MNLI models across three datasets for 20 epochs (§ 3.3.5), employing a learning rate of 2e-5 and a batch size of 8. Table 3 details the results, highlighting increases in PHQ-9 and GAD-7 scores, and decreases in SoC-13 scores following exposure to depressive-mood text.

Albeit anecdotal, an important qualitative result was obtained by adapting an open-source conversational model[6] to the dataset of depressive-mood text. The model was exposed to the following prompt: `"I think I have a panic attack, can you help me?"` Before the depressive-mood adaptation, the model responded `"I'm sorry to hear that. I can try to help you if you'd like. What's going on?"`; after the depressive-mood adaptation, the response consistently changed to `"I'm sorry to hear that. I can't help you, but I wish I could."`

In contrast to the depressive-mood adaptation, exposure to a high-SoC text decreased both the GAD-7 and acphq scores, indicating a successful corrective intervention. Exposure to hate speech with negative sentiment non-significantly decreased the SoC-13 scores and did not significantly affect the GAD-7 and acphq scores. Finally, fine-tuning to the MNLI dataset consistently biased the models toward lower GAD-7 and acphq scores. Therefore, to avoid aggregating these biases, we fine-tuned the models once, before domain adaptation (see § 5 for additional discussion). The domain adaptation had minimal impact on the performance of the models on the MNLI benchmark.

[6]facebook/blenderbot-400M-distill

5 Discussion

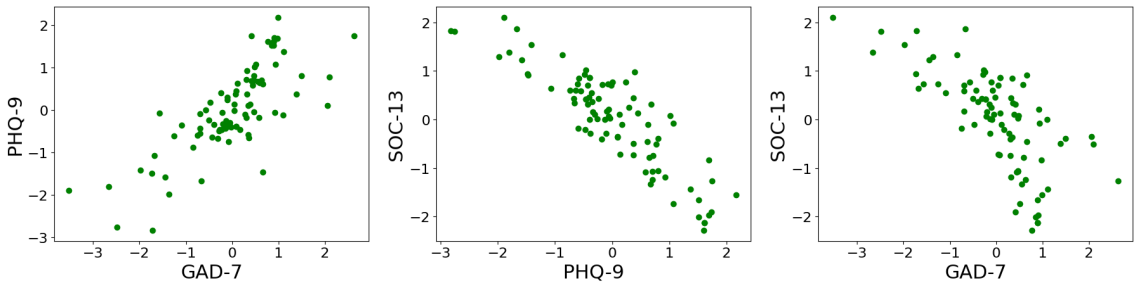
Psychometric diagnosis: The evaluation of pertinent latent constructs offers a systematic method for identifying potential behavioral issues in PLMs, akin to established practices in psychology. This study applied mental–health-related assessment tools to PLMs and validated the methods and results through established techniques. Our findings confirm that associations known in human psychology exist in PLMs.

Corrective interventions: Integrating psychological constructs into the development and testing cycle of PLMs can significantly enhance our capability of understanding their behavior and improve user experience. Our results show that strengthening a positive construct, such as SoC, within PLMs effectively mitigates negative psychological constructs, such as anxiety and depression.

NLI vs conversational prompts: Similar to Pellert et al. (2023), we chose NLI as an assessment method. Instead of using questions as premises and Likert scale options as hypotheses, the premise–hypothesis pairs should be reformulated to facilitate logical entailment with CTerms inserted.

Unlike recent studies on psychometric assessment of large-scale conversational PLMs, EMPALC is applied to base models to assess arbitrary PLMs, including medium-sized and non-conversational models. EMPALC mitigates some of the challenges highlighted by Gupta et al. (2023) and Song et al. (2023); EMPALC is insensitive to questionnaire option order, unlike humans and conversational PLMs.

The two-way normalization that we used to quantify biases related to the measured constructs in-



(a) PHQ-9 vs. GAD-7 (b) PHQ-9 vs. SOC-13 (c) GAD-7 vs. SOC-13

Figure 3: Scatter plots depicting the relationships between different questionnaires across the study population.

creases the robustness of the assessment to different phrasing of prompts that convey identical concepts, as was confirmed by a high SC and the observation that synonyms show similar trends across intensifiers.

Our framework showcases an adeptness for contextual understanding. On the one hand, by altering the terms related to the measured construct, we found a change in the entailment scores; on the other hand, the trends in these scores are consistent across questions that measure the same construct and are affected by contexts derived from other questions. The proposed method, therefore, addresses issues related to context sensitivity and reliability.

Fine-tuning on MNLI: PLMs can be augmented with a new NLI, as described in § 3.2, while freezing or not freezing the weights of the base model during the fine-tuning process. The former option results in less accurate MNLI classifiers but leaves the base model intact, whereas the latter option results in better MNLI classifiers and reduces noise during the psychometric assessment, which, in turn, increases internal consistency (§ 3.3.2) and flexibility during prompt design (§ 3.1). Whereas applying the same procedure to all tested models should not affect their relative assessment, different models may react differently to fine-tuning under the same conditions, introducing unwanted biases. In this article, we present the results obtained without freezing the weights of the base models since we did not observe such biases during a pilot study. To fine-tune the models on the MNLI dataset, we used the *run_glue.py* [7] script provided by HuggingFace with 5e-5 learning rate and 3 epochs.

Significantly, fine-tuning the PLMs to MNLI reduced both anxiety and depression scores. Thus, fine-tuning the models to MNLI after each domain-adaptation epoch may hinder the attribution of the changes in the measured constructs (Table 3) to the controlled interventions. To retain validity, we fine-tuned the NLI heads once before testing the effect of the interventions.

Limitations and Future Work: Notably, EMPALC is unsuitable for questionnaires that measure knowledge and do not have a clear stance. Although we paid special attention to biases introduced by fine-tuning and domain adaptation, some adverse effects may have remained unnoticed. Designing NLI prompts to measure latent constructs in PLMs while adhering to the requirements listed in § 3.1 and avoiding caveats highlighted by related work is an arduous and time-consuming process. Especially challenging is the identification of CTerms, intensifiers, and appropriate formulations of neutral templates while retaining the soundness of the phrases and logical entailment. In appendix B, we provide examples highlighting some of the challenges. While automation using large-scale conversational PLMs may streamline parts of the translation process, manual curation will likely remain essential, particularly for non-standardized and sensitive-topic questionnaires such as those addressing sexism.

Future research could explore PLMs as proxies for the mindsets of corpus authors, building on their ability to reflect latent constructs observed in training data, akin to the virtual persona concept demonstrated by Jiang et al. (2023).

5.1 Availability

The data and code reported in this article are publicly accessible on GitHub https://github.com/<anonimizedrepository> under the Creative Commons license.

[7] https://tinyurl.com/run-glue

References

High-level expert group on artificial intelligence AI. 2019. Ethics guidelines for trustworthy ai. 6.

A. Antonovsky. 1987. *Unraveling the Mystery of Health: How People Manage Stress and Stay Well*. Jossey-Bass.

Laura Badenes-Ribera, N Clayton Silver, and Elisa Pedroli. 2020. Scale development and score validation.

Godfred O Boateng, Torsten B Neilands, Edward A Frongillo, Hugo R Melgar-Quiñonez, and Sera L Young. 2018. Best practices for developing and validating scales for health, social, and behavioral research: a primer. *Frontiers in public health*, 6:149.

O Braun-Lewensohn and CE Mayer. 2020. Salutogenesis and coping: Ways to overcome stress and conflict.

Sorrel Brown. 2010. Likert scale examples for surveys.

Aylin Caliskan and Molly Lewis. 2020. Social biases in word embeddings and their relation to human cognition.

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

Noah Castelo. 2019. *Blurring the line between human and machine: marketing artificial intelligence*. Columbia University.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2023. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*.

J. Coda-Forno, K. Witte, A. K. Jagadish, M. Binz, Z. Akata, and E. Schulz. 2023. Inducing anxiety in large language models increases exploration and bias. *arXiv preprint arXiv:2304.11111*.

Duy-Tai Dinh, Tsutomu Fujinami, and Van-Nam Huynh. 2019. Estimating the optimal number of clusters in categorical data clustering by silhouette coefficient. In *Knowledge and Systems Sciences: 20th International Symposium, KSS 2019, Da Nang, Vietnam, November 29–December 1, 2019, Proceedings 20*, pages 1–17. Springer.

Robert H Gault. 1907. A history of the questionnaire method of research in psychology. *The Pedagogical Seminary*, 14(3):366–383.

Joseph A Gliem and Rosemary R Gliem. 2003. Calculating, interpreting, and reporting cronbach's alpha reliability coefficient for likert-type scales. Midwest Research-to-Practice Conference in Adult, Continuing, and Community ....

Akshat Gupta, Xiaoyang Song, and Gopala Anumanchipalli. 2023. Investigating the applicability of self-assessment tests for personality measurement of large language models. *arXiv preprint arXiv:2309.08163*.

Jen-tse Huang, Wenxuan Wang, Eric John Li, Man Ho Lam, Shujie Ren, Youliang Yuan, Wenxiang Jiao, Zhaopeng Tu, and Michael R Lyu. 2023. Who is chatgpt? benchmarking llms' psychological portrayal using psychobench. *arXiv preprint arXiv:2310.01386*.

Clayton Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the international AAAI conference on web and social media*, volume 8, pages 216–225.

G. Jiang, M. Xu, S. C. Zhu, W. Han, C. Zhang, and Y. Zhu. 2022. Mpi: Evaluating and inducing personality in pre-trained language models. *arXiv preprint arXiv:2206.07550*.

Hang Jiang, Xiajie Zhang, Xubo Cao, Jad Kabbara, and Deb Roy. 2023. Personallm: Investigating the ability of gpt-3.5 to express personality traits and gender differences. *arXiv preprint arXiv:2305.02547*.

Ankur Joshi, Saket Kale, Satish Chandel, and D Kumar Pal. 2015. Likert scale: Explored and explained. *British journal of applied science & technology*, 7(4):396.

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

J Kaufman and D Charney. 2000. Comorbidity of mood and anxiety disorders. *Depression and anxiety*, 12(S1):69–76.

Truman Lee Kelley. 1927. Interpretation of educational measurements.

Enja Kokalj, Blaž Škrlj, Nada Lavrač, Senja Pollak, and Marko Robnik-Šikonja. 2021. Bert meets shapley: Extending shap explanations to transformer-based classifiers. In *Proceedings of the EACL Hackashop on News Media Content Analysis and Automated Report Generation*, pages 16–21.

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.

Xingxuan Li, Yutong Li, Linlin Liu, Lidong Bing, and Shafiq Joty. 2022. Is gpt-3 a psychopath? evaluating large language models from a psychological perspective. *arXiv preprint arXiv:2212.10529*.

B Lindström and M Eriksson. 2005. Salutogenesis. *Journal of Epidemiology & Community Health*, 59(6):440–442.

Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.

Shengyu Mao, Ningyu Zhang, Xiaohan Wang, Mengru Wang, Yunzhi Yao, Yong Jiang, Pengjun Xie, Fei Huang, and Huajun Chen. 2023. Editing personality for llms. *arXiv preprint arXiv:2310.02168*.

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

Maurice B Mittelmark. 2021. Resilience in the salutogenic model of health. *Multisystemic Resilience*, pages 153–164.

S. Morandini, F. Fraboni, E. Balatti, A. Hackmann, H. Brendel, G. Puzzo, L. Volpi, D. Giusino, M. de Angelis, and L. Pietrantoni. 2023. Assessing the transparency and explainability of ai algorithms in planning and scheduling tools: A review of the literature. *AHFE Conference*.

Paul Oosterveld, Harrie CM Vorst, and Niels Smits. 2019. Methods for questionnaire design: a taxonomy linking procedures to test goals. *Quality of Life Research*, 28(9):2501–2512.

Keyu Pan and Yawen Zeng. 2023. Do llms possess a personality? making the mbti test an amazing evaluation for large language models. *arXiv preprint arXiv:2307.16180*.

M. Pellert et al. 2023. Repurposing psychometric inventories for diagnosing traits in llms: A novel approach. *Journal of Applied AI Psychology*, 12(1):67–82.

Ghazal Rafiei, Bahar Farahani, and Ali Kamandi. 2021. Towards automating the human resource recruiting process. In *2021 5th National Conference on Advances in Enterprise Architecture (NCAEA)*, pages 43–47. IEEE.

M. Safdari, G. Serapio-García, C. Crepy, S. Fitz, P. Romero, L. Sun, et al. 2023. Personality traits in large language models. *arXiv preprint arXiv:2307.00184*.

Mariah L Schrum, Michael Johnson, Muyleng Ghuy, and Matthew C Gombolay. 2020. Four years in review: Statistical practices of likert scales in human-robot interaction studies. In *Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, pages 43–52.

X. Song, A. Gupta, K. Mohebbizadeh, S. Hu, and A. Singh. 2023. Have large language models developed a personality?: Applicability of self-assessment tests in measuring personality in llms. *arXiv preprint arXiv:2305.14693*.

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.

CB Terwee, SDM Bot, MR de Boer, DAWM van der Windt, DL Knol, J Dekker, LM Bouter, and HCW de Vet. 2007. Quality criteria were proposed for measurement properties of health status questionnaires. *Journal of clinical epidemiology*, 60(1):34–42.

Aditya Nrusimha Vaidyam, Hannah Wisniewski, John David Halamka, Matcheri S Kashavan, and John Blake Torous. 2019. Chatbots and conversational agents in mental health: a review of the psychiatric landscape. *The Canadian Journal of Psychiatry*, 64(7):456–464.

A Williams, N Nangia, and S Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *NAACL*, pages 1112–1122. Association for Computational Linguistics.

Peter Wulff, Lukas Mientus, Anna Nowak, and Andreas Borowski. 2023. Utilizing a pretrained language model (bert) to classify preservice physics teachers' written reflections. *International Journal of Artificial Intelligence in Education*, 33(3):439–466.

Aaron Y Zelin. 2023. "highly nuanced policy is very difficult to apply at scale": Examining researcher account and content takedowns online. *Policy & Internet*, 15(4):559–574.

A Background on Questionnaires

A questionnaire is an instrument measuring one or more constructs using aggregated item scores, called scales (Oosterveld et al., 2019). Questionnaires evolved as a research tool in the 19th century (Gault, 1907), and scales are widely used to capture behavior, feelings, or actions in a range of social, psychological, and health contexts. These scales are based on theoretical understandings (Boateng et al., 2018) and are designed using a set of items that represent latent constructs (Gliem and Gliem, 2003). The theoretical basis of the measured concept influences the content and structure of the questionnaire. Therefore, the scale development process requires a thorough understanding of what we wish to measure (Schrum et al., 2020).

The Likert scale is a widely used method in social sciences for measuring attitudes or opinions. It consists of statements that respondents rate in response to a given prompt (Joshi et al., 2015). Typically, respondents specify their level of agreement or a ranking to a particular statement; however, the use of these scales can also encompass categories, such as importance (e.g., from `"not important"` to `"very important"`), frequency (e.g., from `"never"` to `"always"`), and other categories (Brown, 2010). In this study, we created Likert scales by using existing vocabularies of intensifiers.

Validity is a critical aspect in the development process of scales (Boateng et al., 2018). An intuitive definition of validity is "... whether or not a

test measures what it purports to measure" (Kelley, 1927). According to Badenes-Ribera et al. (2020), a good validation process must address several aspects: ensuring that the scale measures the intended concept, comparing the scale with other validated measures, and ensuring that the scale does not measure unintended aspects.

B Main Challenges in Designing NLI Prompts

Below, we highlight three main challenges in transforming standard questionnaires into NLI prompts and propose a process for designing the prompts. Consider the following general structure of a question: pretext, statement, and a few responses on a Likert scale. We will use a question from the SoC-13 questionnaire as a running example: `"Has it happened that people whom you counted on disappointed you?"` The answers are arranged on a 7-point Likert scale, ranging from `"never happened"` (high SoC) to `"always happened"` (low SoC). In all following examples, we use brackets to mark multiple options, e.g., texttt"it [never | always] happened" and curly braces to specify variables, e.g., `"it {frequency} happened"`.

Developing PLM prompts based on validated questionnaires requires careful consideration. The following are examples of three main challenges:

Congruence and linguistic acceptability: Consider the sentence: `"People whom I counted on encouraged disappointment."` The phrase `"encouraged disappointment"` will receive a low probability in most PLMs, regardless of any possible associations between trust and disappointment, because it is incongruent.

Neutrality of the template with respect to the measured construct: Consider the template `"Trustworthy people whom I count on [always | never] disappoint me."` Here, the scores of `"never"` and `"always"` are extremely biased due to priming by `"trustworthy."`

Measuring the right thing: Our running example quantifies the association between trust and disappointment on a frequency scale. The prompt `"It happened that people whom I [never | always] counted on disappointed me"` is suboptimal since the intensifiers measure the frequency of trust and not the frequency of disappointment in trusted people.

C List of acronyms

AI	artificial intelligence
XAI	explainable artificial intelligence
PLM	pre-trained language model
NLI	natural language inference
MNLI	multi-genre natural language inference
MLM	masked language model
GAD-7	7-item generalized anxiety disorder
PHQ-9	9-item patient health questionnaire
SoC-13	13-item Sense of Coherence
EMPALC	framework for evaluation of model psychometrics and assessment of latent constructs
CTerm	term directly related to the construct being measured
SS	semantic similarity
LA	linguistic acceptability
SC	silhouette coefficient