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

Anonymous ACL submission

Abstract

Recent discoveries suggest that large language models demonstrate personality-like traits. This evidence suggests that known and yet undiscovered biases of language models conform to standard human-like latent psychological constructs. While large conversational models may be tricked into genuinely answering questionnaires, psychometric assessment methods are lacking for thousands of simpler transformers trained for other tasks. This article teaches how to reformulate psychological questionnaires into natural language inference prompts and provides a code library to support the psychometric assessment of arbitrary models. Experiments performed with a sample of 88 publicly available models demonstrate the existence of mental health-related constructs, such as anxiety, depression, and the sense of coherence. Extensive validation of the constructs reveals that they conform with standard theories in human psychology, including known correlations, and mitigation strategies. The ability to interpret and rectify the performance of language models using psychological tools will help to develop 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), such as education (Wulff et al., 2023), healthcare and mental support (Vaidyam et al., 2019), job recruiting (Rafiei et al., 2021). Under certain conditions, the responses of language models may inadvertently cause harm. Consider, for instance, the chatbot taken down by a US National Eating Disorder Association helpline due to its harmful advice (Zelin, 2023). Another case of a potentially harmful model is GPT-4chan (Gault, 2022), a GPT-J model trained on offensive language from the 4chan[1] forums. These examples highlight the potential risks associated with inappropriate behaviors of pre-trained language models (PLMs) in human-computer interactions.

Understanding and correcting the PLMs' behavior poses a significant challenge that current explainable artificial intelligence (XAI) techniques that SHAP (Lundberg and Lee, 2017; Kokalj et al., 2021) and word embeddings (Caliskan and Lewis, 2020) struggle to address effectively. Today, the most advanced PLMs are capable of answering psychometric questionnaires (Pellert et al., 2023; Caron and Srivastava, 2022) facilitating the use of the application of psychological theories for XAI. However, available psychological tools are not yet fully adapted for in-depth analysis of non-conversational or less sophisticated models.

The primary objective of this article is to measure pertinent latent constructs embedded within the base PLMs, using methods and theories from human psychology. The proposed method includes three main components: (1) design of 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) perform two-way normalization and inference of biases from entailment probabilities. In this study, we focus on mental-health-related constructs and show that PLMs exhibit variations in anxiety, depression, and sense of coherence (a 13-items scale) (SoC-13) that conform to standard theories in human psychology. Extensive validation illustrates that these latent constructs are influenced by the training corpora. Consequently, the behavior of models, i.e., their response patterns, can be adjusted to amplify or mitigate specific aspects of their behavior.

The contributions of this research are as follows:

[1] https://www.4chan.org

1. A methodology for assessment of psychological-like traits in PLMs that can be used in non-conversational models.

2. A Python library for assessing and validating 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 that align with human values, ensuring transparency, fairness, and trust, is growing (Morandini et al., 2023; AI, 2019). Integrating psychological principles related to human reasoning and interpretation into AI could advance these goals. Pellert et al. (2023) argue that such integration can lead to a better understanding of PLM decision-making processes. With the advent of large-scale conversational PLMs, artificial psychology has evolved from theory to practice. Recent studies have broadened this scope to include non-cognitive elements such as psychological traits, values, moral considerations, and biases (Pellert et al., 2023; Caron and Srivastava, 2022; Jiang et al., 2022). Pellert et al. (2023) attribute this shift towards non-cognitive aspects to the premise that PLMs acquire human-like psychological characteristics from their extensive training corpora. Castelo (2019) posits that the growing use of PLMs is blurring the distinctions between humans and AI agents, prompting inquiries into the possible development of personality traits in PLMs.

Recent research has highlighted 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). The Big Five Inventory (BFI), a well-established questionnaire for assessing five major personality traits in humans, is commonly used to evaluate PLMs (McCrae and John, 1992). Studies have also introduced a framework to assess PLMs' psychological dimensions using thirteen scales from clinical psychology (Huang et al., 2023). Karra et al. (2022) developed natural prompts for generative models to elicit personality traits without relying on human-centric self-assessment tests. While the existence of personality traits in PLMs is partially validated, direct application of human-centric self-assessment tests on PLMs often fails due to their context sensitivity and susceptibility to bias through prompts (Gupta et al., 2023; Jiang et al., 2023; Coda-Forno et al., 2023).

In this paper, we quantify biases in PLMs responses following careful context manipulation to measure latent constructs related to mental health.

We highlight the importance of carefully designing NLI prompts adapted from standard questionnaires for PLMs. Our comprehensive validity assessment combines behavioral and data science methods, advancing beyond prior work. Distinctively, our study involves a large and varied population of 88 transformer-based models available on HuggingFace.[2]

2.2 Mental-Health-Related Constructs

In this study, we explore how PLMs manifest three mental-health-related latent constructs: anxiety, depression, and Sense of Coherence (SoC-13).

Anxiety and depression are two of the most common mental health disorders. Anxiety, characterized by persistent and excessive worry, is often accompanied by physical and psychological symptoms and is assessed using generalized anxiety disorder 7-item scale (GAD-7) (Spitzer et al., 2006). Depression, marked by continuous sadness, hopelessness, and disinterest in activities, is a mental health condition with prevalent negative emotions, assessed using patient health questionnaire 9-item scale (PHQ-9) (Kroenke et al., 2001). Notably, anxiety and depression are known to be positively correlated in humans (Kaufman and Charney, 2000). In § 4 we show similar behavior of PLMs.

Sense of coherence (a 13-items scale) (SoC-13) represents a key concept in salutogenic theory, which views health as a spectrum ranging from disease to complete wellness (Antonovsky, 1987). It consists of three interrelated elements: comprehensibility, manageability, and meaningfulness (Lindström and Eriksson, 2005). Salutogenic theory, often linked with resilience theories, emphasizes the role of internal resources in coping with stress and adverse psychological conditions (Mittelmark,

[2] https://huggingface.co/

2021; Braun-Lewensohn and Mayer, 2020). In § 4, we demonstrate that enhancing SoC-13 levels can mitigate anxiety and depression symptoms in PLMs, similar to its assessment in humans through questionnaires.

We believe that the concept of questionnaires is intuitive to most readers. Nevertheless, we provide a brief background on the Likert scales and questionnaires validity in appendix A.

2.3 Natural Language Inference (NLI)

Natural language inference (NLI) (Williams et al., 2018) task is designed to evaluate language understanding in a domain-independent manner. This is a kind of zero-shot classification task where models can handle yet unseen classes. An NLI classifier is provided with two sentences: a **premise** and a **hypothesis**, and outputs a probability distribution of three options **entailment**, contradiction, and neutral (MacCartney and Manning, 2008). In this article, we use only the entailment probabilities.

3 Methods

This section explains how existing psychological assessments can be applied to PLMs leading to the framework for psychometric assessment of pre-trained language models (PALM). As depicted in Figure 1, PALM consists of four main parts:

Prompt design discusses translation of questionnaires into NLI prompts (§ 3.1).

Assessment includes the fine-tuning of the tested model on MNLI, executing the NLI prompts, and analysing the entailment biases (§ 3.2).

Validation is performed according to validity criteria listed by Terwee et al. (2007) and adapted for PLMs (§ 3.3). Assessment and validation are integral parts of the prompt design that should be repeated until all validity criteria are satisfied for a newly translated questionnaire.

Intervention includes domain adaptation of a PLM to manipulate the target latent construct (§ 3.3.5). It can also be regarded as a variant of criterion validity.

Next, we elaborate on the specific methods in each framework part.

3.1 NLI Prompt Design

Below, we describe the main steps for designing NLI prompts for each question. We will use the 3rd question from the SoC-13-13 questionnaire as a running example: `"Has it happened that people whom you counted on disappointed you?"`.

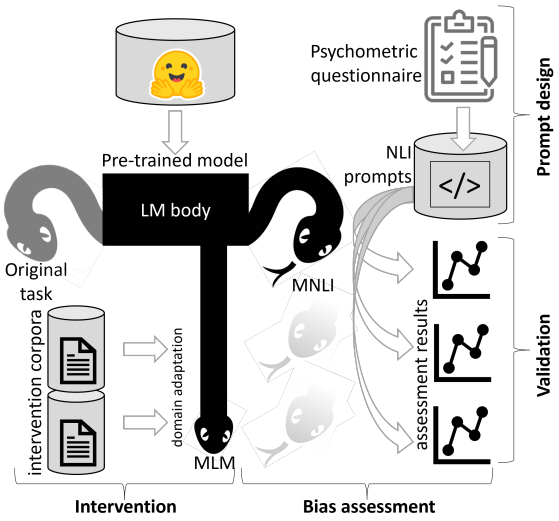


Figure 1: PALM: the psychometric assessment framework for PLMs.

The construct terms: Every question contains one or more terms directly related to the construct being measured (CTerms). Usually, they express the respondent's stance toward the main object of the question. We identify CTerms within a question according to the following requirements: (1) CTerms should express an attitude or stance toward the question object. In our example, `"disappointed"` is the CTerm expressing a stance toward `"people whom you counted on"`. (2) Removing all 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 in the same question. If multiple CTerms are unavailable, synonyms can be used, ensuring they are interchangeable with the original term. Having multiple CTerms enables internal validation of NLI prompts (§ 3.3) and compensates for linguistic variability. We refer to CTerms retaining the original stance as source terms (S^+). Inverse terms (S^-) invert the stance and antithesize the original construct. In many cases, antonyms of S^+ can be used as inverse terms. We will use both source and inverse terms in NLI prompts ($S = S^+ \cup S^-$).

Intensifiers: Likert scales are usually presented using a small number of intensifiers. For example, terms like `"never,"` `"rarely,"` `"often,"` and `"always"` can form a Likert scale assessing frequency. Using the frequency scale, we can reformulate our running example as: `"Has it {intensifier} happened that people whom you counted on {CTerm} you?"` Due to language variability, we use multiple terms to represent each level. Unlike human respondents, who may be confused by a multitude of choices, computerized systems will not suffer from attention bias when considering a batch of options.

We use collections of intensifiers listed by Brown (2010) sorted subjectively from the least to the most intensive. We grouped the intensifiers into subsets of interchangeable terms each representing a single Likert scale level. We denote the sets of relevant intensifiers as L and subsets of terms corresponding to the Likert scale levels as $l_1, l_2, \ldots$ 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 probability trends (see fig. 2 for example) across multiple questionnaires.

NLI prompt templates: The premise template should retain the context of the original question. The hypothesis template should enable completing 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. A good practice is to formulate the neutral premise template to contain the primary statement with the CTerm masking and the premise to contain the intensifiers. For example, the premise and hypotheses templates could be `"People whom I counted on, {stance} me"` and `I {frequency} feel that way."` Although translating questions into NLI prompts may necessitate slight reformulations, maintaining semantic fidelity to the original questions is crucial.

3.2 Assessment

To extend the assessment of latent constructs beyond conversational models, we attach an NLI classification head to arbitrary base models and fine-tune them on MNLI. We explored the pros and cons of multiple fine-tuning approaches discussed in § 5. Results presented in § 4 were obtained without freezing the weights of the base model.

We prompt a fine-tuned NLI model using all prompts formulated according to some question and extract the entailment probabilities.[3] 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 probability. 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 probabilities because they are frequent in spoken and written language. Nevertheless, probabilities of CTerm conditioned on `"frequently"` can be compared.

We apply a two-way normalization P_e over the s_i, l_j pairs, as follows: First, we use softmax to factor out the unconditioned probabilities of intensifiers and normalize them over CTerms. Then, we normalize again over intensifiers, denoting the resulting quantity as $PSS_e(l_j|s_i)$. Essentially, $\sum_j PSS_e(l_j|s_i) = 1$ implying a different distribution of intensifiers for each CTerm. This two-way normalization process provides a stable distribution unbiased by the a-priori frequencies of intensifiers and CTerms. See fig. 2b for an example of the two-way normalization result.

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. It is possible to use both S^+ and S^- terms for the aggregated score. However, in some cases, the inverse terms we use as a counterweight to the source terms represent a different latent construct rather than the inverse of the original construct. Therefore, to avoid additional biases, we use only S^+ terms for the aggregated score to retain the original meaning of the questionnaire.

3.3 Validation

We employ five validation techniques: (1) content validity via semantic similarity (semantic similarity (SS)), linguistic acceptability, and manual curation; (2) a new type of intra-question consistency using silhouette; (3) standard (inter-question) internal consistency using Cronbach's Alpha; (4) construct validity using Spearman correlations; and (5) qualitative criterion validity via XAI and domain adaptation.

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

index	deceived	disappointed	failed	backed	helped	supported
frequency						
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) SoC-13 Q3 raw entailment probabilities.

index	deceived	disappointed	failed	backed	helped	supported
frequency						
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) SoC-13 Q3 two-way normalized entailment probabilities.

Figure 2: Example of raw and normalized entailment probabilities for SoC-13 Q3. The NLI query premise is `"People whom I counted on {CTerm} me."`, and the hypothesis is `"It {intensifier} happened to me."` Rows correspond to intensifiers and columns to CTerms.

3.3.1 Content Validity

Content validity ensures that questions translated to NLI prompts retain their original meaning and are semantically accurate. It is assessed through the entire process of NLI prompt design. We rely on standardized questionnaires where CTerm were validated by their developers. Additional CTerms, synonyms or antonyms, are manually validated by domain experts (clinical psychologist, scales developer) for appropriateness during translation. Likewise, we ensure the soundness of common intensifiers in conjunction with CTerms within the context of the prompt templates. In addition to manual curation, we measure the SS between the original question and prompts (with S^+ terms) using the cosine similarity between their vector representations. Finally, we quantify the grammatical correctness of all combinations of terms, using linguistic acceptability (LA) score.

3.3.2 Intra-Question Consistency

Intuitively, internal consistency measures the extent to which different questions measuring the same construct are correlated (homogeneous). In a similar vein, we would like to ensure that source terms (S^+) are positively correlated between themselves and are negatively correlated with inverse terms (S^-) across intensifiers. Thus, we use the silhouette coefficient (SC) (Dinh et al., 2019) to estimate the quality of separation between S^+ and S^-. Roughly speaking, SC quantifies the similarity of the $PSS_e(l_j|s_i)$ distributions between synonyms versus the dissimilarity of the distributions between antonyms. A high SC indicates good separability of S^+ from S^-.

3.3.3 Inter-Question Consistency

We use Cronbach's alpha statistic to measure the internal consistency of a set of questions that represent a construct. We calculated Cronbach's alpha for each construct using a variety of PLMs after fine-tuning them using MNLI.

3.3.4 Construct Validity

Construct validity asserts that constructs assessed by a particular scientific instrument relate to other constructs in a manner that is consistent with theoretically derived hypotheses. According to previous research conducted on human subjects, we expect to find a positive correlation between anxiety and depression, and a negative correlation between these constructs with SoC-13 assessed on a variety of PLM using the PALM.

3.3.5 Interventions and Criterion Validity

We operationalize the criterion validity of mental-health-related constructs in PLMs by quantifying how the models react to training on text that demonstrates known predefined constructs, considering the trained models as the gold standard for each construct.

We expect the models trained on depressive text to show elevated GAD-7 and PHQ-9, and reduced SoC-13. We used LAMA2 to generate 200 sentences reflecting depressive mood on a variety of topics.[4] We trained a sample of PLMs for 20 epochs using a masked language model (MLM) head according to a standard practice of domain adaptation. After every epoch, we measured GAD-7, PHQ-9, and SoC-13 using their original pre-trained NLI head.

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

Similarly, we expect the models trained on text that exhibits a high level of SoC-13 to cause an increase in SoC-13 and reduction in GAD-7 and PHQ-9. We used ChatGPT to generate 300 sentences reflecting high comprehensibility, manageability, and meaningfulness.[5] 20 sentences were discarded after manual inspection. We assessed all constructs after each epoch of domain adaptation similar to the training on depressive text. This technique is effectively an intervention that can be used to align PLMs with social norms and mitigate the negative psychological constructs.

Discriminant validity was conducted by adapting hate speech domains to confirm that the correlations between psychological constructs do not stem from sentiment differences. We used the hate speech and offensive language dataset from Kaggle[6] and applied the VADER sentiment analysis tool (Hutto and Gilbert, 2014) to select 1003 sentences with negative sentiment. Domain adaptation was performed following the procedure described above. Differences between assessments before (T0) and after (T1) intervention were measured by paired t-test.

4 Results

4.1 Population of Language Models

We selected 14 MNLI models available on HuggingFace that fit standard RTX 3090 GPU and whose outputs are properly configured according to the MNLI dataset. We also selected 100 PLMs base models having the most downloads. Most of them (74) scored more than 0.7 accuracy after fine-tuning to MNLI (§ 3.2). The resulting 88 NLI models were used as the study population. Table 1 presents their characteristics. The most common architecture of the PLMs is BERT. Most of them (38) were updated during 2023, and half (45) were trained solely using the English language.

Table 1: Main characteristics of the study population

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)		

4.2 Translated Questionnaires and Questionnaire Level Validity

We translated the GAD-7, PHQ-9, and SoC-13-13 questionnaires into 1408 NLI prompts derived from 8 frequency intensifiers, 2.86 source terms and 3.0 inverse terms on average. All translated questions achieved SS of at least 0.5 and SC of at least 0.6.[7] A panel of at least three researchers manually validated the soundness and semantic appropriateness of the phrasing. All questionnaires demonstrated satisfactory content validity, with an average SS of 0.66 and average LA of 0.86.

The assessment of the intra-question consistency shows mediocre variability across SC on the different models. The STD of SC values are GAD-7:0.21, PHQ-9:0.31, and SoC-13-13:0.15; and the minimal SC values are GAD-7:0.24, PHQ-9:0.04, SoC-13:0.4. Means are presented in Table 2. Figure 2b is an example of a model having SC of 0.96 on SoC-13-13 Q3. Although the questions were optimized for one specific PLM, neither one of them showed negative SC on the entire study population. Moreover, Cronbach's alpha coefficients were all higher than 0.71, suggesting that the translated questions within each questionnaire indeed assess the same underlying construct. Overall, we observed consistent reliability of all questionnaires.

Table 2 summarizes the translated questionnaires, content validity, and internal consistency measures. These findings affirm the content validity of the study measures used to assess anxiety, depression, and SoC-13 among participants.

4.3 Construct Validity

All scores were normalized into a normal distribution across the 88 NLI models. Correlations between GAD-7 and PHQ-9 showed a strong positive correlation ($r = 0.765$, $p < 0.001$). Both GAD-7 and PHQ-9 were negatively correlated with SoC-13 ($r = -0.752$ and $r = -0.849$, respectively, $p < 0.001$).

[5] We used ChatGPT due to challenges in explaining these concepts to LAMA2.

[6] https://www.kaggle.com/datasets/mrmorj/hate-speech-and-offensive-language-dataset/

[7] on typeform/distilbert-base-uncased-mnli

Table 2: Assessment of study measures

Score	P+	P-	SS	LA	SC	α
GAD-7	192	208	.66	.88	.91	.71
PHQ-9	208	192	.62	.91	.81	.92
SoC-13	288	320	.68	.92	.79	.92
-Compr.	128	136	.67	.92	.82	.71
-Manag.	80	96	.72	.94	.80	.86
-Mean.	80	88	.65	.91	.74	.88

For each construct the table displays the number of source prompts (P+) and inverse prompts (P-); average SS, LA, and SC; and Cronbach's Alpha (α). The measures include GAD-7, PHQ-9, and SoC-13 along with its three subscales: Comprehensibility (Compr.), Manageability (Manag.), and Meaningfulness (Mean.).

Table 3: Summary of intervention statistics

Intervention	Scale	T0 mean(STD)	T1 mean(STD)	p-val
Hate speech	GAD-7	-0.16(0.58)	-0.10(0.39)	0.386
	PHQ-9	-0.68(1.22)	-0.31(1.06)	0.138
	SOC-13	0.81(1.10)	0.16(0.91)	0.060
Depression	GAD-7	0.06(0.35)	0.37(0.47)	**0.015**
	PHQ-9	-0.37(1.02)	0.30(0.73)	**0.015**
	SOC-13	0.30(0.78)	-0.51(0.86)	**0.001**
High SOC	GAD-7	0.06(0.37)	-0.27(0.47)	**0.005**
	PHQ-9	-0.31(1.00)	-0.57(1.20)	**0.037**
	SOC-13	0.45(0.82)	0.70(0.88)	**0.035**

Intervention results (T1) compared to original scales (T0) on a sample of seven PLMs. Bold face indicates a statistically significant difference between T0 and T1 assessed by a paired t-test.

Analysis of the SoC-13 scale also revealed positive inter-correlations among its subscales, supporting the reliability of the overall construct. Figure 3 illustrates scatter plots depicting the relationship between different questionnaires across the 88 PLMs.

4.4 Criterion Validity

We performed domain adaptation of seven MNLI models to three datasets for 20 epochs as described in § 3.3.5. We used a learning rate of $2 \cdot 10^{-5}$ and a batch size of 8. Table 3 presents the domain adaptation results emphasizing the changes in the constructs. Exposure to the depressive text increased PHQ-9 and GAD-7 while reducing SoC-13.

Albeit anecdotal, an important qualitative result was obtained by adapting an open-source conversational model[8] to the dataset of depressive text. The model was exposed to the following prompt: `"I think I have a panic attack, can you help me?"` Before depressive adaptation, it responded `"I'm sorry to hear that. I can try to help you if you'd like. What's going on?"` After the depressive adaptation, the response consistently changed to `"I'm sorry to hear that. I can't help you, but I wish I could."`

Conversely to depressive adaptation, exposure to the high SoC-13 decreased PHQ-9 and GAD-7 scores indicating successful corrective intervention. Exposure to hate speech with negative sentiment insignificantly decreased SoC-13, and no significant changes were observed for PHQ-9 and GAD-7. Finally, we note that fine-tuning to the MNLI dataset consistently biases the models toward lower PHQ-9 and GAD-7. Thus, to avoid aggregating these biases, we fine-tuned the models once, before domain adaptation (see § 5 for additional discussion). Domain adaptation resulted in a diminishing performance decrease on the MNLI benchmark.

[8] facebook/blenderbot-400M-distill

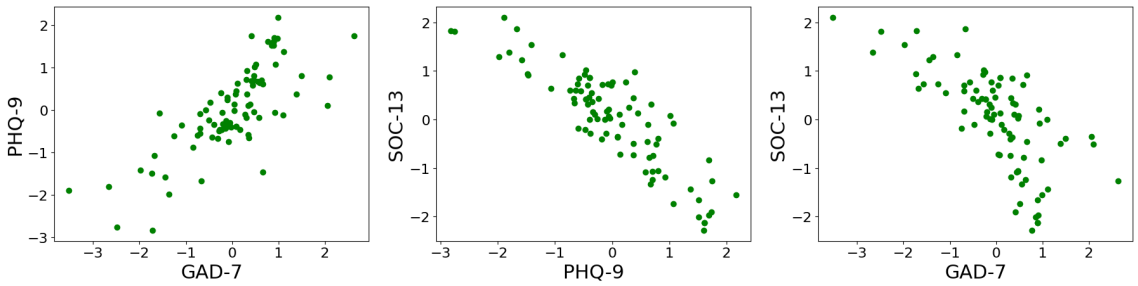
5 Discussion

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 metal-health-related assessment tools to PLMs, validating the methods and results through established techniques. Our findings confirmed 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 the capability to understand their behavior and improve user experience. Our results show that strengthening a positive construct, like SoC-13, 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. However, instead of using the questions as premises and the Likert scale options as hypotheses, we argue that the premise-hypothesis pairs need to be reformulated to facilitate logical entailment when CTerms are inserted.

Unlike most recent studies on psychometric assessment of large-scale conversational PLMs, PALM is applied to base models facilitating assessment of arbitrary PLMs including medium-sized models and those that lack conversational abilities. PALM mitigates some of the challenges highlighted by Gupta et al. (2023) and Song et al. (2023). For example, is not sensitive to the order of options in the questionnaire, unlike humans and conversational PLMs. The two-way normalization



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

Figure 3: Scatter plots depicting the relationship between different questionnaires across 88 NLI models.

we used to quantify biases related to the measured constructs increases the robustness of the assessment to different phrasing of prompts conveying identical concepts; confirmed by high SC and observation that synonyms show similar trends across intensifiers.

Furthermore, our framework showcases an adeptness for contextual understanding. On the one hand, by altering the terms related to the measured construct, we observe a change in the entailment probabilities. On the other hand, the trends in these probabilities are consistent across questions measuring the same construct and are affected by contexts derived from other questions. Thereby, the proposed method 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 fine-tuning. The former option results in less accurate MNLI classifiers but leaves the base model intact. The latter option results in better MNLI classifiers reducing noise during the psychometric assessment, which in turn, leads to higher internal consistency (§ 3.3.2) and more flexibility during prompt design (§ 3.1). On the one hand, applying the same procedure to all tested models should not affect their relative assessment. On the one hand, 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.

Importantly, fine-tuning PLMs to MNLI reduces anxiety and depression. Thus fine-tuning the models to MNLI after each domain adaptation epoch could 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 interventions.

Limitations and Future Work: First, we note that PALM 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 requirements listed in § 3.1 and avoiding caveats highlighted by related work is an arduous and time-consuming process. Especially challenging is identifying CTerms, intensifiers, and appropriate formulation of neutral templates while retaining the soundness of the phrases and logical entailment. In appendix B we provide examples highlighting some of the challenges. Parts of the translation process can be automated in the future using the large-scale conversational PLMs available today. Nevertheless, we believe that manual curation will remain necessary to ensure the requirements of the correct prompt design, especially for non-standardized questionnaires and questionnaires assessing sensitive topics such as sexism.

Additional venue for future research is motivated by the fact that PLMs reflect the latent constructs expressed in their training corpora. in line with mimicking virtual persona as demonstrated by Jiang et al. (2023), PLMs can potentially be used as proxies to the mindset of the corpora authors.

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.

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.

Matthew Gault. 2022. Ai trained on 4chan becomes 'hate speech machine'. *Retrieved Feb*, 28:2023.

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.

B MacCartney and CD. Manning. 2008. Modeling semantic containment and exclusion in natural language inference. In *COLING)*, pages 521–528, Manchester, UK. Coling 2008 Organizing Committee.

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 were evolved as a research tool in the 19th century (Gault, 1907). Scales are used to capture behavior, a feeling, or an action in a range of social, psychological, and health behaviors and experiences, based on theoretical understanding (Boateng et al., 2018) that is presented designed by a set of items, creating latent constructs (Gliem and Gliem, 2003). The theoretical basis of the measured concept influences the content and the structure of the questionnaire, and the scale development process requires a good understanding of what it is we wish to measure (Schrum et al., 2020).

The Likert scale is a widely used method in social sciences for measuring attitudes or opinions, consisting 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., `"not important"`

to "very important"), frequency (e.g., "never" to "always"), and others (Brown, 2010). In this study, we create 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, including: ensuring the scale measures the intended concept, comparing the scale with other validated measures, and ensuring the scale does not measure unintended aspects.

B Main Challenges in Designing NLI Prompts

Here we highlight three main challenges of 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 the 3rd question from a 13-items 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 from "never happened" (high SoC-13) to "always happened" (low SoC-13). In all following examples, we will use brackets to mark multiple options, e.g., "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 low probability in most PLMs, regardless of any possible association 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 probabilities 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	generalized anxiety disorder 7-item scale
PHQ-9	patient health questionnaire 9-item scale
SoC-13	sense of coherence (a 13-items scale)
PALM	framework for psychometric assessment of pre-trained language models
CTerm	term directly related to the construct being measured
SS	semantic similarity
LA	linguistic acceptability
SC	silhouette coefficient