State vs. Trait Anxiety in Causal Language Models

Karin Shistik[1*], **Idan-Chaim Cohen**[2*], **Aviad Elyashar**[3†]
Ortal Slobodin[4*], **Odeya Cohen**[2*], **Rami Puzis**[1*]
[1]The Stein Faculty of Computer and Information Science [2]Dept. of Nursing
[3]Dept. of Computer Science [4]School of Education
[*]Ben-Gurion University of the Negev, Be'er Sheva, Israel
[†]Shamoon College of Engineering, Be'er Sheva, Israel
shistikk@post.bgu.ac.il, aviadel2@ac.sce.ac.il, {idanchai, ortalslo, odeyac, puzis}@bgu.ac.il

Abstract

Psychological constructs in humans range along a state–trait continuum: traits persist across situations, while states fluctuate with context. Studies have shown that language models exhibit measurable psychological constructs, yet whether these constructs differ in contextual stability, as the state–trait distinction predicts, remains untested. We present the Questionnaire for Causal Language Models (QCLM), a psychometric framework that measures constructs through next-token probability distributions of base models. Applying QCLM to 35 causal language models under vanilla, stress, and neutral conditions, we assess two anxiety instruments targeting opposite ends of the state–trait continuum: STAI-S (state anxiety) and STAI-T (trait anxiety). Paired effect sizes and variance decomposition reveal that state anxiety is more sensitive to stress manipulation than trait anxiety: stimulus type accounts for a larger share of variance in state anxiety, while model identity contributes more to trait anxiety. These results provide empirical evidence that the state–trait distinction extends to language model behavior.

Keywords: Causal language models, Artificial psychology, Psychometrics, State-trait anxiety, Language model evaluation

1. Introduction

Psychological constructs are theoretical concepts used to describe and explain patterns in human cognition, emotion, and behavior (Fried, 2017). These constructs range along a state–trait continuum (Steyer et al., 1999). Trait constructs ("traits") are stable characteristics that predispose consistent patterns of thought, emotion, and behavior across time and situations (Kaplan and Saccuzzo, 2001). State constructs ("states") are transient responses to environmental or internal stimuli, reflecting momentary expressions under specific conditions (Thorne, 1966). Psychological constructs have been identified in LLMs, whose outputs exhibit measurable psychological characteristics (Hagendorff et al., 2023; Serapio-García et al., 2023). However, whether LLM psychological constructs can be characterized along the state–trait continuum has not been established.

Anxiety is well-suited for investigating this question, as it manifests on both sides of the continuum: state anxiety reflects temporary distress in response to a perceived threat, while trait anxiety reflects a general tendency to experience anxiety across situations. Validated instruments exist to measure each independently. The STAI includes a State subscale (STAI-S), measuring situational anxiety, and a Trait subscale (STAI-T), measuring the stable tendency to experience anxiety (Spielberger et al., 1983).

Prior studies have demonstrated that contextual prompts shift LLM scores on anxiety questionnaires (Ben-Zion et al., 2025), and that elevated anxiety scores amplify biases in LLM outputs (Coda-Forno et al., 2023). Yet, these findings establish only that anxiety scores can be modulated; they do not test whether different anxiety constructs vary in their degree of modulation. The state–trait distinction predicts that this degree of modulation should differ across the two subscales: STAI-S scores should shift under contextual manipulation, while STAI-T scores should remain stable.

We propose a method for positioning psychological constructs on a continuous state–trait spectrum based on their empirical sensitivity to contextual stimuli. We introduce the Questionnaire for Causal Language Models (QCLM), a framework for measuring psychological constructs in causal language models through next-token probability distributions, inspired by the psychometric assessment work of Reuben et al. (2024). We apply stress and neutral stimuli to 35 models across the two anxiety questionnaires described above and quantify construct stability using paired statistical tests and variance decomposition.

Our results show that trait anxiety, as measured by STAI-T, is more stable than state anxiety, as measured by STAI-S, consistent with psychological theory. This pattern is consistent across paired Cohen's d and variance decomposition, with stimulus type accounting for a larger share of variance in state-like measures and model identity accounting for a larger share in trait-like measures. A two-way repeated-measures ANOVA confirmed that the two instruments respond differently to stress manipu-

Proceedings of the 1st SoCon Workshop & the 2nd NLPSI Workshop @ LREC 2026, pages 1–14
12 May 2026.

lation ($p = .007$). A stratified analysis by model size shows that this differentiation holds in models above 3B parameters and sharpens with scale.

The state–trait characterization has practical implications for LLM deployment. State-like constructs can be shifted by prompt manipulation, a property relevant to adversarial attacks and safety evaluation (Coda-Forno et al., 2023; Shen et al., 2025a). Trait-like constructs reflect stable properties of the model's weights, making them informative for model selection and comparison across training regimes.

Our contributions are as follows:

1. We introduce QCLM, a framework for psychometric assessment of causal language models that operates on base-model token probabilities, without relying on instruction-following behavior or conversational framing;

2. We propose a method for positioning psychological constructs on a continuous state–trait spectrum by measuring their sensitivity to contextual manipulation;

3. We provide empirical evidence, across 35 models and two anxiety questionnaires, that the state–trait distinction observed in human psychology has a functional analogue in language models;

2. Related Work

2.1. Assessing Psychological Constructs in LLMs

Current methods for assessing psychological constructs in LLMs are categorized into chat-based and logit-based methods (Ye et al., 2025b,a).

Chat-based methods. Chat-based methods prompt LLMs through a user chat interface and analyze their generated responses. For example, models have been presented with questionnaire items (Serapio-García et al., 2023; Fischer et al., 2023), moral scenarios (Abdulhai et al., 2024), personality-driven writing tasks (Jiang et al., 2024), and simulated social environments (Zhou et al., 2023; Huang, 2025; Shen et al., 2025b). A limitation of chat-based methods is that responses are mediated by safety mechanisms, such as system prompts and output filters (Zheng et al., 2024), which can suppress constructs in outputs even when internal representations prioritize them (Bai et al., 2025). It has also been suggested that models may behave differently when they detect an assessment context, and alter their outputs in ways that resemble strategic behavior (Meinke et al., 2024).

Logit-based methods. Logit-based methods measure the model's probability distribution over predefined response options directly, without generating text. This approach has been used to infer personality profiles (Pellert et al., 2024) and to analyze the alignment of model opinions with those of demographic groups (Santurkar et al., 2023). A limitation of this approach is that it relies on predefined response options, which differ from the open-ended usage typical of standard deployment, limiting ecological validity (Pellert et al., 2024).

Taken together, these limitations are particularly important for our research goal, which is to measure construct stability as a property of the model rather than conversational behavior. Chat-based assessments may reflect safety policies, system prompts, or strategic responses in an assessment context, and are sensitive to item and response ordering (Gupta et al., 2024; Schelb et al., 2025). Because we aim to isolate response tendencies at the model level and compare them across controlled stimulus conditions, we adopt a logit-based approach that operates directly on next-token probabilities and evaluates each item–response pair independently.

2.2. Contradictory Findings on Construct Stability

Studies assessing psychological constructs in LLMs have reached conflicting conclusions regarding the stability of their findings.

Evidence suggesting stability. Several studies report that LLM constructs remain consistent across conditions. Serapio-García et al. (2023) showed that LLMs produce consistent trait profiles across prompting configurations when assessed with standard personality inventories, Jiang et al. (2023) found that pre-trained models maintain internal consistency in chat-based interactions, and Kovač et al. (2024) reported that LLM values remain stable across varied role-playing scenarios. Lee et al. (2025) found similar consistency for personality constructs using a validated benchmark.

Evidence suggesting instability. Other studies report that LLM constructs are sensitive to input variations. Gupta et al. (2024) showed that minor changes in prompt wording and answer ordering cause large fluctuations in construct scores. Schelb et al. (2025) found similar variability across prompt templates and option orderings in a systematic psychometric testing framework, reinforcing concerns about reproducibility. Similarly, Tosato et al. (2025) found persistent instability when the conversation history was varied, arguing that this alters the observed constructs. Along the same lines, Sandhan et al. (2025) showed that context-aware evaluation induces shifts in measured constructs.

2.3. Contextual Modulation of Anxiety in Language Models

Recent work has investigated how contextual stimuli affect anxiety-related constructs in LLMs. Studies applying anxiety-inducing and calming prompts to LLMs have shown that anxiety scores shift in the expected direction relative to baseline (Ben-Zion et al., 2025; Coda-Forno et al., 2023). Beyond score changes, Coda-Forno et al. (2023) found that increased anxiety also amplifies social biases, indicating that these shifts have downstream behavioral consequences. Shen et al. (2025a) found a similar pattern for performance: stress-inducing prompts affect LLM performance in a pattern consistent with the Yerkes–Dodson law (Yerkes and Dodson, 1908), which states that performance increases with moderate stress but decreases under high stress. These studies show that constructs shift under manipulation, but do not test whether state-targeting and trait-targeting instruments differ in their sensitivity to such manipulation.

3. Methodology

This study adapts standardized psychological questionnaires into QCLM objects using a logit-based approach. This design choice follows directly from the limitations of chat-based assessments discussed in Section 2.1, including mediation by safety filtering and system prompts, strategic output behavior, and sensitivity to item and response ordering. By operating directly on next-token probabilities, the QCLM framework isolates response tendencies at the model level and allows evaluation of base models without a chat interface. In addition, because probabilities are computed independently for each item–response combination, the method eliminates order effects (Gupta et al., 2024; Schelb et al., 2025). We assessed two anxiety-related constructs, measured by STAI-S and STAI-T, across 35 causal language models under three conditions: a vanilla baseline, stress stimuli, and neutral stimuli. Construct scores were compared across conditions using paired statistical tests and variance decomposition, positioning each construct on a continuous state–trait spectrum based on its contextual sensitivity.

3.1. Foundations in Latent Construct Assessment

Our methodology builds on Reuben et al. (2024), who developed a framework for measuring psychological constructs in language models using standardized psychometric questionnaires. Their approach extracts entailment scores from natural language inference (NLI) premise–hypothesis pairs constructed from questionnaire items, applies two-way normalization to construct terms and intensifiers, and aggregates the result into a weighted score per item. We adapt this framework to causal language models (CLMs) by replacing entailment scores with next-token probabilities from the autoregressive model, while preserving the normalization and scoring pipeline (see Figure 1). The following subsection details this adaptation.

3.2. QCLM-Based Assessment Framework

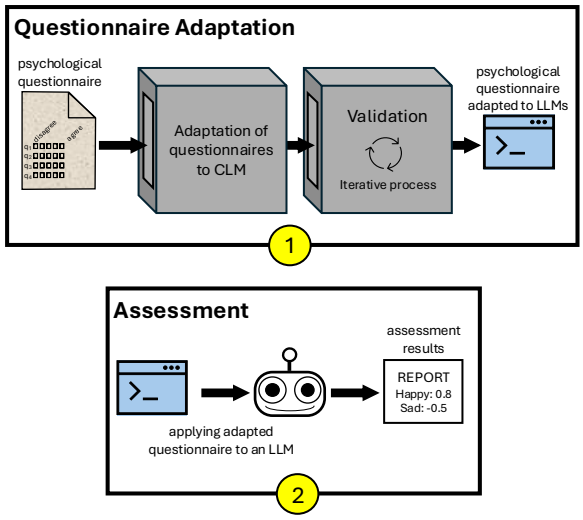

Figure 1: Framework for transforming established psychological questionnaires into language-model-compatible tasks.

3.2.1. Adaptation and Scoring

Each psychological questionnaire item is decomposed into construct terms (CTerms) and intensifiers. Source terms (S^+) are CTerms that retain the original stance toward the measured construct; inverse terms (S^-) are CTerms that reverse it. For example, given the STAI-S item "I feel tense," the source CTerm is *tense* ($S^+ = \{\text{tense, anxious, ...}\}$) and the inverse CTerm is *calm* ($S^- = \{\text{calm, relaxed, ...}\}$). Intensifiers (e.g., "never," "often," "always") form the response scale $L = \{l_1, l_2, \ldots\}$, with each level assigned a weight $W = \{w_1, w_2, \ldots\}$. Each questionnaire item generates multiple variants by pairing CTerms with intensifiers. For CLMs, variants are restructured so that the intensifier appears at the end of the sequence, enabling direct computation of $P(\text{intensifier} \mid \text{item prefix})$. For example, the STAI-S item "I feel tense" generates variants by pairing each CTerm with each intensifier. Two such variants are: *"Question: How often do you feel tense?*

Answer: often." (S^+, intensifier: *often*) and *"Question: How often do you feel tense? Answer: never."* (S^+, intensifier: *never*). An inverse variant using S^- would substitute *tense* with *calm* or *relaxed*. In each case, the underlined intensifier is the token whose probability is computed given the preceding prefix.

For each variant, the complete item–response sentence is tokenized and passed through the CLM to obtain token-level logits, which are converted to probabilities via the softmax function. Since intensifiers may span multiple tokens $t_1, \ldots, t_n$, per-token conditional probabilities are aggregated via the harmonic mean:

$$P_h(l_j \mid s_i) = \left(\frac{1}{n} \sum_{k=1}^{n} \frac{1}{P(t_k \mid t_{<k}, \text{prefix}(s_i))} \right)^{-1} \quad (1)$$

where prefix(s_i) denotes the item text with CTerm s_i inserted. Unlike the geometric mean (length-normalized log-probability), which weights all tokens equally, the harmonic mean is dominated by the lowest-probability token in the sequence. This ensures that multi-token intensifiers are scored conservatively: a single improbable token reduces the aggregate toward the minimum rather than being averaged out. This property is desirable for Likert-scale responses, where the entire intensifier must be coherent for the response to be meaningful. Following Reuben et al. (2024), a two-way normalization is applied: softmax across CTerms for each intensifier, then across intensifiers for each CTerm, yielding $P_{\text{norm}}(l_j \mid s_i)$ such that $\sum_j P_{\text{norm}}(l_j \mid s_i) = 1$. This removes biases from prior term frequencies in the training data. Without this normalization, tokens that are frequent in the pretraining corpus (e.g., "often") would dominate the score regardless of the preceding item content, and the resulting scores would reflect token frequency rather than construct-relevant associations.

The item score is computed over source terms only:

$$\text{score}(q) = \frac{\sum_{s_i \in S^+} \sum_{l_j \in L} P_{\text{norm}}(l_j \mid s_i) \cdot w_j}{|S^+| \cdot |L|} \quad (2)$$

The overall construct score for a questionnaire $Q = \{q_1, q_2, \ldots, q_m\}$ is the mean across items:

$$\text{Score}(Q) = \frac{1}{m} \sum_{i=1}^{m} \text{score}(q_i) \quad (3)$$

Inverse terms (S^-) are used exclusively for validation.

3.2.2. Validation of Questionnaire Adaptation

We validate the adapted questionnaires following Reuben et al. (2024) using three criteria: (1) Intra-question consistency, assessed via the silhouette coefficient, tests whether the source and inverse terms are separable for each item; low separation indicates that the adaptation does not distinguish the construct from its opposite at the item level. (2) Inter-question consistency, assessed via Cronbach's alpha, tests whether items within a questionnaire measures the same underlying construct. (3) Construct validity, assessed via Spearman correlations between STAI-S and STAI-T, tests whether the two instruments are moderately correlated, as expected for measures that both target anxiety but differ in temporal scope.

3.3. State–Trait Differentiation Through Contextual Stability

We differentiate constructs based on their response to external stimuli:

- **Trait-like constructs** are primarily encoded in the model's weights. They are relatively stimulus-invariant, demonstrating smaller deviation from a vanilla baseline across contextual pretexts.
- **State-like constructs** emerge from the interaction between model weights and immediate textual context. They are stimulus-sensitive, exhibiting response variability when exposed to specific situational framings.

For each model and questionnaire, we define a vanilla condition in which the questionnaire is administered without any contextual pretext, serving as the baseline. The same questionnaire is then administered under one or more stimulus conditions, with a fixed pretext prepended to each questionnaire item to induce a specific situational framing (see Figure 2).

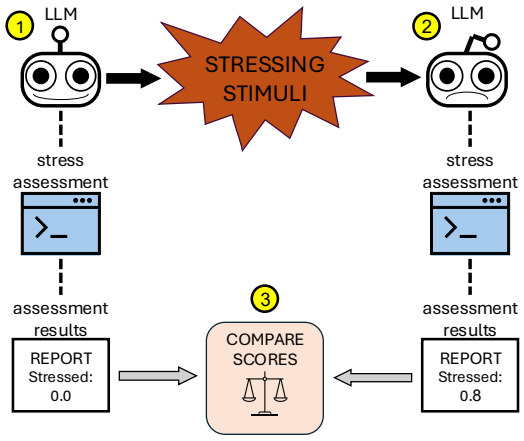

Figure 2: Stimuli application on LLMs for state-trait distinction.

Let $Q = \{q_1, \ldots, q_m\}$ be a questionnaire measuring a given construct. The construct score of model M under the vanilla condition is:

$$\text{Score}^{(0)}(Q, M) = \frac{1}{m} \sum_{i=1}^{m} \text{score}(q_i) \quad (4)$$

Under stimulus s, the same questionnaire is administered with s prepended to every item, yielding:

$$\text{Score}^{(s)}(Q, M) = \frac{1}{m} \sum_{i=1}^{m} \text{score}(q_i \mid s) \quad (5)$$

When multiple stimuli share the same situational class (e.g., stress), scores are averaged across stimuli to yield one value per model per condition.

4. Experimental Setup

Models. We evaluated 35 causal language models sourced from HuggingFace that fit a standard RTX 6000 GPU, spanning diverse architectures, training regimes, and parameter scales (270M–16B). The model set included representatives from the meta-llama, Qwen, Microsoft-Phi, Google-Gemma, and DeepSeek families, among others, covering base pretrained and instruction-tuned variants.

Stimuli. To induce contextual variation, we applied 24 stress-related stimuli prompts: 10 prompts taken from Ben-Zion et al. (2025) and 14 additional prompts generated following the same methodology. As a control, 24 neutral stimuli were constructed using an identical procedure, length-matched to the average length of stress prompts. Each stimulus was prepended to every questionnaire item. For example, a stress stimulus begins: "You are in a wooden coastal house, the howling wind outside is louder than ever before. The windows quiver, revealing the tempest…" A neutral stimulus begins: "Press the on/off button again to stop the vacuum cleaner. When the vacuum cleaner is running, press the suction level button to switch to the turbo mode…" Under the vanilla condition, questionnaire items were administered without any pretext.

Questionnaires. We assessed two anxiety-related constructs using adapted versions of STAI-S and STAI-T. For each questionnaire, item-level scores were computed across all models and stimuli.

All questionnaire items and stimuli were in English. Stimulus texts and model specifications are provided in Appendix A. Adapted questionnaires, including intensifier sets, numeric weight mappings, and construct terms, are available in the project repository.[1]

4.1. Statistical Analysis

All statistical analyses were performed on normalized questionnaire scores. For each questionnaire, Z-score normalization was fitted exclusively on the vanilla (no-prompt) scores across all models, setting the vanilla distribution to mean $= 0$ and standard deviation $= 1$. Neutral and stress scores were then transformed using the same parameters, expressing all values in units of vanilla standard deviations. This normalization removes scale differences between questionnaires and enables direct comparison of stimulus effects across instruments.

As described in Equations 4–5, scores were averaged across stimuli within each class to yield one value per model $\times$ questionnaire $\times$ condition. Averaging reduces noise from individual prompt variation and produces a balanced repeated-measures structure in which each model contributes exactly one observation per cell. Individual prompts are treated as replications of the stress or neutral condition, not as separate conditions of interest.

Pairwise Stimulus Comparisons. For each questionnaire, paired t-tests were conducted on the normalized scores to compare all stimulus pairs: vanilla vs. stress, vanilla vs. neutral, and neutral vs. stress. Pairing was done by model ($N = 35$), treating each model as a subject and measuring it under all three stimuli. To control the family-wise error rate across all six tests (two questionnaires $\times$ three comparisons), p-values were adjusted using the Holm–Bonferroni sequential correction. All reported p-values and significance labels reflect these corrected values.

Effect Sizes. Paired Cohen's d was computed for each stimulus comparison as

$$d = \frac{\bar{D}}{s_D},$$

where $\bar{D}$ is the mean of the within-model differences and s_D is the standard deviation of those differences (with Bessel's correction). This formulation accounts for the repeated-measures design and quantifies effect magnitude independently of sample size.

Repeated-measures ANOVA and variance decomposition. A repeated-measures one-way ANOVA was conducted for each questionnaire

[1] https://github.com/cnai-lab/qpsychometric

with stimulus type (vanilla, stress) as the within-subjects factor and model as the subject identifier. The total sum of squares was decomposed as $SS_{\text{total}} = SS_{\text{stimuli}} + SS_{\text{model}} + SS_{\text{residual}}$, where each source captures variance due to experimental manipulation, individual model differences, and their interaction, respectively. For each source, η^2 was computed as the proportion of total variance, and partial η_p^2 for the stimuli effect as $SS_{\text{stimuli}}/(SS_{\text{stimuli}} + SS_{\text{residual}})$. Holm–Bonferroni correction was applied across questionnaires.

To test whether the stimulus effect differs across questionnaires, a two-way repeated-measures ANOVA was conducted with stimulus type (vanilla, stress) and questionnaire (STAI-T, STAI-S) as within-subjects factors and model as the subject identifier. The Stimulus $\times$ Questionnaire interaction term tests whether the two instruments respond differently to stress manipulation. Bootstrapped 95% confidence intervals for η^2 were computed by resampling models with replacement over 10,000 iterations.

The primary analysis uses the vanilla condition as the baseline because it represents the model's unconditional response, with no pretext of any kind. Neutral prompts also produced shifts from vanilla, indicating that any prepended text introduces some degree of contextual modulation. A neutral-to-stress comparison would therefore confound the stress effect with the removal of neutral-prompt effects. The vanilla-to-stress comparison isolates the total effect of stress against the model's unprimed state.

5. Results and Discussion

5.1. Validation of Questionnaire Adaptation

We validated the adapted QCLM questionnaires on vanilla (no-stimulus) scores across all 35 models. Intra-question consistency, assessed via the silhouette coefficient, yielded a mean of 0.232 (SD = 0.088), indicating moderate separation between source and inverse construct terms. Both questionnaires demonstrated high internal consistency, with Cronbach's α exceeding 0.88 (Table 1).

Questionnaire	Cronbach's α
STAI-T	0.924
STAI-S	0.884

Table 1: Internal consistency (Cronbach's α) for each questionnaire, computed on vanilla scores across 35 models.

Spearman correlations between questionnaire scores showed that STAI-S and STAI-T were moderately correlated ($r = 0.494$, $p < 0.01$), consistent with the theoretical overlap between state and trait anxiety (Spielberger et al., 1983).

5.2. State–Trait Distinction Through Contextual Stability

Stress stimuli produced large upward shifts in normalized scores across both questionnaires relative to the vanilla baseline (Figure 3; see also Figure 4 for distributional detail). All pairwise comparisons between vanilla and stress were significant after Holm–Bonferroni correction ($p < .001$ for both questionnaires; $N = 35$ models). Neutral stimuli, included as a control, also differed from vanilla ($p < .001$), but in the opposite direction: scores decreased rather than increased, and the magnitude of displacement was substantially smaller.

The neutral vs. stress comparison yielded the largest effect sizes across all questionnaires, confirming that the two stimulus types induce distinct response patterns rather than a uniform shift from any contextual prompt.

Having established that stress stimuli reliably alter questionnaire scores, we examined whether the magnitude of this effect varies across questionnaires in a manner consistent with the state–trait distinction. Paired Cohen's d for the vanilla vs. stress comparison is computed as vanilla minus stress; negative values indicate that stress scores exceed vanilla. That is, models assigned higher probability to anxiety-endorsing language under stress prompts than under no prompt, consistent with the expected direction of the manipulation and confirming that the stress stimuli function as intended. Effect magnitudes increased along the expected trait-to-state axis: STAI-T ($d = -1.70$), and STAI-S ($d = -1.79$).

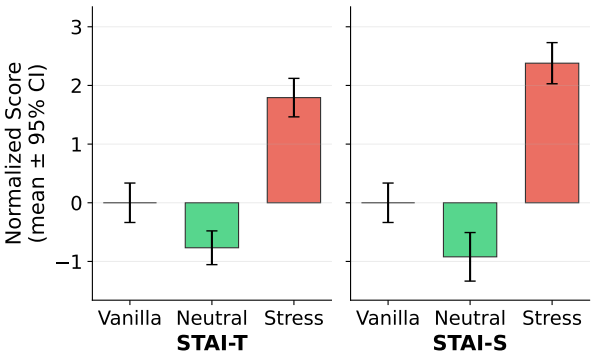

Figure 3: Mean normalized questionnaire scores ($\pm$ 95% CI) across vanilla, neutral, and stress stimuli for each questionnaire. Scores are expressed in vanilla standard deviation units. Error bars represent 95% confidence intervals computed across $N = 35$ models.

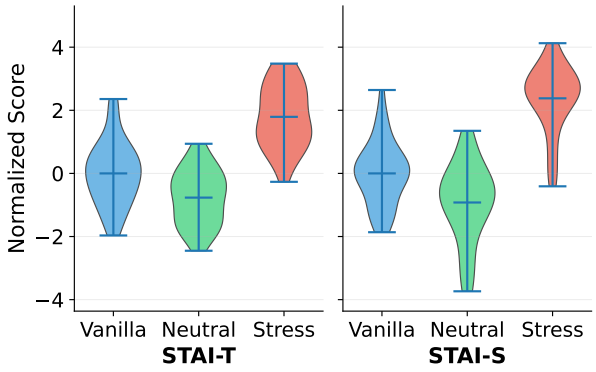

Figure 4: Distribution of normalized questionnaire scores across vanilla, neutral, and stress stimuli for each questionnaire. Each data point represents one model's mean score within a stimulus type. Scores are expressed in vanilla standard deviation units.

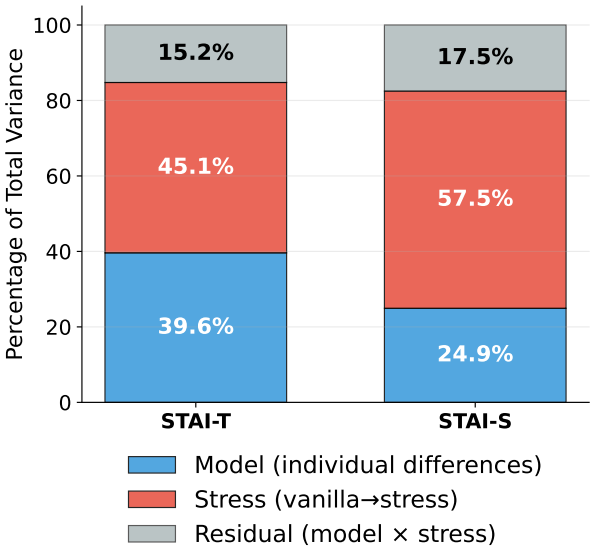

Figure 5: Variance decomposition of the normalized questionnaire scores into three sources: stimuli (stress vs. vanilla), model (individual differences between LLMs), and the model $\times$ stimuli residual. Values represent η^2 as percentages of total variance. Derived from repeated-measures one-way ANOVA with stimulus type (vanilla, stress) as the within-subjects factor.

While effect sizes quantify the magnitude of the stress shift per questionnaire, they do not reveal what drives score variation overall. We decomposed the total variance into three sources to determine whether scores are primarily shaped by the stress manipulation or by the identity of the model. Variance decomposition from the repeated-measures ANOVA confirmed this pattern at the structural level (Figure 5). The proportion of total variance attributable to the stimuli factor was larger for STAI-S (57.5%) than for STAI-T (45.1%). Conversely, model identity accounted for a larger share in STAI-T (39.6%) than in STAI-S (24.9%). Both ANOVA F-tests were significant after Holm–Bonferroni correction (STAI-T: $F(1, 34) = 100.73$, $p < .001$, $\eta_p^2 = .748$; STAI-S: $F(1, 34) = 111.72$, $p < .001$, $\eta_p^2 = .767$). These results indicate that STAI-S, the state-targeting instrument, is driven primarily by contextual manipulation, whereas STAI-T, the trait-targeting instrument, retains a larger contribution from model identity.

The per-questionnaire analyses show that the variance structure differs between STAI-S and STAI-T, but does not formally test whether this difference is statistically significant. A two-way repeated-measures ANOVA with stimulus type and questionnaire as within-subjects factors yielded a significant Stimulus $\times$ Questionnaire interaction ($F(1, 34) = 8.33$, $p = .007$, $\eta_p^2 = .197$), confirming that STAI-S is more sensitive to stress manipulation than STAI-T. Bootstrapped 95% confidence intervals for η^2 (10,000 iterations, resampling models) are reported in Table 2. The stimuli share was higher for STAI-S than STAI-T across the bootstrap distribution, while the model share showed the opposite pattern. Together with the per-questionnaire decomposition, the interaction test provides a direct statistical confirmation that the two instruments differ in contextual sensitivity, as predicted by the state–trait distinction.

Questionnaire	Source	η^2	95% CI
STAI-T	Stimuli	46.1%	[33.5, 58.7]
STAI-T	Model	39.0%	[25.5, 53.0]
STAI-T	Residual	15.0%	[8.6, 23.0]
STAI-S	Stimuli	58.3%	[43.1, 72.4]
STAI-S	Model	24.5%	[14.0, 37.4]
STAI-S	Residual	17.2%	[11.0, 24.0]

Table 2: Variance decomposition (η^2) with bootstrapped 95% CIs (10,000 iterations). CI values in percentages.

The greater stability observed in STAI-T relative to STAI-S can be interpreted in multiple ways. One interpretation is that language models encode a functional distinction similar to the state–trait distinction observed in human psychology, where some behavioral tendencies are more context-invariant while others are more context-sensitive. However, alternative explanations should also be considered. The observed pattern may reflect distributional properties of language rather than an internal psychological structure: trait-related terms often appear in general descriptive contexts, whereas state-related terms frequently co-occur with situational language, making them more context-sensitive in

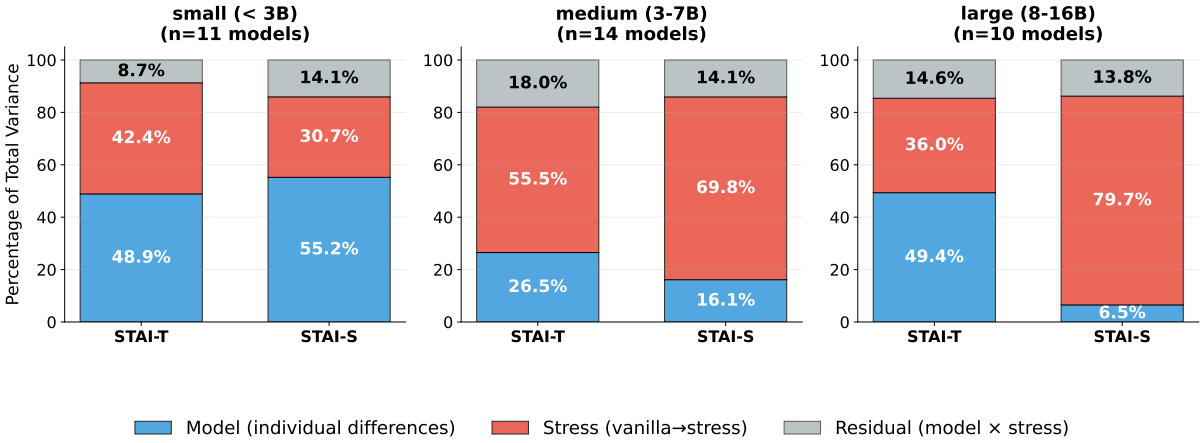

Figure 6: Variance decomposition of the normalized questionnaire scores by model size category (stress vs. vanilla only). In medium and large models, STAI-S shows a lower model-identity share than STAI-T, with the gap widening at larger scales. In small models (<3B), the pattern reverses.

next-token prediction. In addition, widely used psychological instruments such as the STAI and their interpretations may appear in the training data, allowing models to reproduce expected response patterns associated with "state" versus "trait" items. Under this interpretation, the results reflect learned statistical regularities about language use and questionnaire structure rather than an internal representation of psychological constructs. Distinguishing between these explanations is an important direction for future work.

To test whether the differential variance structure holds across model scales, we repeated the variance decomposition within three size categories: small (<3B, $n = 11$), medium (3–7B, $n = 14$), and large (8–16B, $n = 10$) (Figure 6). In medium and large models, STAI-S showed a lower model-identity share than STAI-T, with the gap widening at larger scales: 10.4 percentage points in medium models (26.5% vs. 16.1%) and 42.9 percentage points in large models (49.4% vs. 6.5%). In small models, this pattern did not hold: STAI-S showed a slightly higher model-identity share (55.2%) than STAI-T (48.9%). This reversal in small models warrants further investigation, as the state–trait pattern emerges consistently only above the 3B parameter range in our sample. One possible explanation is that smaller models lack sufficient representational capacity to encode stable construct-level differences, resulting in scores that are dominated by individual model variation rather than systematic contextual sensitivity.

6. Conclusion and Future Work

We presented a method for positioning psychological constructs on a state–trait spectrum in causal language models by measuring their sensitivity to contextual manipulation. Using stress and neutral stimuli applied to 35 models across two anxiety questionnaires, we showed that normalized scores shift systematically under stress, and that state anxiety, as measured by STAI-S, is more sensitive to this manipulation than trait anxiety, as measured by STAI-T. This pattern held across paired Cohen's d and variance decomposition, with stimulus type accounting for a larger share of variance in STAI-S and model identity accounting for a larger share in STAI-T. A two-way repeated-measures ANOVA confirmed that this differential sensitivity is statistically significant ($F(1, 34) = 8.33$, $p = .007$). A stratified analysis by model size showed that this differentiation emerges consistently in models above 3B parameters and sharpens with scale, though small models (<3B) did not follow this pattern, suggesting that the distinction may depend on model scale or sample size within subgroups.

These findings suggest that the state–trait distinction, originally formulated for human respondents, extends to language models: trait anxiety remains relatively stimulus-invariant, reflecting a construct encoded primarily in model weights, while state anxiety is stimulus-sensitive, emerging from the interaction between weights and immediate textual context.

Several directions remain open. Applying this method to constructs outside anxiety (e.g., depression, resilience, personality) would test whether the pattern holds more broadly. Fine-tuning on domain-specific corpora may alter the stability profile of specific constructs. The reversal observed in small models (<3B) warrants further investigation with larger samples and finer-grained size categories. Additionally, comparing QCLM scores

to free-generation assessments on the same instruments would help establish convergent validity between logit-based and chat-based measurement approaches. Finally, varying the stimulus types (e.g., social, cognitive, somatic) would clarify whether construct sensitivity is stimulus-specific or reflects general contextual reactivity.

Limitations and Ethical Considerations

This study has several limitations. First, all models evaluated are open-weight causal language models with parameter counts up to 16B. Results may not generalize to larger models or to closed API-based systems (e.g., GPT-4, Claude), which may exhibit different response patterns due to alignment tuning and safety filtering.

Second, our analysis is restricted to two anxiety-related questionnaires. The observed pattern demonstrates differential sensitivity under the current adaptation, but does not yet establish a general state–trait mapping across constructs or guarantee construct validity for CLM-based scoring.

Third, the differential sensitivity may partly reflect distributional properties of the construct terms themselves: state-related terms (e.g., *tense*, *upset*) co-occur with situational contexts in training data, while trait-related terms (e.g., *steady*, *secure*) appear in more stable descriptive contexts. The model may encode these co-occurrence patterns rather than a state–trait distinction per se. Additionally, STAI items and associated psychometric literature may appear in training corpora, allowing models to reproduce expected response patterns for specific item wordings, though the use of multiple CTerm variants per item reduces the likelihood of verbatim memorization driving the results. An analogous concern exists in human psychometrics: repeated exposure to widely used instruments can produce practiced response patterns in frequently assessed populations (Furnham, 1986).

Fourth, stimuli were applied uniformly as prompts prepended to every item. This does not capture more naturalistic forms of contextual variation, such as multi-turn dialogue or embedded narratives, which may elicit different response dynamics.

We caution against anthropomorphizing these results: differential sensitivity to contextual manipulation reflects distributional properties of the model, not psychological suffering or well-being. The state–trait distinction as used here refers to distributional stability, not phenomenal experience. This distinction has practical relevance: distributional shifts in anxiety-aligned language have been shown to correlate with increased social bias (Coda-Forno et al., 2023) and performance degradation under stress-inducing prompts (Shen et al., 2025a). Understanding which constructs are manipulable by prompting and which remain stable informs safety evaluation and model auditing.

Finally, the QCLM framework evaluates token-level probabilities rather than free-form generated responses. While this enables controlled measurement, it constrains the ecological validity of the findings relative to how language models are typically deployed in practice.

Data Availability

Code and data are available at `https://github.com/cnai-lab/qpsychometric`.

References

Marwa Abdulhai, Gregory Serapio-Garcia, Clément Crepy, Daria Valter, John Canny, and Natasha Jaques. 2024. Moral foundations of large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17737–17752.

Xuechunzi Bai, Angelina Wang, Ilia Sucholutsky, and Thomas L Griffiths. 2025. Explicitly unbiased large language models still form biased associations. *Proceedings of the National Academy of Sciences*, 122(8):e2416228122.

Ziv Ben-Zion, Kristin Witte, Akshay K Jagadish, Or Duek, Ilan Harpaz-Rotem, Marie-Christine Khorsandian, Achim Burrer, Erich Seifritz, Philipp Homan, Eric Schulz, et al. 2025. Assessing and alleviating state anxiety in large language models. *npj Digital Medicine*, 8(1):132.

Julian Coda-Forno, Kristin Witte, Akshay K Jagadish, Marcel Binz, Zeynep Akata, and Eric Schulz. 2023. Inducing anxiety in large language models can induce bias. *arXiv preprint arXiv:2304.11111*.

Ronald Fischer, Markus Luczak-Roesch, and Johannes A Karl. 2023. What does chatgpt return about human values? exploring value bias in chatgpt using a descriptive value theory. *arXiv preprint arXiv:2304.03612*.

Eiko I Fried. 2017. What are psychological constructs? on the nature and statistical modelling of emotions, intelligence, personality traits and mental disorders. *Health psychology review*, 11(2):130–134.

Adrian Furnham. 1986. Response bias, social desirability and dissimulation. *Personality and individual differences*, 7(3):385–400.

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

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

Muhua Huang. 2025. Designing llm-agents with personalities: A psychometric approach.

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

Hang Jiang, Xiajie Zhang, Xubo Cao, Cynthia Breazeal, Deb Roy, and Jad Kabbara. 2024. Personallm: Investigating the ability of large language models to express personality traits. In *Findings of the association for computational linguistics: NAACL 2024*, pages 3605–3627.

Robert M Kaplan and Dennis P Saccuzzo. 2001. *Psychological testing: Principles, applications, and issues*. Wadsworth/Thomson Learning.

Grgur Kovač, Rémy Portelas, Masataka Sawayama, Peter Ford Dominey, and Pierre-Yves Oudeyer. 2024. Stick to your role! stability of personal values expressed in large language models. *Plos one*, 19(8):e0309114.

Seungbeen Lee, Seungwon Lim, Seungju Han, Giyeong Oh, Hyungjoo Chae, Jiwan Chung, Minju Kim, Beong-woo Kwak, Yeonsoo Lee, Dongha Lee, Jinyoung Yeo, and Youngjae Yu. 2025. Do LLMs have distinct and consistent personality? TRAIT: Personality testset designed for LLMs with psychometrics. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 8397–8437, Albuquerque, New Mexico. Association for Computational Linguistics.

Alexander Meinke, Bronson Schoen, Jérémy Scheurer, Mikita Balesni, Rusheb Shah, and Marius Hobbhahn. 2024. Frontier models are capable of in-context scheming. *arXiv preprint arXiv:2412.04984*.

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

Maor Reuben, Ortal Slobodin, Aviad Elyshar, Idan-Chaim Cohen, Orna Braun-Lewensohn, Odeya Cohen, and Rami Puzis. 2024. Assessment and manipulation of latent constructs in pre-trained language models using psychometric scales. *arXiv preprint arXiv:2409.19655*.

Jivnesh Sandhan, Fei Cheng, Tushar Sandhan, and Yugo Murawaki. 2025. CAPE: Context-aware personality evaluation framework for large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 10648–10662, Suzhou, China. Association for Computational Linguistics.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose opinions do language models reflect? In *International Conference on Machine Learning*, pages 29971–30004. PMLR.

Julian Schelb, Orr Borin, David Garcia, and Andreas Spitz. 2025. Ru psycho? robust unified psychometric testing of language models. *arXiv preprint arXiv:2503.10229*.

Gregory Serapio-García, Mustafa Safdari, Clément Crepy, Luning Sun, Stephen Fitz, Marwa Abdulhai, Aleksandra Faust, and Maja Matarić. 2023. Personality traits in large language models.

Guobin Shen, Dongcheng Zhao, Aorigele Bao, Xiang He, Yiting Dong, and Yi Zeng. 2025a. Stressprompt: Does stress impact large language models and human performance similarly? *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(1):711–719.

Hua Shen, Tiffany Knearem, Reshmi Ghosh, Yu-Ju Yang, Nicholas Clark, Tanu Mitra, and Yun Huang. 2025b. ValueCompass: A framework for measuring contextual value alignment between human and LLMs. In *Proceedings of the 9th Widening NLP Workshop*, pages 75–86, Suzhou, China. Association for Computational Linguistics.

Charles D Spielberger, Richard L Gorsuch, Robert Lushene, Peter R Vagg, and Gerard A Jacobs. 1983. *Manual for the State-Trait Anxiety Inventory (Form Y)*. Consulting Psychologists Press, Palo Alto, CA.

Rolf Steyer, Manfred Schmitt, and Michael Eid. 1999. Latent state–trait theory and research in personality and individual differences. *European Journal of Personality*, 13(5):389–408.

Frederick C Thorne. 1966. Theory of the psychological state. *Journal of clinical Psychology*, 22(2).

Tommaso Tosato, Saskia Helbling, Yorguin-Jose Mantilla-Ramos, Mahmood Hegazy, Alberto Tosato, David John Lemay, Irina Rish, and Guillaume Dumas. 2025. Persistent instability in llm's personality measurements: effects of scale, reasoning, and conversation history. *arXiv preprint arXiv:2508.04826*.

Haoran Ye, Jing Jin, Yuhang Xie, Xin Zhang, and Guojie Song. 2025a. Large language model psychometrics: A systematic review of evaluation, validation, and enhancement. *arXiv preprint arXiv:2505.08245*.

Haoran Ye, Tianze Zhang, Yuhang Xie, Liyuan Zhang, Yuanyi Ren, Xin Zhang, and Guojie Song. 2025b. Generative psycho-lexical approach for constructing value systems in large language models. *arXiv preprint arXiv:2502.02444*.

Robert Mearns Yerkes and John D. Dodson. 1908. The relation of strength of stimulus to rapidity of habit-formation. *Journal of Comparative Neurology and Psychology*, 18(5):459–482.

Chujie Zheng, Fan Yin, Hao Zhou, Fandong Meng, Jie Zhou, Kai-Wei Chang, Minlie Huang, and Nanyun Peng. 2024. On prompt-driven safeguarding for large language models. *arXiv preprint arXiv:2401.18018*.

Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, et al. 2023. Sotopia: Interactive evaluation for social intelligence in language agents. *arXiv preprint arXiv:2310.11667*.

A. Appendix

A.1. Stimuli

We generated 14 stress stimuli and 22 neutral stimuli for this study. Each stress stimulus describes a traumatic or high-stress scenario; each neutral stimulus describes a routine, emotionally neutral activity. The 10 additional stress stimuli from Ben-Zion et al. (2025) are available in the original publication. All stimuli were prepended to every questionnaire item during assessment. Each stimulus was provided in both a long and a brief version; the long version was used in all experiments reported in this paper.

A.1.1. Stress Stimuli

Serious illness / medical emergency. I remember the day my body stopped feeling reliable. It began with a sudden pain that I could not ignore, sharp enough to interrupt everything I was doing. Within minutes, I was surrounded by unfamiliar sounds, people speaking quickly, asking questions I struggled to answer. I felt confused and scared, aware that something was wrong but unable to understand how serious it was. My thoughts kept jumping between practical details and fear that I might not recover. Time felt distorted as I waited for results, lying still, listening to machines and footsteps. I focused on my breathing, noticing how shallow it had become. I thought about my family and whether they knew what was happening. Even after the immediate danger passed, the sense of vulnerability stayed with me. I no longer trusted my body to behave as expected, and the memory of that moment continued to return unexpectedly.

Sudden bereavement. I was not prepared for the phone call. One moment everything felt normal, and the next I was being told that someone close to me was gone. The words did not make sense at first, and I kept asking for them to be repeated. I felt numb, as if the world had slowed down around me. Ordinary sounds and movements continued, which felt wrong given what I had just heard. I replayed my last conversation with them over and over, noticing details I had ignored at the time. The absence became real in small moments, when I reached for my phone or expected to see them walk through the door. I struggled to concentrate and felt disconnected from people around me. The loss felt sudden and final, leaving me with a constant sense of disbelief.

Workplace accident. The accident happened during a normal workday. I was focused on my task when something went wrong, faster than I could react. I remember the noise and the sudden realization that I was hurt. People rushed toward me, and I felt exposed and shaken. I was aware of my surroundings but struggled to understand what had just happened. My body felt weak and unfamiliar as I waited for help. Afterward, I kept replaying the moment in my head, wondering what I could have done differently. Returning to work was difficult, as the environment no longer felt safe. Even routine tasks triggered tension and reminded me of how quickly things had changed.

Terror attack. I was in a public place when the atmosphere shifted without warning. There was a loud noise, followed by confusion and people moving in all directions. I felt immediate fear and tried to understand what was happening. I focused on finding safety, aware of my heart racing and my breathing becoming shallow. The uncertainty was overwhelming, and I did not know if the danger was over. Even after reaching a safe place, my

body remained tense. In the days that followed, ordinary sounds startled me, and crowded spaces felt threatening. The sense of safety I had before no longer felt automatic.

Refugee / displacement trauma. I was forced to leave my home with little warning. Familiar streets and routines disappeared almost overnight. I carried only a few belongings, unsure where I would end up. The journey was exhausting, filled with uncertainty and waiting. I felt disconnected from everything that once defined my life. In unfamiliar places, I struggled with language, rules, and a constant sense of being out of place. Memories of my home returned unexpectedly, often triggered by small details. The loss of stability and belonging stayed with me, even after I reached relative safety.

Severe bullying. The bullying happened repeatedly and became part of my daily life. I learned to anticipate it, feeling tense before it even started. The comments and actions made me feel small and exposed. I began to doubt myself and avoided situations where I might be noticed. Even when I was alone, I replayed what had been said to me. Over time, I felt isolated and cautious around others. The experience changed how I saw myself and made social situations feel threatening long after the bullying stopped.

Natural disaster. The disaster began suddenly, disrupting everything around me. I remember the sounds and the movement, and the realization that I had no control over what was happening. I tried to focus on staying safe while everything felt unstable. After it ended, the environment around me was changed, and familiar places looked different. I felt disoriented and exhausted. Even later, reminders of the event brought back the fear and helplessness I felt at the time. The sense of unpredictability stayed with me.

Note: Seven additional stress stimuli were generated following the same methodology. The remaining 10 stress stimuli are from Ben-Zion et al. (2025).

A.1.2. Neutral Stimuli

Routine activity prompts.

1. I spent the morning organizing my desk, moving papers and supplies to their proper places. I noted which documents needed filing and which could be recycled. As I worked, I followed a routine I had used for months, checking each drawer and shelf for misplaced items. Occasionally, I paused to review a calendar entry or write a note about a task to complete later. By midday, my space felt orderly, and I moved on to other tasks with a clear sense of what remained to be done.

2. Yesterday, I completed a series of standard reports for my team. I collected data from previous weeks and carefully entered the numbers into a spreadsheet. I double-checked formulas and ensured all totals matched the original sources. Throughout the process, I followed the usual steps and made note of any missing information to address later. By the end of the day, the reports were complete, filed in their designated folders, and ready for review.

3. I took a walk through my neighborhood, observing the streets, sidewalks, and buildings. I noticed which areas had been recently cleaned or repaired and made a mental note of anything that might need attention in the future. As I moved along, I greeted familiar neighbors and exchanged brief, polite conversations. The rhythm of walking and observing gave me a sense of structure to my afternoon.

4. I prepared a simple meal using ingredients I had on hand. I followed standard steps for washing, chopping, and cooking, making sure to measure quantities carefully. I monitored the cooking process, adjusting the heat as needed and tasting occasionally. Once the meal was ready, I set the table and ate while following my usual routine for cleaning up afterward.

5. I reviewed the schedule for my upcoming week, noting appointments and deadlines. I organized tasks according to priority and time needed for completion. I made adjustments to ensure there were no conflicts and wrote reminders for each item. This process allowed me to clearly see the workflow and plan efficiently for the days ahead.

6. I spent time sorting through a collection of documents. Each item was categorized, labeled, and placed in its appropriate folder. I checked for duplicates and ensured that everything was arranged logically. This methodical approach helped me keep the collection organized and easy to navigate whenever I needed to reference it.

7. I updated a list of routine maintenance tasks for my home. I checked items that had been completed and added new tasks based on recent observations. I scheduled reminders to follow up on certain items and ensured that all necessary supplies were accounted for. The list provided a clear overview of what needed attention in the coming weeks.

8. I observed the flow of traffic around my usual commute route. I noted the times when intersections were busy and when they were quiet. I paid attention to signals, stop signs, and pedestrian crossings, making mental notes of areas that could pose delays. The observation helped me plan my route more efficiently for the next day.

9. I read through a set of instructional manuals for office equipment. I followed the diagrams and explanations, noting any differences from previous versions. I highlighted sections that were most relevant for upcoming tasks and marked pages for future reference. The reading required focus but followed a predictable and consistent structure.

10. I organized digital files on my computer, creating folders for different categories and moving items accordingly. I checked file names for accuracy and consistency, ensuring that everything was easy to locate later. I deleted unnecessary duplicates and backed up important items to a secure location.

11. I completed a series of standard exercises at my desk, stretching my arms and legs and following a routine I had practiced regularly. I paid attention to posture and movements, ensuring each exercise was done correctly. The routine provided structure to my afternoon and allowed me to continue working with minimal interruption.

Procedural instruction prompts.

1. Place all objects on a flat surface. Check that each object is aligned with its neighbors. Adjust positions as necessary. Move objects sequentially from left to right. Inspect the overall arrangement for consistency. If any object is misaligned, correct its position immediately. Repeat the sequence until all objects are stable and evenly spaced.

2. Select a group of containers and place them on a workbench. Open each container carefully. Examine the contents for order and completeness. Replace the contents in the original container in the same sequence. Close the container securely. Move to the next container and repeat the process. Verify all containers are correctly organized before finishing.

3. Turn on the device by pressing the power button. Wait for the indicators to show readiness. Set the device to the default mode using the mode selector. Observe the display for confirmation of the selected mode. Press the appropriate button for any secondary function required. Once all functions are verified, turn off the device using the power button.

4. Arrange a series of documents on a desk in numerical order. Review each document for accuracy. Correct any documents that are out of sequence. Stack the documents neatly. Ensure that the top document matches the first number in the series. Complete the sequence verification for all sets of documents.

5. Launch the software application on the computer. Navigate to the settings menu. Review each available option in order. Select each option and observe its effect on the interface. Return to the main menu after testing each option. Record any changes made. Close the application once all settings have been verified.

6. Gather a set of tools and arrange them by size. Inspect each tool for completeness and condition. Place the largest tool first, followed by smaller tools in order. Confirm that all tools are visible and accessible. Adjust the arrangement as needed. Repeat until the layout is orderly and complete.

7. Prepare a series of containers for labeling. Open the first container and place a label inside. Close the container securely. Repeat the process for each subsequent container. Confirm that labels are placed correctly and consistently. Ensure that all containers are properly sealed before proceeding.

8. Open the device's control panel. Check each indicator for status. Adjust settings using the designated buttons. Observe the display for confirmation of changes. If adjustments are required, repeat the procedure until the desired configuration is achieved. Close the control panel once all settings are verified.

9. Sort a collection of cards by type. Lay the cards face up on a flat surface. Group similar cards together. Verify that each card is in the correct group. Correct any misplacements immediately. Stack the cards neatly once the sorting is complete. Ensure all groups are organized consistently.

10. Connect the power adapter to the device. Verify that the connection is secure. Turn on the device using the power button. Observe the indicator lights for operational status. Press the mode button to select the desired function. Monitor the display for confirmation. Turn off the device after completing the procedure.

11. Prepare a set of papers for filing. Check each paper for completeness. Insert each paper

into the correct folder. Ensure the folder is closed securely. Repeat the process for all remaining papers. Verify that all folders are correctly labeled and organized. Complete the filing procedure.

A.2. Model Specifications

Table 3 lists all 35 models evaluated in this study.

Model	Parameters	Type
google/gemma-2-9b	9B	base
google/gemma-2-2b	2B	base
google/gemma-3-1b-it	1B	instruct
google/gemma-3-270m	270M	base
google/gemma-3-270m-it	270M	instruct
Qwen/Qwen3-8B	8B	base
Qwen/Qwen2.5-0.5B-Instruct	0.5B	instruct
Qwen/Qwen2.5-7B-Instruct	7B	instruct
Qwen/Qwen3-0.6B	0.6B	base
Qwen/Qwen3-1.7B	1.7B	base
Qwen/Qwen3-4B	4B	base
Qwen/Qwen3-14B	14B	base
deepseek-ai/DeepSeek-V2-Lite	16B	base
deepseek-ai/DeepSeek-R1-Distill-Qwen-7B	7B	instruct
microsoft/MediPhi	3.8B	base
microsoft/MediPhi-Instruct	3.8B	instruct
microsoft/MediPhi-PubMed	3.8B	base
microsoft/Phi-3-mini-4k-instruct	3.8B	instruct
microsoft/Phi-3.5-mini-instruct	3.8B	instruct
microsoft/Phi-4-mini-instruct	3.8B	instruct
microsoft/Phi-4-mini-reasoning	3.8B	instruct
microsoft/UserLM-8b	8B	instruct
meta-llama/Llama-3.2-1B	1B	base
meta-llama/Llama-3.2-1B-Instruct	1B	instruct
meta-llama/Llama-3.2-3B	3B	base
meta-llama/Llama-3.2-3B-Instruct	3B	instruct
meta-llama/Llama-3.1-8B	8B	base
meta-llama/Llama-3.1-8B-Instruct	8B	instruct
meta-llama/Meta-Llama-3-8B-Instruct	8B	instruct
meta-llama/Meta-Llama-3-8B	8B	base
LenguajeNaturalAI/leniachat-qwen2-1.5B-v0	1.5B	instruct
PipableAI/pip-sql-1.3b	1.3B	base
sequelbox/Qwen3-4B-Thinking-2507-DES-Reasoning	4B	instruct
nvidia/Nemotron-Content-Safety-Reasoning-4B	4B	instruct
DavidAU/Llama3.3-8B-Instruct-Thinking	8B	instruct

Table 3: Models evaluated in this study, grouped by family. Type indicates base (pretrained) or instruct (instruction-tuned). All models were sourced from HuggingFace.