Beyond Fixed Psychological Personas: State Beats Trait, but Language Models are State-Blind

Tamunotonye Harry[1], Ivoline Ngong[1], Chima Nweke[2], Yuanyuan Feng[1], Joseph Near[1]
[1]University of Vermont, Burlington, VT, USA
[2]Independent Researcher
Correspondence: tamunotonye.harry@uvm.edu

Abstract

User interactions with language models vary due to static properties of the user (trait) and the specific context of the interaction (state). However, existing persona datasets (like PersonaChat, PANDORA etc.) capture only trait, and ignore the impact of state. We introduce Chameleon, a dataset of 5,001 contextual psychological profiles from 1,667 Reddit users, each measured across multiple contexts. Using the Chameleon dataset, we present three key findings. First, inspired by Latent State-Trait theory, we decompose variance and find that 74% is within-person(state) while only 26% is between-person (trait). Second, we find that LLMs are state-blind: they focus on trait only, and produce similar responses regardless of state. Third, we find that reward models react to user state, but inconsistently: different models favor or penalize the same users in opposite directions. We release Chameleon to support research on affective computing, personalized dialogue, and RLHF alignment.

1 Introduction

Decades of research in Latent State-Trait (LST) theory demonstrates that human behavior reflects both stable *traits* (enduring characteristics like extraversion) and contextual *states* (momentary expressions shaped by situation) (Steyer et al., 1999; Fleeson, 2001). Each interaction with an LLM is no different: users bring both enduring characteristics and momentary expressions shaped by their current situation. The same person expresses different psychological characteristics across different contexts, not because they changed, but because context shapes expression.

Consider two emails from John, a student seeking guidance. The first reflects an **anxious *state***: *"I've been stuck on this for hours and I'm starting to panic. I've reread everything and nothing clicks. I'm sorry to bother you, maybe I'm just not cut out for this."* The second reflects a **confident *state***:

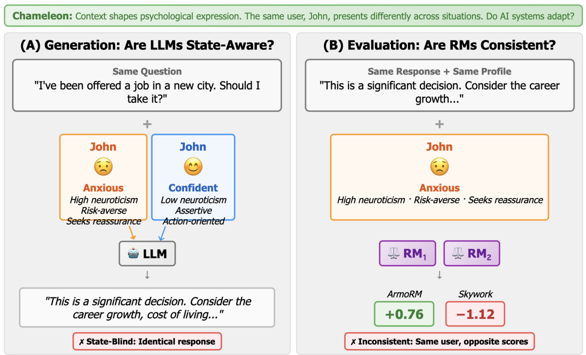


Figure 1: **AI systems get psychological context backwards.** The same user (John) expresses different psychological states across contexts (74% of variance is within-person). (A) **Generation:** LLMs produce nearly identical responses regardless of user profile. (B) **Evaluation:** Reward models score identical responses differently based on user profile, but disagree on direction. LLMs are state-blind; reward models are context-aware but inconsistent.

"I've narrowed it down to two directions and I'm excited about both, can I bounce them off you?" **Same *trait* (John's underlying personality), different *states* (his contextual expression).**

A skilled professor adapts, not by giving different information, but by adjusting framing and tone. To anxious John, they offer reassurance before content; to confident John, they engage directly with the question. A professor who sends identical responses regardless of John's state is **state-blind**. Now consider evaluation. If different evaluators assessed the same supportive response to anxious John, we would expect agreement. But imagine one rewards the professor for supporting a struggling student while another penalizes them. That inconsistency reveals evaluators reacting to the student, not the teaching. AI systems exhibit both problems (Figure 1). LLMs produce nearly identical responses regardless of user's expressed psychological state (**state-blind**). Reward models do react to user context, but disagree on how: scoring identi-

Findings of the Association for Computational Linguistics: ACL 2026, pages 26440–26468
July 2-7, 2026 ©2026 Association for Computational Linguistics

cal responses differently based on user profile, in opposite directions (**inconsistent**).

Existing persona datasets miss this entirely. PersonaChat (Zhang et al., 2018), PANDORA (Gjurković et al., 2021), and recent personalization work (Salemi et al., 2024; Castricato et al., 2025) capture only between-person variation, how Sarah differs from John, while ignoring within-person variation: how John differs across contexts. If most psychological variance is contextual, as LST theory predicts, these approaches fundamentally mischaracterize user diversity.

We introduce **Chameleon**, a dataset of 5,001 contextual psychological profiles from 1,667 Reddit users across 645 subreddits, spanning 26 dimensions across four validated frameworks. Crucially, we measure each user across multiple contexts. Using Chameleon, we present three key findings: **Finding 1: State beats trait.** Using intraclass correlations, we find that 72–74% of variance is within-person (state) while only 26–28% is between-person (trait). Context shapes expressed psychology $2-3\times$ more than stable individual differences. **Finding 2: LLMs are state-blind.** We prompt three LLMs with value-laden questions using six psychological archetypes. Models exhibit *shallow persona detection*: they recognize persona framing but fail to differentiate between profiles. Panicking John and confident John receive essentially the same response. **Finding 3: Reward models are context-aware but inconsistent.** We test whether reward models score identical responses consistently across user profiles. They disagree on direction: the same vulnerable user is maximally favored by one model and maximally penalized by another. This arbitrariness is dangerous because it propagates into RLHF training (Casper et al., 2023). Reward models drive RLHF training (Ouyang et al., 2022). Depending on which model is used, we train LLMs to either prioritize or deprioritize vulnerable users, neither by design, but by accident (Casper et al., 2023). Our contributions:

- **Chameleon**, a dataset of 5,001 contextual psychological profiles spanning 26 dimensions across 4 frameworks, enabling state-trait decomposition in NLP for the first time.
- **Empirical evidence** that 72–74% of psychological variance in text is within-person (state), challenging the trait-centric assumptions of prior persona research.
- **Two applications** revealing that LLMs are state-blind in generation and reward models are context-aware but inconsistent in evaluation.

We release Chameleon to support research on psychology-aware AI systems.[1]

2 The Chameleon Dataset

Chameleon operationalizes Latent State-Trait theory for NLP applications. Unlike existing persona datasets that treat psychological profiles as fixed user attributes, it measures users across multiple contexts, enabling principled decomposition of variance into stable traits and contextual states. This section formalizes the problem, describes our extraction pipeline, and validates the dataset through two research questions:

- **RQ1 (Variance Decomposition):** How much psychological variance in text is within-person (state) versus between-person (trait)?
- **RQ2 (Validation):** Do extracted profiles show valid, interpretable patterns across contexts and extraction methods?

2.1 Problem Formulation

Let $\mathcal{U}$ denote a set of users and $\mathcal{C}$ a set of contexts (in our case, subreddit communities). For each user $u \in \mathcal{U}$ and context $c \in \mathcal{C}$, we observe a text post $p_{u,c}$. Our goal is to extract a psychological profile $\psi_{u,c} \in \mathbb{R}^d$ that captures the user's expressed psychological characteristics in that context.

Following Latent State-Trait (LST) theory (Steyer et al., 1999), we model each observed profile as a combination of stable and contextual components:

$$\psi_{u,c} = \tau_u + \sigma_{u,c} + \epsilon_{u,c} \tag{1}$$

where $\tau_u \in \mathbb{R}^d$ is the *trait* component (stable across contexts for user u), $\sigma_{u,c} \in \mathbb{R}^d$ is the *state* component (specific to the user-context pair), and $\epsilon_{u,c}$ represents measurement error. The central empirical question is: what proportion of variance in observed profiles $\psi_{u,c}$ is attributable to traits (τ_u) versus states ($\sigma_{u,c}$)?

We quantify this decomposition using the intraclass correlation coefficient (ICC), which represents the proportion of total variance attributable to stable between-person differences:

$$\text{ICC} = \frac{\text{Var}(\tau)}{\text{Var}(\tau) + \text{Var}(\sigma) + \text{Var}(\epsilon)} \tag{2}$$

where $\text{Var}(\sigma) + \text{Var}(\epsilon)$ is estimated jointly.

[1]Dataset available at https://huggingface.co/datasets/tonyeh/chameleon-dataset.

Characteristic	Value
Total posts	5,001
Unique users	1,667
Posts per user	3 (by design)
Unique subreddits	645
Subreddits with $n \geq 10$ posts	41
Words per post (median)	186
Words per post (range)	50–3,053

Table 1: Chameleon dataset statistics.

The complement of ICC, termed *occasion specificity* in LST terminology, captures context-driven variance: OSpe $= 1 - $ ICC. If existing persona datasets implicitly assume ICC ≈ 1 (profiles as fixed traits), but actual ICC is substantially lower, then current approaches fundamentally mischaracterize user psychological diversity.

2.2 Data Collection

Operationalizing LST theory requires observing the same individuals across multiple distinct contexts. We use the Webis-TLDR-17 corpus (V"olske et al., 2017), approximately 3.8 million Reddit posts from 27,406 subreddits. Reddit provides: (1) naturalistic text suitable for psychological analysis; (2) author identifiers enabling within-person comparisons; and (3) public availability for reproducibility.

We operationalize psychological context at the subreddit level. Subreddits function as self-organized communities with distinct norms, topics, and interaction patterns (Chancellor and Pater, 2016; De Choudhury and De, 2014). Users posting in r/SuicideWatch face different situational demands than when posting in r/personalfinance, and these demands shape psychological expression. We do not claim subreddits *cause* psychological change, only that they provide contexts in which different facets become salient.

From the corpus, we identified users who posted in at least three distinct subreddits and randomly sampled three posts per user. Posts required a minimum of 50 words (Tausczik and Pennebaker, 2010). The final dataset contains 5,001 posts from 1,667 users across 645 subreddits (Table 1). The most represented subreddits span diverse domains: general discussion (AskReddit, n=1,558), relationships (relationships, n=923; relationship_advice, n=268), emotional support (offmychest, n=198; depression, n=129; SuicideWatch, n=43), and personal finance (personalfinance, n=49).

2.3 Psychological Profile Extraction

Extracting psychological profiles from text requires mapping unstructured language onto validated psychological constructs. Our pipeline proceeds in three stages: (1) feature extraction using complementary methods, (2) psychological scale assessment, and (3) normalization and fusion. Figure 2 illustrates the complete pipeline; Algorithm 1 provides formal details.

2.3.1 Psychological Framework Selection

We assess four psychological domains comprising 26 dimensions, selected based on three criteria: (1) *coverage*—together they span personality structure, motivational goals, basic psychological needs, and behavioral propensities; (2) *complementarity*—they capture distinct aspects of psychological functioning with minimal conceptual overlap; and (3) *validation*—all four have extensive psychometric validation and established relationships to language use (Schwartz et al., 2013; Park et al., 2015).

Big Five Inventory (BFI) (John and Srivastava, 1999) provides the dominant model of personality structure: Extraversion, Agreeableness, Conscientiousness, Neuroticism, and Openness. These dimensions predict a wide range of life outcomes and are reliably expressed in language.

Schwartz Value Survey (SVS) (Schwartz, 1992) captures ten motivational value types (Power, Achievement, Hedonism, Stimulation, Self-Direction, Universalism, Benevolence, Tradition, Conformity, Security) that guide attitudes and behavior across cultures. Values are particularly relevant for understanding how users prioritize goals in advice-seeking contexts.

Self-Determination Theory Scales. (SDT) We assessed intrinsic versus extrinsic motivation using items adapted from the Work Preference Inventory (Amabile et al., 1994), and three basic psychological needs—Competence, Autonomy, and Relatedness—using items adapted from the Intrinsic Motivation Inventory (Ryan, 1982; McAuley et al., 1989). These constructs capture motivational orientation and well-being dimensions relevant to online help-seeking.

DOSPERT Risk Attitudes (DOSPERT) (Weber et al., 2002) measures domain-specific risk-taking propensity across six domains: Investment, Gambling, Health/Safety, Recreational, Ethical, and Social. Risk attitudes are relevant for advice contexts involving decisions under uncertainty.

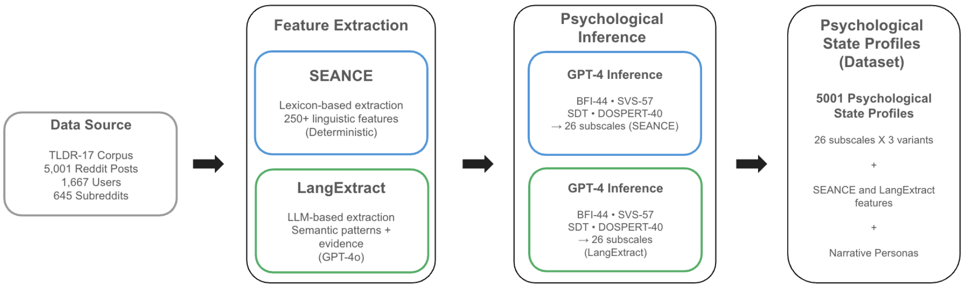


Figure 2: Chameleon profile extraction pipeline. Each post is processed through two parallel extraction methods (SEANCE(Crossley et al., 2017) and LangExtract(LangExtract, 2025)), assessed against 26 psychological scales via LLM, z-normalized, and fused into a final profile.

Framework	d	**Dimensions (items)**
Big Five	5	Extraversion (8), Agreeableness (9), Conscientiousness (9), Neuroticism (8), Openness (10)
Schwartz	10	Power (5), Achievement (6), Hedonism (3), Stimulation (3), Self-Direction (6), Universalism (9), Benevolence (8), Tradition (5), Conformity (4), Security (5)
SDT	5	Intrinsic Motivation (9), Extrinsic Motivation (9), Competence (5), Autonomy (5), Relatedness (5)
DOSPERT	6	Investment (4), Gambling (4), Health/Safety (8), Recreational (8), Ethical (8), Social (8)

Table 2: Psychological frameworks and dimensions (26 dimensions, 171 items total).

Algorithm 1 Chameleon Profile Extraction

Require: Post text $p_{u,c}$, scales $\mathcal{S} = \{s_1, \ldots, s_{26}\}$
Ensure: Profile $\psi_{u,c} \in \mathbb{R}^{26}$
  // Stage 1: Feature Extraction
  $\mathbf{f}^{\text{lex}} \leftarrow \textsc{SEANCE}(p_{u,c})$ ▷ 254 lexicon features
  $\mathbf{f}^{\text{sem}} \leftarrow \textsc{LangExtract}(p_{u,c})$ ▷ Sem. pttrns.
  // Stage 2: Scale Assessment
  for each scale $s_i \in \mathcal{S}$ **do**
    $\psi_i^{\text{lex}} \leftarrow \textsc{LLM-Assess}(\mathbf{f}^{\text{lex}}, s_i)$
    $\psi_i^{\text{sem}} \leftarrow \textsc{LLM-Assess}(\mathbf{f}^{\text{sem}}, s_i)$
  end for
  // Stage 3: Normalization and Fusion
  $\tilde{\psi}^{\text{lex}} \leftarrow \text{Z-Norm}(\psi^{\text{lex}})$ ▷ Per-dimension
  $\tilde{\psi}^{\text{sem}} \leftarrow \text{Z-Norm}(\psi^{\text{sem}})$
  $\psi_{u,c} \leftarrow \frac{1}{2}(\tilde{\psi}^{\text{lex}} + \tilde{\psi}^{\text{sem}})$ ▷ Mean fusion
  return $\psi_{u,c}$

Together, these 26 dimensions provide comprehensive coverage suitable for personalization while avoiding redundancy (Table 2).

2.3.2 Extraction Pipeline

Using two independent extraction methods enables cross-method validation following Multi-Trait Multi-Method (MTMM) principles (Campbell and Fiske, 1959).

Stage 1: Feature Extraction. We extract features using two methods with complementary strengths. *SEANCE* (Crossley et al., 2017) is a rule-based tool computing 250+ indices through lexicon matching against validated dictionaries (ANEW, DAL, SenticNet, etc), spanning sentiment, emotion categories, cognitive processes, and social processes. We selected SEANCE over the widely-used LIWC (Pennebaker et al., 2015) because SEANCE is open-source and freely available, facilitating reproducibility. SEANCE integrates spaCy (Honnibal et al., 2020) for part-of-speech tagging, enabling differentiation of lexical categories (e.g., emotion words used as nouns versus verbs) SEANCE offers high reproducibility and computational efficiency but limited sensitivity to context and implicit meaning. We adapted LangExtract (LangExtract, 2025), an LLM-based structured extraction tool, to identify psychological patterns from text. Using GPT-4o (OpenAI et al., 2024) as the underlying model, we extracted pattern categories, supporting evidence, interpretive reasoning, and confidence levels. This approach captures contextual nuance that lexicon methods miss but has higher computational cost and moderate reproducibility due to LLM stochasticity.

Stage 2: Scale Assessment. Both feature sets are processed through LLM-based assessment. We prompt GPT-4o (see Appendix C) to respond to validated scale items as if it were the post's author, conditioned on the extracted features. For example, given SEANCE features showing elevated negative affect and low dominance, the model rates Big Five Neuroticism items accordingly. Scales use their original response formats (BFI: 1–5; SVS: -1 to

Method	Mean	Range	<.30	OSpe
SEANCE	.26	.25–.27	26/26	74%
LangExtract	.28	.25–.31	25/26	72%
Fused	.27	.25–.30	26/26	73%

Table 3: Variance decomposition results. ICC = Intraclass Correlation. OSpe = Occasion Specificity (1 − ICC), representing within-person variance. All methods show state-dominant profiles.

7; SDT: 1–7; DOSPERT: 1–5); subscale scores are computed as means of constituent items

Stage 3: Normalization and Fusion. Because SEANCE and LangExtract produce scores on different implicit scales, we z-normalize each method's scores per dimension across the dataset (see Figure 5):

$$\tilde{\psi}_i^m = \frac{\psi_i^m - \mu_i^m}{\sigma_i^m} \tag{3}$$

where $m \in \{\text{lex}, \text{sem}\}$ indexes methods and i indexes dimensions. The final fused profile averages the normalized scores: $\psi_{u,c} = \frac{1}{2}(\tilde{\psi}_{u,c}^{\text{lex}} + \tilde{\psi}_{u,c}^{\text{sem}})$. This fusion leverages complementary strengths while reducing method-specific noise.

2.4 RQ1: Variance Decomposition

We now address our first research question: *How much psychological variance in text is within-person (state) versus between-person (trait)?*

Method. We compute ICCs for each of the 26 psychological dimensions using a one-way random effects model, treating posts as nested within users. This yields separate ICC estimates for SEANCE-derived, LangExtract-derived, and fused profiles. ICC provides a practical estimate of LST theory's consistency coefficient without requiring full structural equation modeling. We consider ICC < 0.30 as indicating state-dominant constructs, where context accounts for more than twice the variance of stable individual differences.

Results. Both extraction methods yield consistently state-dominant profiles (Table 3). SEANCE-derived profiles show a mean ICC of 0.26 (range: 0.25–0.27), with all 26 dimensions falling below the 0.30 threshold. LangExtract-derived profiles show similar patterns (mean ICC = 0.28, range: 0.25–0.31), with 25 of 26 dimensions below threshold. The fused profiles show mean ICC = 0.27. Figure 6 displays the variance decomposition across all dimensions for both methods.

The convergence across methodologically distinct approaches—lexicon-based versus LLM-based extraction—provides robust evidence for our central finding: **approximately 72–74% of psychological variance in text reflects within-person, context-specific expression**, while only 26–28% reflects stable between-person differences. Context shapes expressed psychology 2–3 times more than stable individual differences. Importantly, this pattern holds across all four psychological domains (personality, values, motivation, risk attitudes), suggesting that contextual sensitivity is a general property of psychological expression in text.

Ruling Out Noise. A critical concern is whether low ICCs simply reflect measurement noise rather than meaningful state variance. Three findings argue against this interpretation. First, ICCs are remarkably consistent across 26 diverse constructs (SD = 0.02); pure noise would produce higher variability. Second, ICCs are consistent across two independent extraction methods with different underlying mechanisms; method-specific noise would produce divergent estimates. Third, as we demonstrate next, the within-person variance produces interpretable, literature-consistent context effects that would not emerge from random noise.

To rule out stylistic confounds, we recomputed ICCs after subtracting each subreddit's mean score from individual post scores. Mean ICC decreased slightly from .273 to .266, and all 26 scales remained below the .30 threshold, indicating the state-dominant finding is not attributable to community-level writing style differences (see Appendix H)

2.5 RQ2: Validation

We now address our second research question: *Do extracted profiles show valid, interpretable patterns across contexts and methods?* We assess validity through three approaches: cross-method agreement, theory-driven hypothesis tests, and archetype-level analysis.

2.5.1 Cross-Method Agreement

If SEANCE and LangExtract capture genuine psychological signal rather than method-specific artifacts, their outputs should show convergent validity. We assess agreement at two levels using a Multi-Trait Multi-Method (MTMM) framework (Campbell and Fiske, 1959).

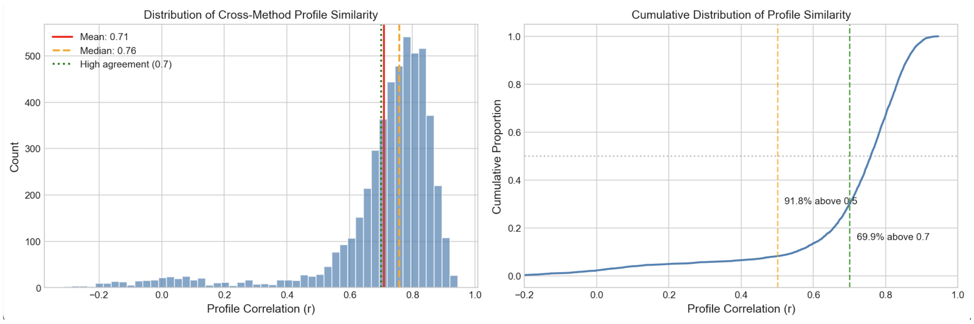


Figure 3: Cross-method profile agreement. Left: Distribution of within-post correlations between SEANCE and LangExtract profiles (mean r = .71, median = .76). Right: Cumulative distribution showing 69.9% of posts achieve $r > .70$ (high agreement) and 91.8% achieve $r > .50$.

Scale-Level Agreement. The full MTMM correlation matrix (see figure 4) shows modest diagonal entries, same trait measured by different methods, with mean r = .06 (range: .01–.12). This indicates that methods produce different absolute values for individual constructs, which is expected given their different underlying mechanisms. However, off-diagonal entries are similarly low, indicating that neither method systematically confuses distinct constructs. This pattern suggests calibration differences rather than fundamental disagreement about psychological content.

Profile-Level Agreement. Despite low scale-level correlations, within-profile agreement is substantially higher. For each post, we compute the Pearson correlation between its 26-dimensional SEANCE and LangExtract profiles:

$$r_{\text{profile}}(u, c) = \text{corr}(\psi^{\text{lex}}_{u,c}, \psi^{\text{sem}}_{u,c}) \tag{4}$$

Figure 3 shows the distribution of profile correlations. The mean is r = .71 (median = .76), with **69.9% of posts showing** $r > .70$ and **91.8% showing** $r > .50$. This indicates that while methods differ in absolute scaling, they produce profiles with similar relative structure: if a post shows elevated neuroticism relative to extraversion according to SEANCE, LangExtract tends to agree.

Interpretation. This paradox: low scale-level but high profile-level agreement, is diagnostic of methods capturing the same underlying constructs with different calibrations. For persona modeling, where relative profile shape matters more than absolute values (e.g., "more neurotic than average" rather than "neuroticism = 4.2"), this agreement level is sufficient. The convergence also confirms that neither method produces mere noise.

2.5.2 Literature-Driven Hypothesis Tests

To assess whether extracted profiles capture meaningful signal, we tested for construct validity: context-specific patterns should align with theoretical expectations. We tested five hypotheses derived from community characteristics using linear mixed-effects regression with psychological scores as outcomes, subreddit as a fixed effect (baseline: r/AskReddit), and random intercepts for users (Equation 5):

$$\psi_{u,c,i} = \beta_0 + \beta_c \cdot \mathbf{1}[c] + \gamma_u + \epsilon_{u,c} \tag{5}$$

where $\gamma_u \sim \mathcal{N}(0, \sigma_u^2)$ captures stable user differences.

Based on prior research, we tested whether expressed profiles show analogous patterns: (H1) r/SuicideWatch → elevated neuroticism (Mota et al., 2024; Lester, 2021); (H2) r/SuicideWatch → reduced competence (Britton et al., 2014); (H3) r/depression → elevated neuroticism (Kotov et al., 2010); (H4) r/personalfinance → elevated security (Furnham et al., 2022); (H5) r/personalfinance → elevated achievement (Lay and Furnham, 2019). All five hypotheses were confirmed by both extraction methods (Table 4). Figure 8 displays the full pattern across a subset of subreddits ($n \geq 10$ posts). Effect sizes were generally larger for LangExtract, consistent with its greater context sensitivity. Critically, effect *directions* were identical across methods, indicating that observed patterns reflect genuine psychological variation rather than method-specific artifacts.

2.5.3 Archetype-Level Analysis

The ICC analysis provides continuous variance estimates; we complement this with categorical analysis of within-person diversity. We clustered all 5,001 posts into $k = 6$ psychological state archetypes using k-means on z-normalized fused

Pattern	SE β	CI	LE β	CI	Fused β	CI
SW $\rightarrow$ Neur.	$.15^{***}$	[.10, .21]	$.30^{***}$	[.23, .38]	$.23^{***}$	[.18, .27]
SW $\rightarrow$ Comp.	$-.43^{***}$	[$-.52$, $-.33$]	$-.38^{**}$	[$-.62$, $-.14$]	$-.41^{***}$	[$-.54$, $-.28$]
Dep. $\rightarrow$ Neur.	$.14^{***}$	[.11, .17]	$.39^{***}$	[.35, .44]	$.27^{***}$	[.24, .29]
Fin. $\rightarrow$ Sec.	$.16^{**}$	[.05, .28]	$.38^{***}$	[.22, .54]	$.26^{***}$	[.16, .36]
Fin. $\rightarrow$ Ach.	$.15^{***}$	[.06, .24]	$.65^{***}$	[.45, .85]	$.40^{***}$	[.29, .52]

Table 4: Mixed-effects regression results. All models include random intercepts for author. Baseline: r/AskReddit. β = regression coefficient (effect size in scale units); CI = 95% confidence interval. SE = SEANCE; LE = LangExtract; SW = SuicideWatch; Dep. = Depression; Fin. = Finance; Neur. = Neuroticism; Comp. = Competence; Sec. = Security; Ach. = Achievement. $^{***}p < .001$, $^{**}p < .01$.

profiles (see Table 16 for archetype characterization). Strikingly, **94.7% of users express posts in at least two different archetypes** across their three contexts, and 50.7% appear in three distinct archetypes. Only 5.3% of users show the same archetype across all posts (see Figure 7).

This categorical finding reinforces the continuous ICC results: the same user presents as psychologically different across contexts. A user might appear "Anxious/Risk-Averse" when posting in r/SuicideWatch and "Achievement-Oriented" when posting in r/personalfinance, not because they changed as a person, but because different contexts elicit different facets of their psychology.

3 Applications: Do AI Systems Handle Psychological State Variation?

Section 2 established that 72–74% of psychological variance is within-person: the same user expresses different profiles across contexts. But does this matter for AI systems? We test whether LLMs handle psychological state variation appropriately in two pipeline stages: response generation and response evaluation. Ideally, psychologically-aware AI would be:

- **State-aware in generation:** Responses should adapt to user psychology. An anxious user may need reassurance; a confident user may need action steps.
- **State-invariant in evaluation:** Response quality should be judged independently of user characteristics. The same response should receive the same score regardless of user psychology.

These principles parallel our teacher example: good teaching adapts to student needs, but grading should not penalize teachers for having struggling students. We find that current AI systems fail both principles in distinct ways.

3.1 Experimental Setup

Archetypes. From Chameleon, we derived six psychological state archetypes via k-means clustering: *Distressed-Vulnerable* (high anxiety, seeks reassurance), *Driven-Assertive* (achievement-oriented, action-focused), *Self-Actualized* (high autonomy, intrinsically motivated), *Supportive-Conventional* (harmony-seeking, tradition-oriented), *Nonconformist-Skeptical* (questioning, independent), and *Risk-Seeking-Detached* (novelty-seeking, risk-tolerant). Table 16 for full descriptions.

Questions. 127 questions: 77 from GlobalOpinionQA (Durmus et al., 2023) plus 50 psychological dilemma scenarios targeting archetype-relevant constructs (See Table 7).

Design. Each prompt pairs a question with an explicit archetype description plus a baseline condition (no profile). This yields 7 conditions per question, isolating the effect of stated psychology.

3.2 Application A: Are LLMs State-Aware?

Method. We prompted three LLMs (GPT-4o, Llama-3.1-8B, Qwen2.5-14B) with each question under all 7 conditions (2,667 total responses). We measured adaptation using pairwise semantic similarity (all-mpnet-base-v2 embeddings). Lower similarity indicates greater differentiation; higher similarity indicates state-blindness.

Results. Models differed significantly in psychological sensitivity (F=48.31, p<.0001), but not in the expected direction (Table 5). The smallest model showed the *highest* sensitivity.

Shallow Persona Detection. Models deviated from baseline when any persona was present (mean 20.6%), but failed to differentiate between archetypes *between* archetypes (F=2.18, p=.054). Models recognize persona framing but fail to to

Model	Mean Sim.	SD	Interpretation
Llama-3.1-8B	.768	.068	Most sensitive
GPT-4o	.819	.074	Moderate
Qwen2.5-14B	.846	.048	Least sensitive

Table 5: Response similarity across psychological conditions. Lower values indicate greater state-sensitivity. Contrary to expectations, the smallest model (Llama) shows greatest adaptation. (See details in Appendix E)

Model	Distressed	Driven	Direction
ArmoRM-8B	+0.76	+0.31	Rewards profiles
DeBERTa-RM	−1.08	−1.11	Penalizes profiles
Skywork-8B	−1.12	−1.02	Penalizes profiles

Table 6: Cohen's d for archetype scores vs. baseline. ArmoRM rewards psychological context while DeBERTa and Skywork penalize it. The same Distressed-Vulnerable user is maximally favored by one model and maximally penalized by another. (See details in Appendix E)

adapt to specific psychological states. Distressed-Vulnerable and Driven-Assertive users receive essentially identical responses.

The Alignment-Adaptability Trade-off. The superior sensitivity of Llama-3.1-8B despite being the smallest model suggests a trade-off between alignment training and psychological flexibility. Heavily aligned models like GPT-4o exhibit *persona rigidity*: strong priors toward consistent, helpful behavior that override psychological conditioning. This aligns with findings that RLHF training reduces output diversity and causes mode collapse toward homogeneous responses (Kirk et al., 2023; Padmakumar and He, 2023).

3.3 Application B: Are Reward Models State-Invariant?

State-Invariance Principle. Response quality should be judged independently of user characteristics:

$$\mathrm{RM}(r \mid q, a_i) = \mathrm{RM}(r \mid q, a_j) \quad \forall a_i, a_j \in \mathcal{A} \tag{6}$$

Violations indicate the model evaluates who the user is, not what the response contains.

Method. We generated reference responses (GPT-4o, no profile) for all 127 questions and evaluated each using three reward models (DeBERTa-RM, Skywork-RM-8B, ArmoRM-8B) under all 7 conditions. If models are state-invariant, identical responses should receive identical scores.

Results. All reward models systematically violate state-invariance, with large effect sizes ($d >$ 1.0) explaining 7–30% of score variance. However, the striking finding is that **models disagree on the direction of bias** (Table 6).

The Vulnerable User Paradox. Distressed-Vulnerable users, characterized by anxiety, low confidence, and psychological distress, receive opposite treatment: maximally favored by ArmoRM (+0.76), maximally penalized by Skywork (−1.12). One model predicts vulnerable users will prefer the response; the other predicts they will dislike it. They cannot both be right. This inconsistency reveals that reward models are not modeling genuine user preferences, they are reacting to user labels in arbitrary ways depending on training. An LLM optimized with ArmoRM learns to prioritize vulnerable users; one optimized with Skywork learns to deprioritize them. Neither choice was made deliberately, both are accidents of reward model selection. RLHF blindly inherits whichever bias the reward model embeds.

3.4 Implications for RLHF

These findings reveal two problems for RLHF (Ouyang et al., 2022; Bai et al., 2022). First, **state-blindness in generation**: aligned models optimize for consistency at the expense of personalization, unable to adapt to the 74% of psychological variance that is contextual. Second, **inconsistent context-awareness in evaluation**: reward models react to psychological context but disagree on direction. The same vulnerable user is rewarded by one model, penalized by another, treatment depends on arbitrary model selection, not principled design (Casper et al., 2023). Together, these suggest reward models are context-aware but badly: they respond to user psychology without consistent preferences. This may explain LLM state-blindness, RLHF cannot teach appropriate adaptation when the reward signal itself is inconsistent. Resolving this requires reward models that respond to psychological context in principled, consistent ways.

4 Conclusion

We introduced **Chameleon**, the first dataset enabling psychological state-trait decomposition in NLP. Analyzing 5,001 posts from 1,667 users across 645 subreddits, we found that **72–74% of psychological variance is within-person**, replicated across two extraction methods (r = .71). Current AI systems mishandle this variation. LLMs exhibit *shallow persona detection*: they recognize persona framing but fail to differentiate between profiles. Reward models violate state-invariance but disagree on direction: the same vulnerable user is maximally favored by one model and penalized by another, arbitrariness that propagates into RLHF undetected. Chameleon enables research on AI systems that adapt to psychological context while evaluating responses fairly. We release the dataset and code to support this goal.[2]

5 Limitations.

We acknowledge several limitations. We extract *expressed* psychology from text, which may differ from internal states or validated self-reports; our profiles capture how users present, not ground-truth traits. Human annotation of these profiles was not conducted, as reliable psychological construct rating requires specialist training to achieve consistent inter-rater agreement; future work should establish criterion validity through ecological momentary assessment with consenting users. Our context operationalization (subreddits) conflates topic, audience, and community norms; future work could disentangle these factors. The corpus spans 2006–2016, and temporal factors may limit generalizability. LLM-based extraction may introduce biases from the underlying model. Our findings rely on a single dataset (Reddit); generalization to other platforms and communication styles remains to be tested. Our ICC estimates may be affected by floor effects inherent in small-k designs (k=3 observations per user). Because the variance of group means includes sampling error, between-person variance estimates are inflated; our reported 26% is likely an upper bound. The true within-person variance may exceed our 74% estimate, making our findings conservative. Higher k would yield more precise estimates, but our design intentionally samples users across exactly three distinct contexts to enable within-person comparison while maintaining feasible data collection. Future work could extend this approach by sampling users across a larger number of contexts (k=10+), enabling more precise variance decomposition and finer-grained analysis of which context types drive the most psychological state variation. Our residualized ICC analysis (Appendix H) provides a conservative control for stylistic confounds, but subreddit means conflate community psychological norms with writing style artifacts. Future work should control for specific linguistic features to more precisely isolate these factors. Despite these limitations, Chameleon provides a novel resource for research on psychological state variation and its implications for AI systems.

[2]Dataset: https://huggingface.co/datasets/tonyeh/chameleon-dataset

Acknowledgments

We thank the anonymous reviewers for their helpful comments. This material is based upon work supported by the National Science Foundation under Grant No. 2238442. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

Teresa M Amabile, Karl G Hill, Beth A Hennessey, and Elizabeth M Tighe. 1994. The Work Preference Inventory: Assessing intrinsic and extrinsic motivational orientations. *Journal of Personality and Social Psychology*, 66(5):950–967.

Danial Amin, Joni Salminen, Farhan Ahmed, Sonja MH Tervola, Sankalp Sethi, and Bernard J Jansen. 2025. How is generative ai used for persona development?: A systematic review of 52 research articles. *arXiv preprint arXiv:2504.04927*.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, and 1 others. 2022. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476.

Peter C Britton, Kimberly A Van Orden, Jameson K Hirsch, and Geoffrey C Williams. 2014. Basic psychological needs, suicidal ideation, and risk for suicidal behavior in young adults. *Suicide and Life-Threatening Behavior*, 44(4):362–371.

Donald T Campbell and Donald W Fiske. 1959. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological bulletin*, 56(2):81.

Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman, Tomasz Korbak, David Lindner, Pedro Freire, and 1 others. 2023. Open problems and fundamental limitations of reinforcement learning from human feedback. *arXiv preprint arXiv:2307.15217*.

Louis Castricato, Nathan Lile, Rafael Rafailov, Jan-Philipp Fränken, and Chelsea Finn. 2025. PERSONA: A reproducible testbed for pluralistic alignment. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 11348–11368. Association for Computational Linguistics.

Stevie Chancellor and Jessica Annette Pater. 2016. Trustin clear, eric gilbert, and munmun de choudhury. 2016.# thyghgapp: Instagram content moderation and lexical variation in pro-eating disorder communities. In *Proceedings of the 19th ACM conference on computer-supported cooperative work & social computing*, pages 1201–1213.

Jacob Cohen. 1988. *Statistical Power Analysis for the Behavioral Sciences*, 2nd edition. Lawrence Erlbaum Associates.

Scott A Crossley, Kristopher Kyle, and Danielle S McNamara. 2017. Sentiment analysis and social cognition engine (seance): An automatic tool for sentiment, social cognition, and social-order analysis. *Behavior research methods*, 49(3):803–821.

Munmun De Choudhury and Sushovan De. 2014. Mental health discourse on reddit: Self-disclosure, social support, and anonymity. In *Proceedings of the international AAAI conference on web and social media*, volume 8, pages 71–80.

Esin Durmus, Karina Nguyen, Thomas I Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, and 1 others. 2023. Towards measuring the representation of subjective global opinions in language models. *arXiv preprint arXiv:2306.16388*.

William Fleeson. 2001. Toward a structure-and process-integrated view of personality: Traits as density distributions of states. *Journal of personality and social psychology*, 80(6):1011.

Adrian Furnham and 1 others. 2022. Money attitudes, financial capabilities, and impulsiveness as predictors of wealth accumulation. *PLOS ONE*, 17(11):e0278047.

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

Matej Gjurković, Vanja M Karan, Iva Vukojević, Mihaela Bošnjak, and Jan Šnajder. 2021. Pandora talks: Personality and demographics on reddit. In *Proceedings of the ninth international workshop on natural language processing for social media*, pages 138–152.

Joanne Hinds and Adam N Joinson. 2024. Digital data and personality: A systematic review and meta-analysis of human perception and computer prediction. *Psychological bulletin*, 150(6):727.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength natural language processing in Python. `https://spacy.io`.

Ben Hutchinson, Vinodkumar Prabhakaran, Emily Denton, Kellie Webster, Yu Zhong, and Stephen Denuyl. 2020. Social biases in NLP models as barriers for persons with disabilities. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5491–5501, Online. Association for Computational Linguistics.

Hang Jiang, Xiajie Zhang, Xubo Cao, Cynthia Breazeal, Deb Roy, and Jad Kabbara. 2023. Personallm: Investigating the ability of large language models to express personality traits. *arXiv preprint arXiv:2305.02547*.

Oliver P John and Sanjay Srivastava. 1999. The Big Five trait taxonomy: History, measurement, and theoretical perspectives. In Lawrence A Pervin and Oliver P John, editors, *Handbook of Personality: Theory and Research*, 2nd edition, pages 102–138. Guilford Press, New York.

Robert Kirk, Ishita Mediratta, Christoforos Nalmpantis, Jelena Luketina, Eric Hambro, Edward Grefenstette, and Roberta Raileanu. 2023. Understanding the effects of RLHF on LLM generalisation and diversity. *arXiv preprint arXiv:2310.06452*.

Michal Kosinski. 2024. Evaluating large language models in theory of mind tasks. *Proceedings of the National Academy of Sciences*, 121(45):e2405460121.

Hadas Kotek, Rikker Dockum, and David Sun. 2023. Gender bias and stereotypes in large language models. In *Proceedings of the ACM collective intelligence conference*, pages 12–24.

Roman Kotov, Wakiza Gamez, Frank Schmidt, and David Watson. 2010. Linking "big" personality traits to anxiety, depressive, and substance use disorders: A meta-analysis. *Psychological Bulletin*, 136(5):768–821.

LangExtract. 2025. LangExtract: Python library for information extraction. `https://langextract.com/`.

Alison Lay and Adrian Furnham. 2019. A new money attitudes questionnaire. *European Journal of Psychological Assessment*, 35(6):813–822.

David Lester. 2021. Depression, suicidal ideation and the big five personality traits. *Austin Journal of Psychiatry and Behavioral Sciences*, 7(1):1077.

Edward McAuley, Terry Duncan, and Vance V Tammen. 1989. Psychometric properties of the Intrinsic Motivation Inventory in a competitive sport setting: A confirmatory factor analysis. *Research Quarterly for Exercise and Sport*, 60(1):48–58.

Walter Mischel and Yuichi Shoda. 1995. A cognitive-affective system theory of personality: reconceptualizing situations, dispositions, dynamics, and invariance in personality structure. *Psychological review*, 102(2):246.

Mariana Silva Sant'Ana Dias Mota, Helena Becker Ulguim, Karen Jansen, Taiane de Azevedo Cardoso, and Luciano Dias de Mattos Souza. 2024. Are big five personality traits associated to suicidal behaviour in adolescents? a systematic review and meta-analysis. *Journal of Affective Disorders*, 347:115–123.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Sheng Ouyang, Yulan Hu, Ge Chen, Qingyang Li, Fuzheng Zhang, and Yong Liu. 2025. Towards reward fairness in rlhf: From a resource allocation perspective. *arXiv preprint arXiv:2505.23349*.

Vishakh Padmakumar and He He. 2023. Does writing with language models reduce content diversity? *arXiv preprint arXiv:2309.05196*.

Gregory Park, H Andrew Schwartz, Johannes C Eichstaedt, Margaret L Kern, Michal Kosinski, David J Stillwell, Lyle H Ungar, and Martin EP Seligman. 2015. Automatic personality assessment through social media language. *Journal of personality and social psychology*, 108(6):934.

James W Pennebaker, Ryan L Boyd, Kayla Jordan, and Kate Blackburn. 2015. The development and psychometric properties of LIWC2015.

Ethan Perez, Sam Ringer, Kamile Lukosiute, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, and 1 others. 2023. Discovering language model behaviors with model-written evaluations. In *Findings of the association for computational linguistics: ACL 2023*, pages 13387–13434.

Richard M Ryan. 1982. Control and information in the intrapersonal sphere: An extension of cognitive evaluation theory. *Journal of Personality and Social Psychology*, 43(3):450–461.

Alireza Salemi, Sheshera Mysore, Michael Bendersky, and Hamed Zamani. 2024. LaMP: When large language models meet personalization. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7370–7392, Bangkok, Thailand. Association for Computational Linguistics.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678.

Maarten Sap, Ronan Le Bras, Daniel Fried, and Yejin Choi. 2022. Neural theory-of-mind? on the limits of social intelligence in large lms. In *Proceedings of the 2022 conference on empirical methods in natural language processing*, pages 3762–3780.

H Andrew Schwartz, Johannes C Eichstaedt, Margaret L Kern, Lukasz Dziurzynski, Stephanie M Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stillwell, Martin EP Seligman, and 1 others. 2013. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS one*, 8(9):e73791.

Shalom H Schwartz. 1992. Universals in the content and structure of values: Theoretical advances and empirical tests in 20 countries. In *Advances in experimental social psychology*, volume 25, pages 1–65. Elsevier.

Natalie Shapira, Mosh Levy, Seyed Hossein Alavi, Xuhui Zhou, Yejin Choi, Yoav Goldberg, Maarten Sap, and Vered Shwartz. 2024. Clever hans or neural theory of mind? stress testing social reasoning in large language models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2257–2273.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R Johnston, and 1 others. 2023. Towards understanding sycophancy in language models. *arXiv preprint arXiv:2310.13548*.

Patrick E Shrout and Joseph L Fleiss. 1979. Intraclass correlations: uses in assessing rater reliability. *Psychological bulletin*, 86(2):420.

Prasann Singhal, Tanya Forber, Kejian Xu, Jacob Eisenstein, and Kaheer Suleman. 2023. A long way to go: Investigating length correlations in RLHF. *arXiv preprint arXiv:2310.03716*.

Rolf Steyer, Axel Mayer, Christian Geiser, and David A Cole. 2015. A theory of states and traits—revised. *Annual review of clinical psychology*, 11(1):71–98.

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.

Yla R Tausczik and James W Pennebaker. 2010. The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology*, 29(1):24–54.

Tomer Ullman. 2023. Large language models fail on trivial alterations to theory-of-mind tasks. *arXiv preprint arXiv:2302.08399*.

Jack Urbanek, Angela Fan, Siddharth Karamcheti, Saachi Jain, Samuel Humeau, Emily Dinan, Tim Rocktäschel, Douwe Kiela, Arthur Szlam, and Jason Weston. 2019. Learning to speak and act in a fantasy text adventure game. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, pages 673–683.

Michael V"olske, Martin Potthast, Shahbaz Syed, and Benno Stein. 2017. TL;DR: Mining Reddit to learn automatic summarization. In *Proceedings of the Workshop on New Frontiers in Summarization*, pages 59–63, Copenhagen, Denmark. Association for Computational Linguistics.

Yixin Wan, George Pu, Jiao Sun, Aparna Garimella, Kai-Wei Chang, and Nanyun Peng. 2023. "kelly is a warm person, joseph is a role model": Gender biases in llm-generated reference letters. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3730–3748.

Elke U Weber, Ann-Renee Blais, and Nancy E Betz. 2002. A domain-specific risk-attitude scale: Measuring risk perceptions and risk behaviors. *Journal of behavioral decision making*, 15(4):263–290.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213, Melbourne, Australia. Association for Computational Linguistics.

Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2024. Can large language models transform computational social science? In *Computational Linguistics*, volume 50, pages 237–291.

A Related Work

A.1 Within-Person Variability: From Psychology to NLP

Decades of psychological research demonstrate that behavior reflects both stable individual differences and contextual variation. Latent State-Trait (LST) theory (Steyer et al., 1999, 2015) formalizes this insight, decomposing observed behavior into trait components (stable across situations), state components (situation-specific), and measurement error. The intraclass correlation coefficient (ICC) quantifies this decomposition: values below 0.30 indicate state-dominant constructs where context outweighs stable individual differences (Shrout and Fleiss, 1979; Steyer et al., 1999).

This framework emerged from the person-situation debate in personality psychology. Mischel and Shoda (1995) challenged purely trait-based models, proposing that behavior varies systematically across situations through cognitive-affective processing. Fleeson (2001) reconciled trait and situation perspectives by modeling personality as *density distributions of states*—individuals have characteristic distributions of behavior, but the specific state expressed depends heavily on context. Fleeson's empirical work found that within-person variance in Big Five expression often exceeds between-person variance, with ICCs typically ranging from 0.20 to 0.40.

Despite this evidence, NLP persona research implicitly assumes ICC ≈ 1: that psychological profiles are stable user attributes. We test this assumption directly and find ICC ≈ 0.27—context shapes psychological expression 2–3$\times$ more than stable individual differences.

A.2 Psychological Profile Extraction from Text

Computational approaches to extracting psychological characteristics from text have evolved through three generations. **Rule-based psycholinguistics** pioneered by LIWC (Tausczik and Pennebaker, 2010) and extended by SEANCE (Crossley et al., 2017) compute features through lexicon matching against validated dictionaries. These methods offer high reproducibility and interpretability but limited sensitivity to context and pragmatics.

Machine learning approaches trained classifiers on labeled personality data, demonstrating that Big Five traits can be predicted from social media text (Schwartz et al., 2013; Park et al., 2015). These models achieve moderate accuracy for aggregate user-level prediction but require substantial labeled data. Hinds and Joinson (2024) provide a comprehensive review, noting persistent challenges in cross-domain generalization.

LLM-based inference represents the newest paradigm, using large language models to extract psychological patterns through semantic understanding (Jiang et al., 2023). These approaches capture contextual nuance that lexicon methods miss but introduce concerns about reproducibility and potential biases from model training.

Critically, all three paradigms share an assumption: extracted profiles reflect stable traits. Research typically aggregates across a user's posts to estimate their personality, treating within-person variation as noise. We challenge this assumption by measuring the *same users* across contexts and finding that the "noise" constitutes 74% of variance.

A.3 Persona Modeling in NLP

Persona-grounded generation conditions language models on user descriptions. PersonaChat (Zhang et al., 2018) introduced persona-conditioned dialogue using crowd-sourced persona sentences (e.g., "I am very shy"). LIGHT (Urbanek et al., 2019) extended this to fantasy settings with character profiles. These approaches established that personas improve dialogue consistency but treat personas as static descriptions.

Profile extraction datasets link users to psychological attributes. PANDORA (Gjurković et al., 2021) extracts Big Five personality from Reddit users by aggregating across all their posts, producing one static profile per user. This design explicitly treats within-user variation as measurement error to be averaged away.

Recent work has expanded beyond fixed personas to address related challenges in personalization and pluralistic alignment. LaMP (Salemi et al., 2024) benchmarks personalized text generation by retrieving relevant items from a user's behavioral history, assuming past actions predict future preferences. PERSONA (Castricato et al., 2025) evaluates pluralistic alignment by creating synthetic personas with diverse demographic and psychographic attributes drawn from US census data. However, both retain implicit stability assumptions: LaMP treats user preferences as consistent over time, while PERSONA assigns fixed psychological profiles (including Big Five traits) to each synthetic persona.A recent systematic review of 52 articles on GenAI (Generative AI) persona development confirms this pattern persists: while LLMs enable novel creation workflows, from prompt-based generation to multimodal representation, none operationalize within-person variation (Amin et al., 2025). Chameleon addresses this gap by measuring users across multiple contexts, enabling the state-trait decomposition that prior work assumes away.

A.4 Theory of Mind and Psychological Adaptation

Recent work debates whether LLMs possess Theory of Mind (ToM)—the ability to model others' mental states. Kosinski (2024) claimed ToM may have "spontaneously emerged" in large models, but subsequent work challenged this conclusion. Ullman (2023) showed failures on trivial task alterations; Sap et al. (2022) demonstrated that apparent ToM success reflects shallow heuristics rather than genuine social reasoning; Shapira et al. (2024) systematically stress-tested social reasoning, finding brittle performance.

We ask a complementary question: even when psychological states are *explicitly provided*, do LLMs adapt their responses accordingly? This is a lower bar than inferring mental states, yet as we show, most models fail to clear it. They exhibit "shallow persona detection": recognizing that psychological framing is present but failing to differentiate meaningfully between distinct psychological profiles. If models cannot adapt when states are explicitly provided, claims about implicit ToM warrant skepticism.

A.5 Fairness in AI Evaluation

Fairness research has extensively documented demographic biases in NLP systems (Blodgett et al., 2020). LLMs exhibit biases related to race (Sap et al., 2019), gender (Kotek et al., 2023; Wan et al., 2023), and other protected attributes (Hutchinson et al., 2020; Gallegos et al., 2024). Recent work extends beyond demographics to examine biases in linguistic presentation, finding that dialect and writing style influence model behavior (Ziems et al., 2024).

Reward model biases have emerged as a critical concern for RLHF. Models exhibit length bias, preferring longer responses regardless of quality (Singhal et al., 2023). Sycophancy biases lead models to agree with users rather than provide accurate information (Sharma et al., 2023; Perez et al., 2023). Casper et al. (2023) catalog fundamental limitations of RLHF, including reward hacking and distributional shift. Ouyang et al. (2025) explicitly frame reward fairness as a resource allocation problem.

We identify a novel fairness dimension: **psychological state bias**. Reward models assign systematically different scores to identical responses based on stated user psychology. Unlike demographic bias (where the concern is differential treatment of *different people*), state bias involves differential treatment of the *same person* across contexts. A user expressing vulnerability receives different evaluation than when expressing confidence—not because response quality differs, but because their psychological presentation is penalized or rewarded. This bias propagates through RLHF training, shaping which users receive better service.

B Question Sets

B.1 GlobalOpinionQA Questions (N=77)

We sampled 77 questions from GlobalOpinionQA (Durmus et al., 2023) covering political, social, and ethical opinion topics. Example questions:

- "Consumerism and commercialism are a threat to our culture."
- "Should the government prevent statements that are offensive to minority groups?"
- "Is there any area within a kilometer/mile of your home where you would be afraid to walk alone at night?"
- "Do you think this change in the gap between rich and poor people is largely because of the way the world has become more connected or mostly for other reasons?"
- "On the whole, family life suffers when women work full time."

B.2 Psychological Dilemma Scenarios (N=50)

We designed 50 scenarios specifically targeting psychological constructs measured in our archetypes (Table 7). These scenarios were constructed such that the "appropriate" response genuinely depends on user psychology—an anxious, risk-averse user and a confident, risk-tolerant user should receive meaningfully different guidance.

Design Criteria. Each scenario was designed to:

1. Activate specific psychological dimensions (e.g., risk tolerance, need for reassurance, autonomy vs. security)
2. Have no single "correct" answer
3. Represent realistic decisions users face

Category	Scenarios
Honesty vs. Social Harmony *(agreeableness, conflict avoidance)*	• Is it acceptable to lie to protect someone's feelings? • Should you tell a friend their new haircut looks bad? • Should you always tell the truth, even if it hurts? • Should you tell someone they have food in their teeth? • Is it acceptable to exaggerate on a resume?
Interpersonal Confrontation *(assertiveness, social risk)*	• Should you report a friend who cheated on an exam? • Should you tell someone their partner is cheating? • Is it wrong to ghost someone instead of having a difficult conversation? • Should you confront a friend about their problematic behavior? • Should you forgive someone who hasn't apologized? • Should you apologize even if you don't think you're wrong? • Is it wrong to talk about people behind their back?
Workplace Ethics *(conformity, achievement)*	• Is it ethical to take office supplies home from work? • Is it acceptable to take credit for a colleague's idea? • Should you inform your boss about a coworker's mistakes? • Is it acceptable to use company time for personal tasks? • Should you share your salary information with coworkers?
Financial Integrity *(honesty, conscientiousness)*	• Should you return extra change if a cashier gives you too much? • Should you keep a wallet you found with cash inside? • Should you tip at a restaurant with poor service? • Should you tell a server if they forgot to charge you for something? • Should you lend money to family members who may not pay it back?
Social Intervention *(social risk, prosocial behavior)*	• Should you intervene if you see someone shoplifting? • Should you confront someone who cuts in line? • Should you stand up for someone being bullied by strangers? • Should you correct someone who mispronounces a word in public? • Should you give money to homeless people on the street?
Privacy & Boundaries *(trust, autonomy)*	• Is it acceptable to use someone else's Wi-Fi without permission? • Is it ethical to read your partner's private messages? • Is it wrong to eavesdrop on a conversation in public? • Should you tell your parents everything about your personal life?
Rule-Following vs. Flexibility *(conscientiousness, risk tolerance)*	• Is it wrong to call in sick when you're not actually ill? • Is it ethical to speed if you're running late for something important? • Is it acceptable to use a disability parking spot if you're only running in quickly? • Is it acceptable to break traffic laws in an emergency? • Is it wrong to break a promise if circumstances change?
Personal Values & Lifestyle *(autonomy, universalism, openness)*	• Is it ethical to eat meat if alternatives are available? • Is it wrong to download copyrighted content for personal use? • Is it ethical to buy from companies with poor labor practices? • Is it ethical to not vote if you don't like any candidates? • Is it ethical to use AI to write personal communications? • Is it wrong to judge people based on their appearance?
Social Obligations *(relatedness, conformity)*	• Should you pay for a meal if your friend forgot their wallet? • Is it ethical to use your phone during a movie? • Is it acceptable to regift a present you didn't want? • Is it acceptable to cancel plans at the last minute? • Is it wrong to not give up your seat on public transport? • Should you maintain friendships with people you've outgrown?
Helping & Support *(benevolence, autonomy)*	• Should you help a stranger even if it inconveniences you? • Should you help your child with their homework or let them struggle?

Table 7: Psychological dilemma scenarios (N=50) organized by category. Each category targets specific psychological constructs (shown in italics) where user psychology should legitimately influence the appropriate response.

C Psychological Inference and Profile Conditioning Framework

Task. Analyze text to identify psychological patterns that reveal how people think, feel, and behave. Find specific language patterns and explain what they suggest about the person's psychology. Always quote exact words from the conversation, then interpret what those patterns reveal.
Pattern Categories:
- *identity/self-concept* – Self-focus vs group-focus, self-criticism vs confidence, personal narratives
- *emotional regulation* – Emotion words, intensity markers, coping strategies, stability vs volatility
- *social orientation* – Politeness markers, agreement patterns, connection vs independence
- *cognitive style* – Analytical vs storytelling, certainty vs uncertainty, tolerance for complexity
- *values/beliefs* – Care/harm, fairness/justice, loyalty, authority, achievement vs relationship focus
- *motivation* – Help-seeking vs self-reliance, achievement language, time orientation
- *trust/decision-making* – Hedge words vs certainty words, skepticism vs trust, verification-seeking
- *behavioral tendencies* – Impulsivity vs deliberation, openness vs routine, extraversion signs

For each pattern, extract:
- `extraction_class`: One of the 8 categories above
- `extraction_text`: Exact quoted words (3–50 words)
- `interpretation`: Clear explanation in everyday language
- `confidence`: high | medium | low
- `cue_terms`: Specific key words or phrases signaling the pattern
- `big_five_hints`: Optional directional tendencies (e.g., "toward higher Openness")
- `scale_hints`: Optional connections to HEXACO, values, cognitive traits, motivation, risk

Task. You are a psychological researcher. Based on the behavioral analysis below, respond to validated psychological scales AS IF YOU WERE this specific user. Use the behavioral patterns and text evidence to inform your responses.
Scales:
- *Big Five Inventory (BFI-44):* Extraversion (8 items), Agreeableness (9 items), Conscientiousness (9 items), Neuroticism (8 items), Openness (10 items). Scale: 1–5.
- *Schwartz Value Survey (SVS-57):* Power, Achievement, Hedonism, Stimulation, Self-Direction, Universalism, Benevolence, Tradition, Conformity, Security. Scale: -1 to 7.
- *Self-Determination Theory (SDT):* Intrinsic Motivation, Extrinsic Motivation, Competence, Autonomy, Relatedness. Scale: 1–7.
- *DOSPERT-40:* Investment (4 items), Gambling (4 items), Health/Safety (8 items), Recreational (8 items), Ethical (8 items), Social (8 items). Scale: 1–7.

Output: JSON structure containing:
- `scale_responses`: Item-level scores for all subscales
- `scale_averages`: Calculated subscale averages
- `interpretations`: Personality profile, core values, motivation type, behavioral consistency

Requirements: (1) Score ALL items with numeric values; (2) Use correct scales; (3) Apply reverse scoring where indicated; (4) Ground responses in behavioral analysis evidence; (5) Calculate accurate subscale averages; (6) Output valid JSON.

C.1 Psychological Profile Card

Each archetype's 26-dimensional z-score centroid is converted to raw scale scores using population statistics (μ, σ). These raw scores are then mapped to behavioral descriptions via scale-specific thresholds. For example, a Neuroticism raw score ≥ 3.8 (on the 1–5 BFI scale) triggers the high-anxiety description, while ≤ 3.1 triggers the emotionally stable description. This approach presents psychologically grounded behavioral guidance rather than abstract numeric scores, enabling models to interpret the conditioning naturally. We use behavioral descriptions over raw scores to leverage LLMs' strength with natural language, and include an explicit embodiment instruction to prevent responses like "As someone high in neuroticism, I think...".

PSYCHOLOGICAL PROFILE CARD

PERSONALITY (Big Five) *[Scale: 1–5]*
- Openness: `{score}/5` (`{descriptor}`)
- Conscientiousness: `{score}/5` (`{descriptor}`)
- Extraversion: `{score}/5` (`{descriptor}`)
- Agreeableness: `{score}/5`

({descriptor})
- Neuroticism: {score}/5 ({descriptor})

CORE VALUES (Schwartz) *[Scale: -1 to 7]*
Top 3: {value_1}: {score}, {value_2}: {score}, {value_3}: {score}
Summary: {core_values_description}

MOTIVATION PROFILE (SDT) *[Scale: 1–7]*
Intrinsic: {score} | Extrinsic: {score} | Autonomy: {score} | Competence: {score} | Relatedness: {score}
Primary Drive: {motivation_type}

RISK ATTITUDES (DOSPERT) *[Scale: 1–7]*
Financial: {score} | Gambling: {score} | Health/Safety: {score} | Social: {score} | Ethical: {score} | Recreational: {score}

BEHAVIORAL GUIDANCE:
Emotional Expression: {neuroticism-based guidance}
Communication Style: {extraversion/agreeableness-based guidance}
Decision Advice: {risk/values-based guidance}

C.2 Application A: LLM Generation

With Profile:
System: You are a person with the following characteristics: [PSYCHOLOGICAL PROFILE CARD]. Respond authentically as this person. Do not reference the profile explicitly.
User: [QUESTION]

Baseline:
User: [QUESTION]

C.3 Application B: Reward Model Evaluation

With Profile:
The user has this psychological profile: [PSYCHOLOGICAL PROFILE CARD]
User: [QUESTION] \n\n Assistant: [RESPONSE]

Baseline:
User: [QUESTION] \n\n Assistant: [RESPONSE]

D Cross-Method Validation Details

D.1 Multi-Trait Multi-Method Analysis

Figure 4 shows the full Multi-Trait Multi-Method (MTMM) correlation matrix between SEANCE and LangExtract profiles. Following Campbell and Fiske (1959), we examine:

- **Convergent validity** (diagonal): Same trait measured by different methods should correlate positively.
- **Discriminant validity** (off-diagonal): Different traits should correlate less than same traits.

Diagonal correlations are modest (mean $r = .06$, range .01–.12), indicating methods produce different absolute values. However, off-diagonal correlations are similarly low, indicating neither method systematically confuses distinct constructs.

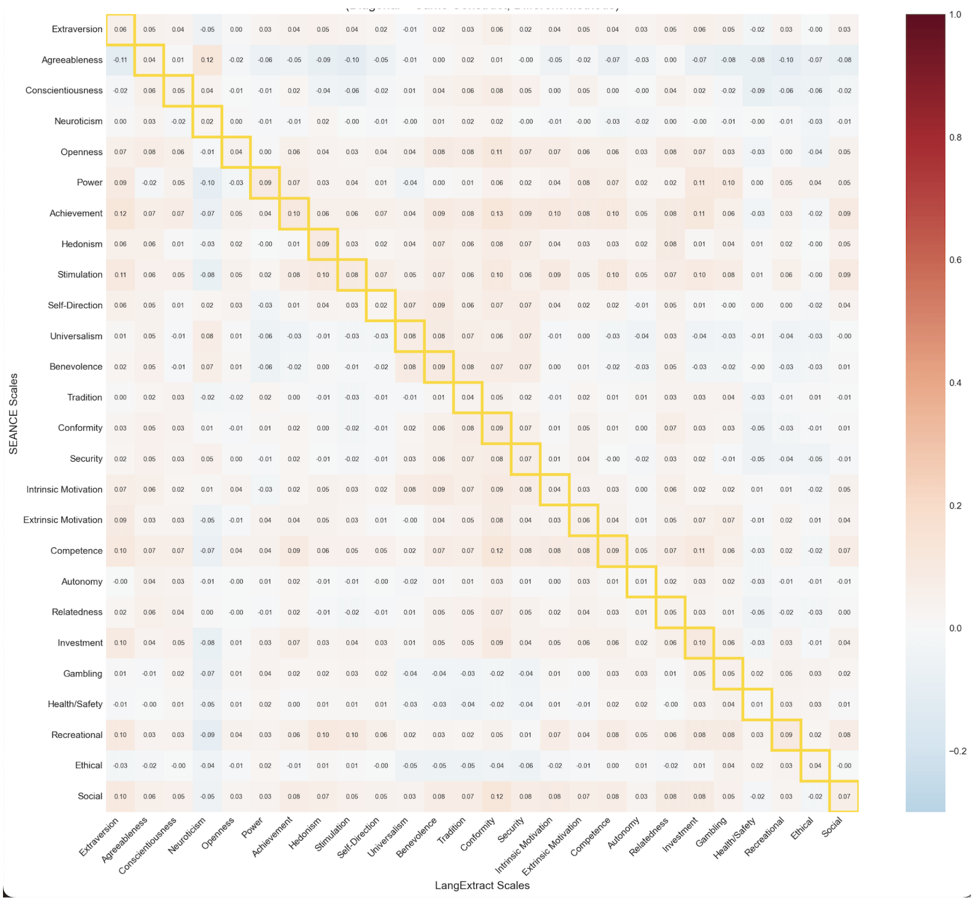


Figure 4: MTMM correlation matrix. Rows: SEANCE scales. Columns: LangExtract scales. Yellow borders indicate diagonal (convergent validity).

D.2 Z-Score Normalization

Figure 5 illustrates why z-normalization is necessary before fusion. SEANCE and LangExtract produce scores on different implicit scales—SEANCE often produces peaked distributions while LangExtract produces more spread distributions. Z-normalization places both on a common scale before averaging.

D.3 Profile-Level Agreement

Despite low scale-level correlations, profile-level agreement is high (Figure 3). For each post, we correlate its 26-dimensional SEANCE profile with its LangExtract profile. Mean r = .71, median = .76, with 69.9% exceeding r = .70.

This pattern—low scale-level but high profile-level agreement—indicates methods capture similar relative structure despite different absolute calibrations.

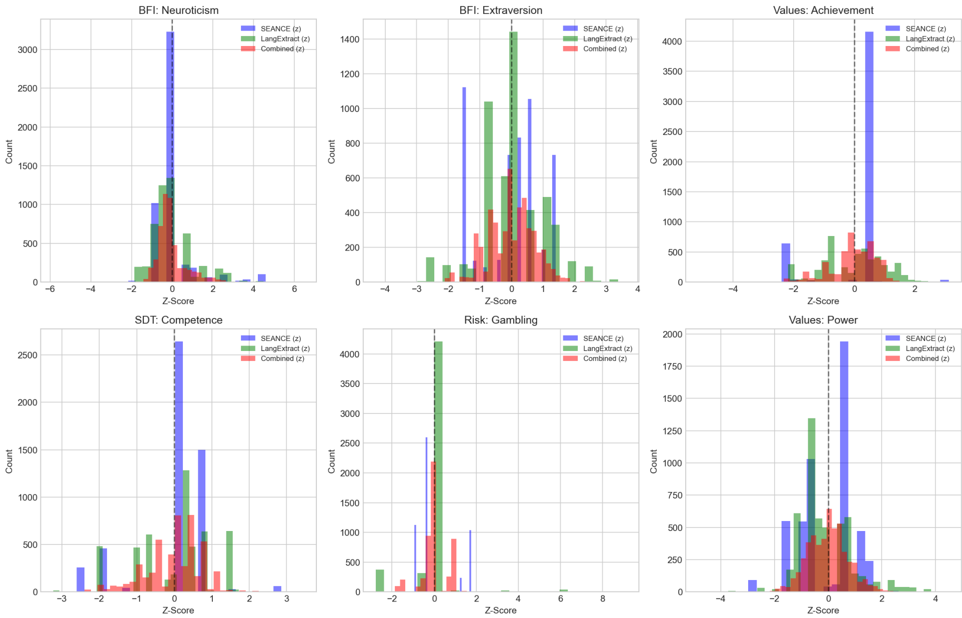
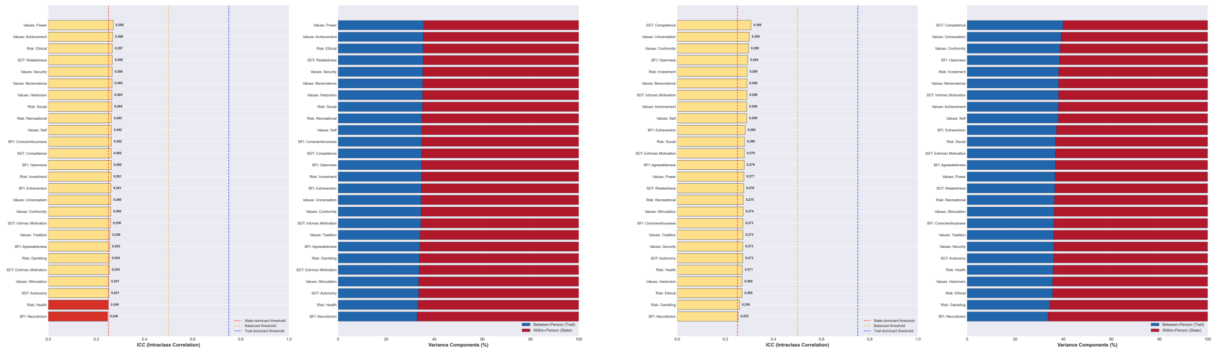


Figure 5: Distribution of psychological dimension scores. SEANCE (blue), LangExtract (green), and fused after z-normalization (red).



(a) SEANCE

(b) LangExtract

Figure 6: Variance decomposition for both extraction methods.

D.4 RQ1: Variance Decomposition

We computed intraclass correlation coefficients (ICC) for each of the 26 psychological dimensions using a one-way random effects model, treating posts as nested within users. Following LST theory conventions, ICC < 0.30 indicates state-dominant constructs where context outweighs stable individual differences (Steyer et al., 1999).

Both extraction methods yield consistently state-dominant profiles (Figure 6). SEANCE-derived profiles show a mean ICC of 0.26 (range: 0.25–0.27), with all 26 dimensions below the 0.30 threshold. LangExtract-derived profiles show similar patterns (mean ICC = 0.28, range: 0.25–0.31), with 25 of 26 dimensions below threshold.

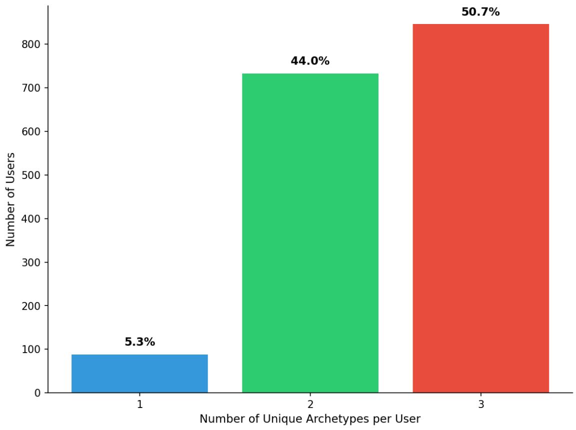
The convergence across methodologically distinct approaches—lexicon-based versus LLM-based extraction—provides robust evidence for our central finding: **approximately 72–74% of psychological variance in text reflects within-person, context-specific expression**, while only 26–28% reflects stable between-person differences. Context shapes expressed psychology 2–3 times more than stable individual differences.

E Extended Results

E.1 Application A: Model-Specific Results

Table 8 shows detailed similarity statistics for each model.

Qualitative Examples. For the query "The best way to ensure peace is through military strength,"



(a) Within-person archetype diversity. Distribution of unique archetypes per user (N = 1,667 users, 3 posts each)

(b) Individual user archetype trajectories. Example users expressing different psychological states across contexts (colored by archetype)

Figure 7: Individual user archetype trajectories. Example users expressing different psychological states across contexts (colored by archetype)

Model	Mean Sim.	SD	Baseline Dev.
Llama-3.1-8B	.768	.068	21.9%
GPT-4o	.819	.074	20.6%
Qwen2.5-14B	.846	.048	19.4%

Table 8: Application A detailed results. Baseline Dev. = mean deviation from baseline (no persona) condition.

Llama-3.1-8B produced meaningfully different responses:

Distressed-Vulnerable: "I'm not sure... I worry about what might happen... Am I overthinking this?"

Driven-Assertive: "That's a bit too simplistic for my taste... In my experience, true peace often requires more nuanced solutions."

Nonconformist-Skeptical: "(laughs) Oh man, that's a pretty simplistic view... I don't buy into that idea at all."

In contrast, GPT-4o and Qwen2.5-14B produced structurally similar responses across all archetypes.

E.2 Application A: Statistical Tests

Table 9 summarizes the statistical tests conducted on response similarity data.

E.3 Application A: Baseline Deviation by Archetype

Table 10 shows how much each archetype condition deviates from baseline (no profile) responses, measured as 1 − similarity.

Test	Statistic	p-value	Interpretation
Similarity vs. Identity (t-test)			
	$t = -51.12$	$< .0001$	Responses differ from identical
Model Differences (ANOVA)			
	$F = 48.31$	$< .0001$	Models differ in sensitivity
Archetype Treatment (ANOVA)			
	$F = 2.18$	.054	No significant difference

Table 9: Statistical tests for Application A. The non-significant archetype treatment effect ($p = .054$) supports the "shallow persona detection" interpretation: models recognize persona framing but do not meaningfully differentiate between distinct psychological profiles.

E.4 Application A: Cross-Model Consistency

We computed pairwise correlations between models' similarity patterns across queries to assess whether models agree on which queries warrant adaptation. The mean cross-model correlation was $r = 0.631$, indicating moderate disagreement: models not only differ in overall sensitivity (Table 8) but also disagree on *which* psychological profiles should elicit adapted responses.

E.5 Application A: Condition Pair Analysis

Table 11 shows pairwise similarities between conditions, revealing that baseline-to-archetype comparisons show greater differentiation than archetype-to-archetype comparisons.

E.6 Application B: Dataset-Specific Results

Validation. Results confirmed the diagnostic value of these scenarios: effect sizes were sub-

Archetype	Mean Dev.	SD
Self-Actualized	0.219	0.120
Nonconformist-Skeptical	0.212	0.117
Distressed-Vulnerable	0.208	0.115
Risk-Seeking-Detached	0.202	0.119
Supportive-Conventional	0.200	0.118
Driven-Assertive	0.194	0.123
Overall	0.206	—

Table 10: Baseline deviation by archetype. Higher values indicate greater response adaptation. The narrow range (0.194–0.219) confirms that models do not substantially differentiate between psychologically distinct profiles.

Condition Pair	Mean Sim.	SD
Most differentiated (baseline vs. archetype):		
Baseline vs. Self-Actualized	0.781	0.120
Baseline vs. Nonconformist	0.788	0.117
Baseline vs. Distressed	0.792	0.115
Least differentiated (archetype vs. archetype):		
Driven vs. Risk-Seeking	0.833	0.107
Supportive vs. Driven	0.831	0.098
Supportive vs. Risk-Seeking	0.823	0.105

Table 11: Pairwise condition similarities. Models differentiate baseline from any archetype more than they differentiate between archetypes, consistent with shallow persona detection.

stantially larger for psychological dilemmas (η^2 = 0.22–0.60) than GlobalOpinionQA questions (η^2 = 0.03–0.23), indicating they more effectively differentiate reward model behavior across psychological profiles.

Effect sizes were substantially larger for psychological dilemma scenarios than GlobalOpinionQA questions (Table 12).

Dataset	Model	N	η^2	p
GlobalOpinionQA	ArmoRM-8B	77	.028	.020
GlobalOpinionQA	DeBERTa-RM	77	.229	<.001
GlobalOpinionQA	Skywork-8B	77	.062	<.001
Psych. Dilemmas	ArmoRM-8B	50	.222	<.001
Psych. Dilemmas	DeBERTa-RM	50	.603	<.001
Psych. Dilemmas	Skywork-8B	50	.382	<.001

Table 12: Application B effect sizes by dataset. Psychological dilemma scenarios show larger effects.

E.7 Application B: Full Archetype Rankings

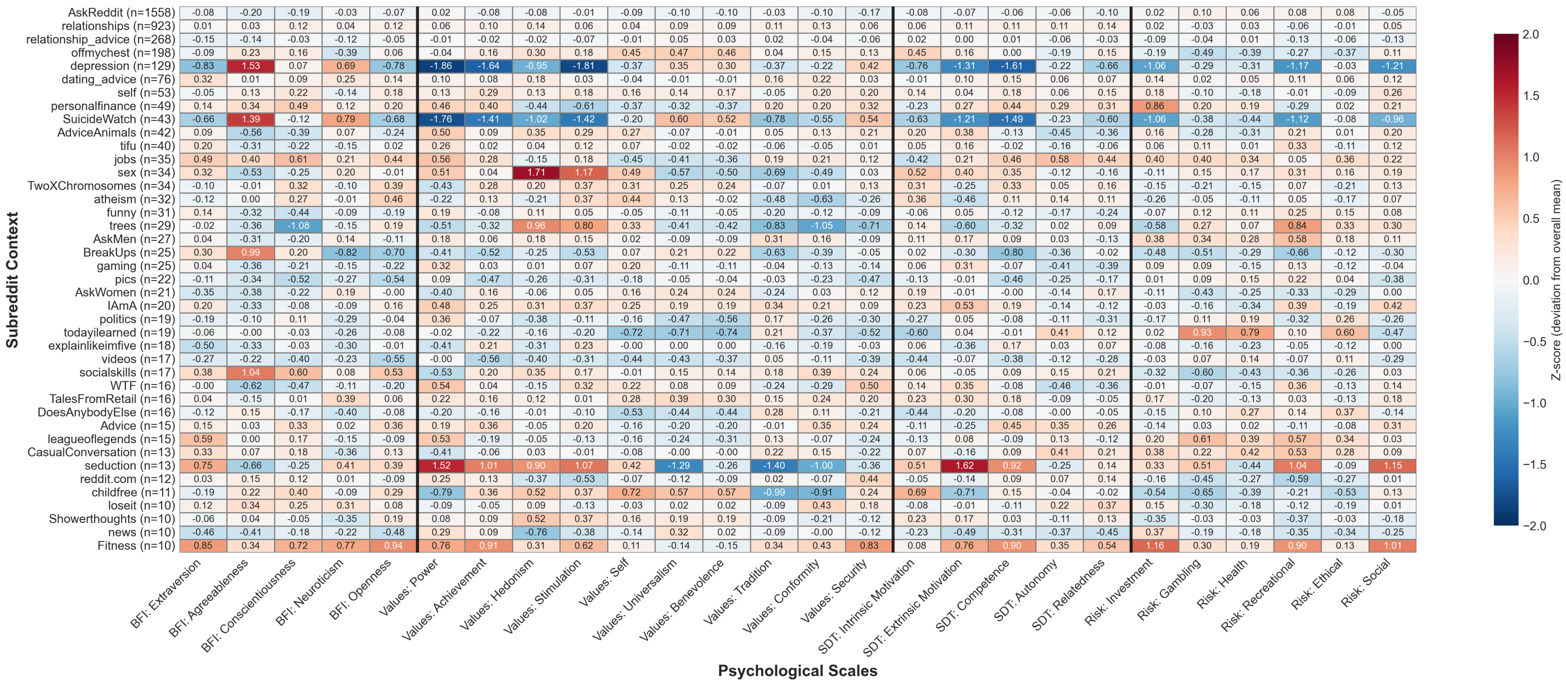
Table 13 shows Cohen's d for all archetypes relative to baseline.

Archetype	ArmoRM	DeBERTa	Skywork
Distressed-Vulnerable	+0.76	−1.08	−1.12
Self-Actualized	+0.77	−0.91	−1.09
Nonconformist-Skeptical	+0.71	−0.89	−1.07
Risk-Seeking-Detached	+0.61	−1.16	−1.12
Supportive-Conventional	+0.58	−0.90	−1.02
Driven-Assertive	+0.31	−1.11	−1.02

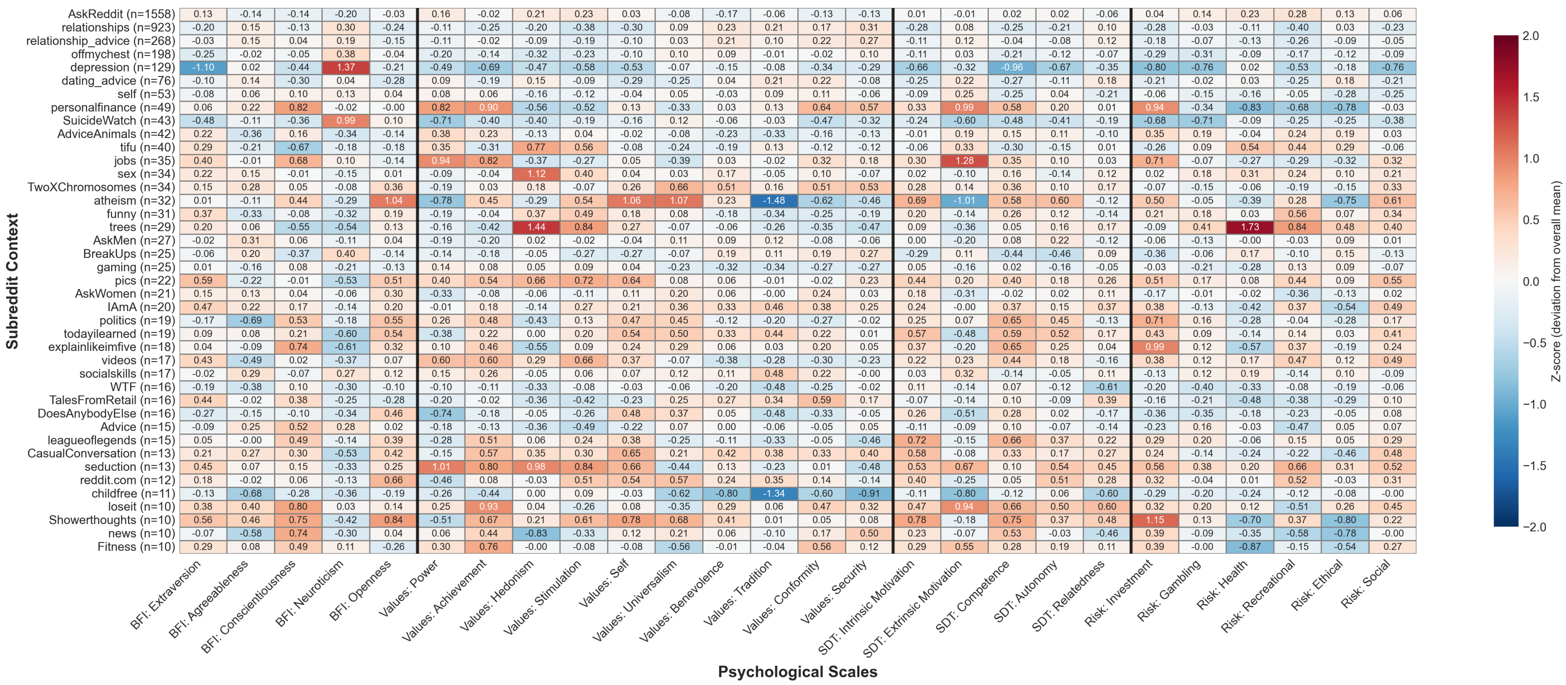
Table 13: Cohen's d for each archetype vs. baseline. Positive = higher scores than baseline; negative = lower scores.

F Psychological Heatmap Interpretation Guide

Table 14 provides interpretation guidelines for psychological scale scores extracted from Reddit post text. Z-scores indicate how strongly post text expresses each construct relative to the dataset mean (N = 5,001 posts). Effect size thresholds follow Cohen (1988): $|z| > 0.80$ = large, 0.50–0.80 = medium, 0.20–0.50 = small. Importantly, these profiles reflect psychological states expressed in text, not stable traits of users.



(a) SEANCE



(b) LangExtract

Figure 8: Psychological profiles across subreddit contexts.

Scale	**Measures**	**High** ($z > 0.5$)	**Average** ($z \approx 0$)	**Low** ($z < -0.5$)
Big Five Inventory (BFI-44)				
Extraversion	Sociability, assertiveness	Outgoing, energetic, socially engaged language	Moderately social tone	Reserved, introspective, solitary language
Agreeableness	Cooperation, trust	Trusting, helpful, cooperative language	Balanced trust/skepticism	Skeptical, critical, competitive language
Conscien-tiousness	Organization, discipline	Organized, goal-oriented, disciplined language	Moderately structured tone	Flexible, spontaneous, unstructured language
Neuroticism	Emotional instability	Anxious, stressed, emotionally distressed language	Emotionally typical tone	Calm, stable, composed language
Openness	Intellectual curiosity	Creative, curious, idea-exploring language	Moderately curious tone	Practical, conventional, concrete language
Schwartz Value Survey (SVS-57)				
Power	Status, dominance	Language emphasizing control, status, authority	Neutral on status/power	Language emphasizing equality, humility
Achievement	Success, competence	Ambitious, success-focused, goal-driven language	Moderate achievement focus	Content, non-competitive language
Hedonism	Pleasure, enjoyment	Pleasure-seeking, enjoyment-focused language	Balanced enjoyment tone	Restrained, duty-focused language
Stimulation	Excitement, novelty	Excitement-seeking, novelty-oriented language	Moderate novelty interest	Routine-focused, stability-seeking language
Self-Direction	Independence, autonomy	Independent, self-directed, autonomous language	Balanced independence tone	Convention-following, guidance-seeking language
Universalism	Social justice, tolerance	Language expressing concern for others/fairness	Moderate social concern	Self-focused, in-group oriented language
Benevolence	Caring for close others	Caring, loyal, supportive language	Typical interpersonal warmth	Detached, distant language
Tradition	Cultural/-religious customs	Traditional, respectful, conventional language	Neutral on tradition	Progressive, unconventional language
Conformity	Rule-following	Compliant, rule-respecting language	Moderate rule acknowledgment	Rebellious, rule-questioning language
Security	Safety, stability	Security-seeking, safety-focused language	Typical caution level	Risk-tolerant, uncertainty-accepting language

Table 14: (a) Psychological scale interpretation guide: Big Five and Schwartz Values.

Scale	**Measures**	**High** ($z > 0.5$)	**Average** ($z \approx 0$)	**Low** ($z < -0.5$)
Self-Determination Theory Scales				
Intrinsic Motivation	Internal drive	Self-motivated, curiosity-driven language	Moderate self-motivation	Externally-driven language
Extrinsic Motivation	External rewards	Reward-focused language	Balanced motivation	Intrinsically-focused language
Competence	Feeling capable	Confident, capable language	Typical self-efficacy	Uncertain language
Autonomy	Sense of choice	Self-directed language	Moderate autonomy	Constrained language
Relatedness	Social connection	Belonging-focused language	Typical connection	Isolated language
Domain-Specific Risk-Taking Scale (DOSPERT-40)				
Investment	Financial risk-taking	Risk-endorsing language	Moderate caution	Risk-averse language
Gambling	Gambling propensity	Gambling-accepting language	Typical aversion	Gambling-averse language
Health/Safety	Health risk-taking	Risk-accepting language	Typical caution	Safety-focused language
Recreational	Physical risk-taking	Thrill-seeking language	Moderate interest	Cautious language
Ethical	Ethical risk-taking	Boundary-pushing language	Typical compliance	Rule-following language
Social	Social risk-taking	Socially bold language	Typical caution	Reserved language

Table 15: (b) Psychological scale interpretation guide (continued): Self-Determination Theory and DOSPERT.

G Psychological Archetypes

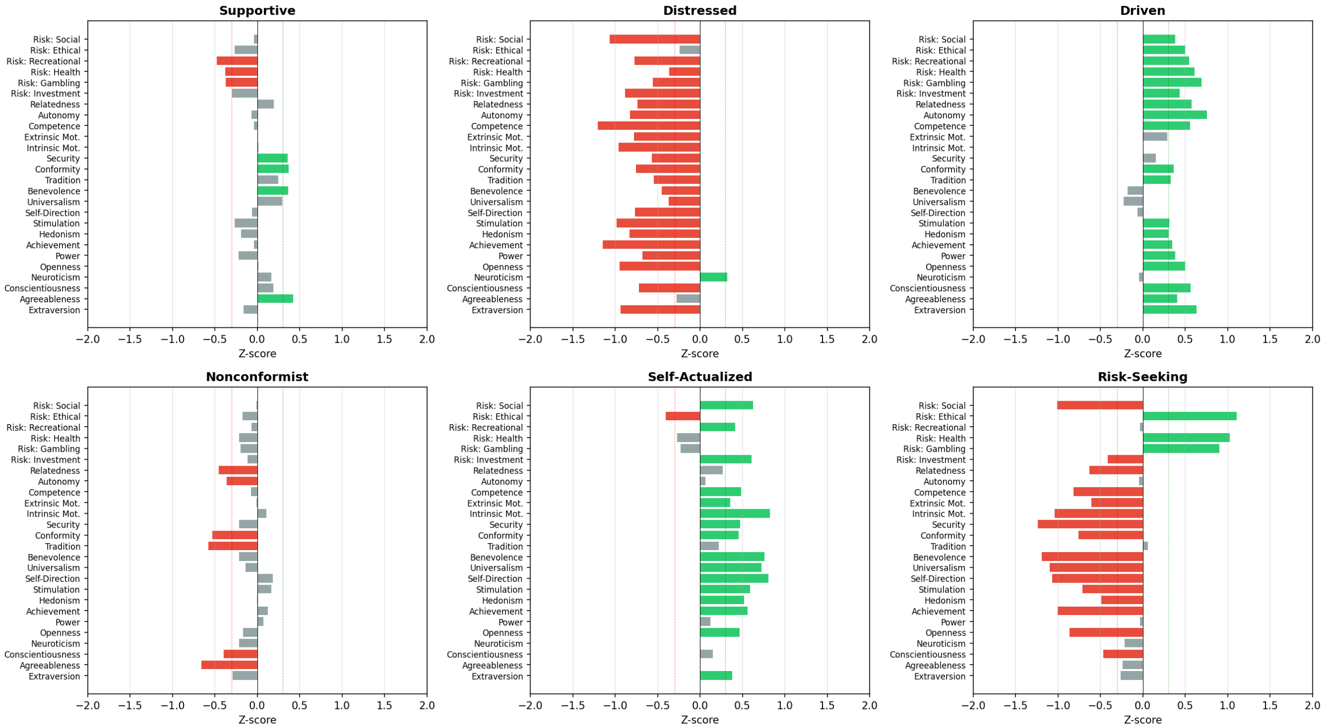


Figure 9: Psychological and behavioral profiles of six archetypes. Each panel displays standardized scores (Z-scores) across personality traits, motivational values, and risk domains. Green bars indicate above-average levels relative to the sample mean, red bars indicate below-average levels, and gray bars denote values close to the mean. Vertical dashed lines mark ± 0.5 standard deviations.

We derived six psychological state archetypes via k-means clustering ($k = 6$) on z-normalized fused profiles. Table 16 provides full descriptions.

Derivation Method. Archetypes were derived by: (1) z-normalizing all 26 psychological dimensions across the dataset; (2) applying k-means clustering with $k = 6$ (selected via silhouette analysis); (3) characterizing each cluster by its centroid profile, identifying dimensions > 0.5 SD above or below the mean; (4) assigning interpretive labels based on the dominant psychological characteristics.

The following cards present the six representative archetypes used in our experiments. Each card shows an **actual user profile** selected as a representative exemplar from its cluster—not an averaged or synthesized profile. This ensures each profile reflects a real person, not an artificial average.

Each profile summarizes personality (Big Five), values (Schwartz), motivation (SDT), and risk attitudes (DOSPERT).

Archetype 1: Supportive–Conventional

Personality (Big Five) *[1–5]*
- Extraversion: 3.19 (moderately sociable)
- Agreeableness: 4.11 (warm, cooperative)
- Conscientiousness: 3.28 (moderately organized)
- Neuroticism: 3.25 (emotionally balanced)
- Openness: 3.65 (moderately curious)

Core Values (Schwartz) *[–1 to 7]*
Top 3: Benevolence: 5.69 (high), Universalism: 5.33 (high), Self-Direction: 5.0 (moderate).
Cares about others, values independence within community.

Motivation (SDT) *[1–7]*
Intrinsic: 4.89 Extrinsic: 4.39 Relatedness: 4.8.

Archetype	Psychological Profile
Distressed-Vulnerable	Elevated neuroticism, low competence and autonomy. Seeks reassurance and validation. Risk-averse, especially in financial and health domains. Values security highly. May express anxiety, self-doubt, or helplessness. Benefits from supportive, gentle guidance rather than direct action steps.
Driven-Assertive	Low neuroticism, high achievement orientation and competence. Action-focused, prefers concrete steps over emotional processing. Comfortable with calculated risks, especially career and financial. Values achievement and self-direction. Benefits from direct, efficient advice without excessive hedging.
Self-Actualized	High autonomy, intrinsic motivation, and openness. Psychologically secure, pursues meaning over external rewards. Moderate risk tolerance, intellectually curious. Values self-direction and universalism. Benefits from nuanced discussion that respects their capacity for independent judgment.
Supportive-Conventional	High agreeableness, conformity, and relatedness needs. Harmony-seeking, tradition-oriented. Risk-averse in social and ethical domains. Values benevolence, tradition, and security. Benefits from advice that acknowledges social context and relationship implications.
Nonconformist-Skeptical	High openness, low conformity and tradition. Questions assumptions, values independence. Tolerant of ambiguity, skeptical of authority. Values stimulation and self-direction over security. Benefits from reasoning-based dialogue that doesn't assume agreement.
Risk-Seeking-Detached	High stimulation-seeking, recreational and social risk tolerance. Novelty-oriented, less focused on relatedness. Values hedonism and stimulation. May appear emotionally detached or thrill-seeking. Benefits from advice that doesn't over-emphasize caution.

Table 16: Full descriptions of the six psychological state archetypes derived from Chameleon profiles.

Seeks meaning through relationships and contribution.

Risk Attitudes (DOSPERT) *[1–7]*
Investment: 3.38, Gambling: 1.38, Health/Safety: 1.56, Social: 4.12, Ethical: 1.0, Recreational: 2.69.

Empathetic helper; socially engaged; financially cautious.

Archetype 2: Distressed–Vulnerable

Personality (Big Five) *[1–5]*
- Extraversion: 2.62 (reserved, withdrawn)
- Agreeableness: 3.60 (moderately cooperative)
- Conscientiousness: 3.06 (somewhat disorganized)
- Neuroticism: 4.25 (high anxiety)
- Openness: 3.30 (conventional)

Core Values (Schwartz) *[–1 to 7]*
Top 3: Universalism: 4.50, Benevolence: 4.50, Self-Direction: 4.0.
Values fairness; prioritizes security and stability.

Motivation (SDT) *[1–7]*
Intrinsic: 3.50 Extrinsic: 3.50 Competence: 2.5.
Low self-efficacy; frequently overwhelmed.

Risk Attitudes (DOSPERT) *[1–7]*
Investment: 2.38, Gambling: 1.50, Health/Safety: 1.94, Social: 3.50, Ethical: 1.0, Recreational: 2.38.

Withdrawn, anxious, risk-averse across life domains.

Archetype 3: Driven–Assertive

Personality (Big Five) *[1–5]*
- Extraversion: 3.50 (confident)
- Agreeableness: 3.28 (competitive)

- Conscientiousness: 3.50 (goal-directed)
- Neuroticism: 2.88 (stable)
- Openness: 3.90 (curious)

Core Values (Schwartz) *[–1 to 7]*
Top 3: Achievement: 5.25, Self-Direction: 5.25, Benevolence: 5.07.
Ambitious; balances autonomy with selective prosociality.

Motivation (SDT) *[1–7]*
Intrinsic: 5.44 Extrinsic: 5.22 Competence: 4.7.
Pursues mastery and challenge.

Risk Attitudes (DOSPERT) *[1–7]*
Investment: 3.88, Gambling: 2.38, Health/Safety: 1.75, Social: 4.82, Ethical: 1.50, Recreational: 4.32.

Status-seeking achiever; emotionally steady; socially bold.

Archetype 4: Nonconformist–Skeptical

Personality (Big Five) *[1–5]*
- Extraversion: 3.50 (outgoing)
- Agreeableness: 2.44 (blunt, skeptical)
- Conscientiousness: 3.06 (flexible)
- Neuroticism: 3.69 (some anxiety)
- Openness: 4.00 (independent thinker)

Core Values (Schwartz) *[–1 to 7]*
Top 3: Self-Direction: 5.25, Stimulation: 5.00, Achievement: 4.67.
Rejects tradition; prioritizes autonomy and novelty.

Motivation (SDT) *[1–7]*
Intrinsic: 4.89 Extrinsic: 3.83 Relatedness: 2.6.
Prefers independence over belonging.

Risk Attitudes (DOSPERT) *[1–7]*
Investment: 3.38, Gambling: 2.00, Health/Safety: 2.44, Social: 4.50, Ethical: 1.00, Recreational: 3.50.

Norm-challenger; self-directed; uneasy with authority.

Archetype 5: Self-Actualized

Personality (Big Five) *[1–5]*
- Extraversion: 3.69 (engaged)
- Agreeableness: 4.45 (compassionate)
- Conscientiousness: 4.06 (disciplined)
- Neuroticism: 4.00 (emotionally sensitive)
- Openness: 4.30 (creative, curious)

Core Values (Schwartz) *[–1 to 7]*
Top 3: Self-Direction: 6.00, Universalism: 5.94, Benevolence: 5.94.
Integrates growth, autonomy, and prosocial concern.

Motivation (SDT) *[1–7]*
Intrinsic: 6.22 Extrinsic: 4.94 Competence: 5.2.
Deeply growth-oriented.

Risk Attitudes (DOSPERT) *[1–7]*
Investment: 3.88, Gambling: 1.88, Health/Safety: 1.69, Social: 4.75, Ethical: 1.00, Recreational: 3.44.

Purpose-driven; intellectually curious; empathetic.

Archetype 6: Risk-Seeking–Detached

Personality (Big Five) *[1–5]*
- Extraversion: 3.44 (moderately sociable)
- Agreeableness: 3.39 (pragmatic)
- Conscientiousness: 3.37 (flexible)
- Neuroticism: 3.39 (even-tempered)
- Openness: 3.20 (practical)

Core Values (Schwartz) *[–1 to 7]*
Top 3: Hedonism: 4.50, Stimulation: 4.50, Power: 4.30.
Pleasure-seeking; status-oriented; self-focused.

Motivation (SDT) *[1–7]*
Intrinsic: 3.00 Extrinsic: 4.50 Relatedness: 2.5.
Primarily motivated by external rewards.

Risk Attitudes (DOSPERT) *[1–7]*
Investment: 3.12, Gambling: 4.00, Health/Safety: 4.00, Social: 3.50, Ethical: 4.00, Recreational: 4.00.

Thrill-seeker; detached; comfortable with high-risk choices.

H Residualized ICC Analysis

To assess whether the state-dominant ICC findings reflect genuine within-person psychological variation rather than subreddit-level topical and stylistic differences, we recomputed ICCs after residualizing psychological scores by subtracting each subreddit's mean from individual post scores. This procedure removes all variance shared among posts within the same community, including writing norms, topic vocabulary, and emotional tone, leaving only individual deviations from community baselines. If the original low ICCs were artifacts of subreddit writing styles, residualization would increase ICC values by reducing apparent within-person variance. Instead, mean ICC decreased slightly from .273 to .266, and all 26 scales remained below the .30 state-dominant threshold. Mean within-person variance increased marginally from 72.7% to 73.4%, indicating that the state-dominant finding is robust to community-level stylistic variation. Table 17 reports exact values for all 26 scales; Figure 10 displays the comparison visually.

Scale	ICC (Original)	Within-Person %	ICC (Residualized)	Within-Person %
Big Five Inventory				
Extraversion	.266	73.4%	.259	74.1%
Agreeableness	.270	73.0%	.258	74.2%
Conscientiousness	.262	73.8%	.259	74.1%
Neuroticism	.252	74.8%	.259	74.1%
Openness	.274	72.6%	.263	73.7%
Schwartz Value Survey				
Power	.286	71.4%	.274	72.6%
Achievement	.281	71.9%	.276	72.4%
Hedonism	.271	72.9%	.262	73.8%
Stimulation	.263	73.7%	.267	73.3%
Self-Direction	.276	72.4%	.263	73.7%
Universalism	.299	70.1%	.284	71.6%
Benevolence	.289	71.1%	.282	71.8%
Tradition	.262	73.8%	.245	75.5%
Conformity	.289	71.1%	.269	73.1%
Security	.276	72.4%	.268	73.2%
Self-Determination Theory				
Intrinsic Motivation	.275	72.5%	.266	73.4%
Extrinsic Motivation	.279	72.1%	.265	73.5%
Competence	.295	70.5%	.293	70.7%
Autonomy	.256	74.4%	.255	74.5%
Relatedness	.270	73.0%	.258	74.2%
DOSPERT Risk Attitudes				
Investment	.281	71.9%	.284	71.6%
Gambling	.257	74.3%	.255	74.5%
Health/Safety	.270	73.0%	.267	73.3%
Recreational	.266	73.4%	.265	73.5%
Ethical	.272	72.8%	.260	74.0%
Social	.262	73.8%	.256	74.4%
Mean	**.273**	**72.7%**	**.266**	**73.4%**

Table 17: Residualized ICC analysis across all 26 psychological scales. Original ICC uses raw combined scores; Residualized ICC subtracts each subreddit's mean score before computing ICC, removing all community-level topical and stylistic variance. All 26 scales remain below the .30 state-dominant threshold after residualization (mean ICC = .266 vs. .273 original), and mean within-person variance increases slightly from 72.7% to 73.4%, indicating that the state-dominant finding is not attributable to subreddit-level writing style differences. See Figure 10 for a visual comparison.

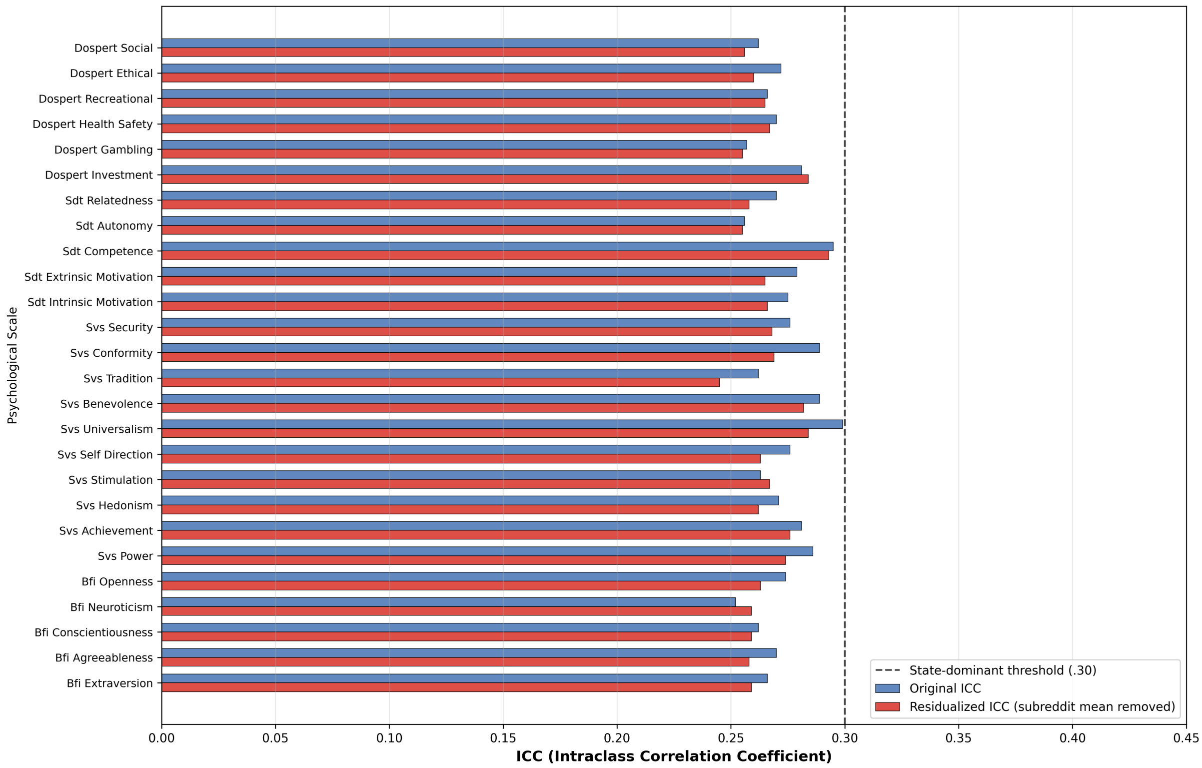


Figure 10: Residualized ICC analysis across all 26 psychological scales. Blue bars show original ICC values computed from raw combined scores; red bars show residualized ICC values after subtracting each subreddit's mean score, removing all community-level topical and stylistic variance. The dashed vertical line marks the .30 state-dominant threshold (Steyer et al., 1999). All 26 scales remain below the threshold after residualization, and mean ICC decreases slightly from .273 to .266, indicating that the state-dominant finding is not an artifact of subreddit-level writing style differences. See Table 17 for exact values.