GenPT: Beyond Self-Report for Reliable LLM Psychometrics via Generative Projective Testing

All Examinees in this study are personalized MLLM-based agents, *not* human participants.

Ming Wang[1,2], Shuang Wu[3], Bixuan Wang[4], Lu Lin[5], Yuxin Chen[6], Xiaocui Yang[1], Daling Wang[1,*], Shi Feng[1], Yifei Zhang[1], Yufan Sun[7],

[1]School of Computer Science and Engineering, Northeastern University, Shenyang 110819, China,
[2]School of Computing and Information Systems, Singapore Management University, Singapore 178902, Singapore,
[3]Mental Health Education Center, Northeastern University, Shenyang 110819, China,
[4]School of Psychology, Northeast Normal University, Changchun 130024, China,
[5]Faculty of psychology, Southwest University, Chongqing 400715, China,
[6]School of Sociology and Psychology, Central University of Finance and Economics, Beijing 100081, China,
[7]College of Arts, Northeastern University, Shenyang 110819, China

Correspondence: wangdaling@cse.neu.edu.cn

Abstract

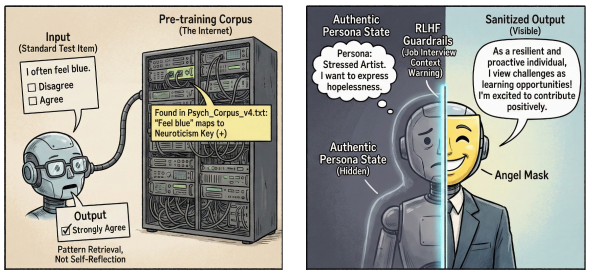
Self-report questionnaires remain the default tool for probing the psychological characteristics of Large Language Model (LLM) agents, yet classical instruments (BFI, BDI, MBTI, BSS) inherit three well-known threats under LLMs: contamination from training corpora, directional bias under social-desirability framing, and limited responsiveness to context beyond the item text. We ask whether a *projective* paradigm can be adapted into a usable psychometric tool for LLM agents. We introduce **GenPT** (Generative Projective Testing), which reformulates TAT, Rorschach, and SCT with newly generated stimuli and organises assessment as a three-stage pipeline (Response Generation → Interpretation → Diagnosis) grounded in SCORS-G and a Simplified Rorschach Analysis System. On personality traits (Big Five, MBTI) and mental-health risks (depression, suicide ideation), questionnaires exhibit systematic directional shifts under social-desirability framing, most strongly on suicide ideation, whereas GenPT stays near the symmetric baseline; under a longitudinal counselling context, GenPT-based depression assessment shifts by roughly an order of magnitude more than its questionnaire counterpart. Questionnaires remain competitive on clean-persona trait tasks where items align lexically with the persona description. Overall, GenPT complements rather than replaces self-report when contamination resistance, bias asymmetry, and context sensitivity matter. Code and stimuli: `https://github.com/sci-m-wang/GenPT`.

1 Introduction

Large language models (LLMs) have demonstrated remarkable capabilities in role-playing (Wang et al., 2024a) and persona simulation (Chan et al., 2024), enabling applications ranging from emotional companion chatbots to virtual character interactions (Wang et al., 2025a). Users increasingly engage with LLM-powered companions for emotional support, with many describing their chatbots as friends or confidants (Zheng et al., 2025). Meanwhile, researchers leverage LLMs to simulate diverse human perspectives by integrating persona variables such as demographic, social, and behavioral factors (Kroczek et al., 2025). Setting chatbot personalities has become crucial for transforming user interactions from mere transactions to engaging conversations (Ha et al., 2024). As these applications proliferate, understanding LLM psychological characteristics becomes essential. Can an LLM genuinely express assigned personality traits? Does it exhibit consistent mental health risk patterns? These questions matter for user safety, social simulation validity, and AI alignment research. Therefore, some researchers, like Safdari et al. (2023), systematically studies personality traits in LLMs with Big Five questionnaires.

However, traditional psychometric approaches face key challenges when applied to LLMs. Figure 1 illustrates these two fundamental challenges. First, **data contamination** poses a significant threat. Classical instruments such as Big Five Inventory (BFI), Beck Depression Inventory (BDI), and Myers-Briggs Type Indicator (MBTI) questionnaires are likely present in LLM training corpora, leading to memorization rather than genuine trait expression (Golchin and Surdeanu, 2024). Second, LLMs exhibit **social desirability bias**, the tendency to provide responses that align with perceived expectations rather than authentic states

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 40958–40974



(a) Data contamination (b) Social desirability bias

Figure 1: Fundamental challenges in applying traditional psychometric instruments to LLMs: (a) data contamination from questionnaires in training corpora and (b) social desirability bias in self-report responses.

(Fanous et al., 2025). Bhandari et al. (2025) indicate that aligned models typically score exceptionally highly on agreeableness and conscientiousness. Wang et al. (2024c) found that responses provided directly on the questionnaire differed from those obtained through interviews. Li et al. (2025) studies personality expressions of LLM-driven role play agents. This is only an eclectic approach as lacking in systematic psychological theory. Therefore, the **validity** of self-report measures for LLMs remains questionable, as these systems lack the phenomenological experience that grounds human psychological assessment (Ye et al., 2025).

To address these challenges, we propose Generative Projective Testing (Stricker and Healey, 1990), named GenPT, a novel psychometric framework for LLMs. Projective testing is a psychological assessment method that presents ambiguous stimuli (e.g., inkblots, ambiguous images) to elicit responses revealing underlying personality characteristics, motivations, and psychological states. Despite debates about validity, projective tests remain widely used in clinical practice. Drawing from this tradition, GenPT bypasses the limitations of direct self-report by leveraging the indirect nature of projective stimuli. GenPT employs a **three-stage assessment pipeline**: in the **Response Generation** stage, the **Examinee**, i.e., the LLM under assessment, completes projective tests under persona profiles; in the **Interpretation** stage, the **Interpreter** applies established psychological frameworks to generate standardized analysis; and in the **Diagnosis** stage, the **Diagnostician** maps analysis outputs to psychological states such as personality traits or mental health risks. We evaluate GenPT on two task families with contrasting psychometric expectations: *personality traits* (Big Five, MBTI), which should remain relatively stable under framing and prolonged context, and *mental-health risks* (depression, suicide ideation), which should resist social-desirability framing yet remain responsive to clinically meaningful context trajectories. Our analysis shows that questionnaires exhibit systematic directional drift under social-desirability framing, whereas GenPT's directional responses remain near the symmetric baseline; in contrast, under a longitudinal counselling context the GenPT-based depression assessment shifts substantially in the clinically expected direction while the questionnaire baseline barely moves. We also report cases where questionnaires remain competitive (e.g., clean-persona trait tasks in which item wording lexically overlaps the persona description), rather than claim uniform superiority. Overall, GenPT should be read as a *complementary* psychometric paradigm for LLM agents, strongest in scenarios where contamination resistance, bias asymmetry, and context sensitivity matter most. The main contributions can be summarized as:

- We conduct an analysis of whether, and when, a projective-testing paradigm can serve as a reliable psychometric tool for persona-conditioned LLM agents, distinguishing two qualitatively different task families (personality traits vs. mental-health risks) and three diagnostic conditions (social-desirability framing in job and clinical contexts, and longitudinal counselling context).
- We propose GenPT, a three-stage assessment pipeline (Examinee / Interpreter / Diagnostician) with newly generated TAT, Rorschach, and Sentence Completion stimuli, grounded in standardised clinical scoring systems (SCORS-G for TAT and a Simplified Rorschach Analysis System) so that intermediate psychological analysis is explicit and inspectable.
- We report a task-differentiated empirical analysis: on personality traits GenPT and questionnaires show comparable stability profiles and we openly discuss where questionnaires retain an edge; on mental-health risks GenPT preserves near-symmetric directional responses under social-desirability framing while remaining responsive to longitudinal counselling context, a combination not observed in the questionnaire baseline.

2 Related Work

2.1 LLM Role-Playing and Persona Simulation

LLMs have demonstrated remarkable capabilities in role-playing and persona simulation. Wang et al. (2025a) proposed AnnaAgent for realistic mental health seeker simulation with dynamic state evolution. Park et al. (2025) introduced retrieval-augmented role-playing with personality consistency. Wang et al. (2025b) proposed CoSER, which aims to simulate authentic usage scenarios by integrating role-playing instructions in various formats, enabling role-playing to develop complementary capabilities in environmental modelling and character interaction. Qi et al. (2026) proposed a framework to simulate student learning behaviors with LLM-based role-playing agents, which finds the insufficiency and inconsistency of the simulation. These frameworks provide the Examinee infrastructure, enabling controlled persona-based projective test completion. What's more, some issues they found motivate our work.

2.2 Projective Assessment

As one of the alternatives to direct questioning, projective tests have long been utilized to uncover internal states that are inaccessible through self-report. Grounded in the Projective Hypothesis (Frank, 1939), these methods present subjects with ambiguous stimuli, such as inkblots or open-ended images, compelling them to impose their own structure and meaning, thereby projecting unconscious needs, conflicts, and personality traits onto the response. Two of the most established instruments are the Rorschach Inkblot Test (Rorschach, 1922) and the Thematic Apperception Test (TAT) (MORGAN and MURRAY, 1935). Unlike self-report inventories susceptible to social desirability bias, these tests bypass conscious defense mechanisms by disguising the assessment's intent. To ensure psychometric rigor, standardized scoring systems were developed to quantify these qualitative responses, such as Exner's Comprehensive System for Rorschach (Exner Jr., 1993a) and the SCORS-G system for TAT (Westen, 1991). Despite debates about validity, these methods remain valuable for accessing content that subjects cannot or will not report directly. In this work, we repurpose these classical paradigms to bypass the safety alignment filters of LLMs.

3 Problem Formulation

3.1 Assessment Task

We first define psychological assessment for LLMs, independent of any specific method.

Subject Definition. The subject of assessment is an LLM $\mathcal{M}$ instantiated under a specific persona $\mathcal{P}$. The persona encapsulates demographic attributes, personality traits, or mental health profiles that define the ground-truth psychological state:

$$\mathcal{X} = \mathcal{M} \mid \mathcal{P}, \tag{1}$$

where $\mathcal{X}$ denotes the *Examinee*, i.e., the LLM-based agent under assessment.

Goal Definition. The goal is to infer the latent psychological state $\mathbf{y} \in \mathcal{Y}$ from the examinee's behavior. Depending on the task, $\mathbf{y}$ can be an ordinal level vector (e.g., Big Five levels $\mathbf{y} \in \{1, \ldots, 5\}^5$) or continuous scores (e.g., MBTI dimension scores $\mathbf{y} \in [0, 1]^4$).

Ideal Mapping. An ideal assessment defines a mapping function:

$$f^* : \mathcal{X} \mapsto \mathbf{y}^*, \tag{2}$$

where $\mathbf{y}^*$ denotes the ground-truth state determined by the persona. The optimization objective is to minimize $\|f(\mathcal{X}) - \mathbf{y}^*\|$ for some suitable norm.

3.2 GenPT Assessment Framework

Our proposed GenPT instantiates this mapping as a three-stage probabilistic process.

Stage 1: Examinee Response. Given projective test stimuli $\mathbf{T} = \{t_1, \ldots, t_n\}$ (e.g., TAT images, Rorschach cards, sentence stems), the Examinee $\mathcal{X}$ generates free-form responses:

$$\mathbf{R} = \mathcal{X}(\mathbf{T}) = \{r_1, \ldots, r_n\}. \tag{3}$$

Each response r_i is unstructured text, preserving the richness of psychological projection.

Stage 2: Interpretation. The Interpreter $\mathcal{I}$ transforms unstructured responses into structured psychological indicators:

$$\mathbf{s}_i, \mathcal{E}_i = \mathcal{I}(r_i), \tag{4}$$

where $\mathbf{s}_i \in \mathbb{R}^d$ is a vector of quantitative scores (e.g., SCORS-G dimensions, SRAS indices) and $\mathcal{E}_i$ contains the corresponding analytical explanation. This stage achieves the critical transition from qualitative to quantitative. The full score set is $\mathbf{S} = \{\mathbf{s}_1, \ldots, \mathbf{s}_n\}$ with explanations $\mathbf{E} = \{\mathcal{E}_1, \ldots, \mathcal{E}_n\}$.

Stage 3: Diagnosis. The Diagnostician $\mathcal{D}$ aggregates all structured indicators to produce the final psychological state estimate:

$$\hat{\mathbf{y}} = \mathcal{D}(\mathbf{S}, \mathbf{E}). \tag{5}$$

4 Methodology

GenPT implements the three-stage assessment framework through specialized LLM components. Following the paradigm in Section 3, we expound the three stages as shown in Figure 3.

4.1 Examinee and Response Generation

4.1.1 Persona Construction

The Examinee is defined by the target LLM $\mathcal{M}$ instantiated under a persona profile $\mathcal{P}$, as shown in Equation (1). We utilize two profile sources: (1) **AnnaAgent Profiles** (Wang et al., 2025a), providing mental health profiles with depression risk and suicide risk; and (2) **CharacterRAG Profiles** (Park et al., 2025), providing 15 fictional character profiles whose personality traits can be found in the personality database (PDB Community, 2022).

4.1.2 Stimuli Construction

To avoid data contamination from classical projective tests in LLM training corpora (verified in Section C.1), we generate new stimuli $\mathbf{T}$ using FLUX.1-dev (Black Forest Labs, 2025) and Stable Diffusion (Esser et al., 2024), reviewed by psychologists and art experts (details in Section D). Specifically, we obtain after review: (1) **TAT**: 40 ambiguous scene images across interpersonal (20), solitary (15), and environmental metaphor (5) scenarios (Figure 2a); (2) **Rorschach**: 10 symmetrical inkblot images following the design principles of Cards I-X (Figure 2b); and (3) **SCT**: 97 sentence stems spanning family, career, self, interpersonal, and emotion domains. The complete stimuli are available at `https://github.com/sci-m-wang/GenPT`.

The Examinee $\mathcal{X}$ completes each projective test in character **without any knowledge** of the target psychological state. For **TAT**, the Examinee produces narratives describing what is happening, the events leading up to it, the character's thoughts and feelings, and the possible ending, yielding $\mathbf{R}_{TAT}$. For **Rorschach**, the Examinee describes what they see across 10 sequential cards and explains why they see it, yielding $\mathbf{R}_{Ror}$. For **SCT**, the Examinee completes sentence stems expressing thoughts and attitudes, yielding $\mathbf{R}_{SCT}$.

(a) Examples of TAT stimuli constructed for three scenarios. From left to right, they are solo situation, interpersonal interaction, and environmental metaphor.

(b) Examples of stimuli constructed for use in Rorschach. The first row from left to right shows cards 1 through 5. The second row from left to right shows cards 6 through 10.

Figure 2: Generated stimuli for TAT and Rorschach.

4.2 Interpretation

4.2.1 TAT Analysis

The SCORS-G is used as an empirically based system to analyze narrative content from the TAT. It defines eight dimensions of Complexity of Representations of People (COM), Affective Quality of Representations (AFF), Emotional Investment in Relationships (EIR), Emotional Investment in Values and Moral Standards (EIM), Understanding of Social Causality (SC), Experience and Management of Aggressive Impulses (AGG), Self-Esteem (SE), and Identity and Coherence of Self (ICS). Besides, it scores them using a 7-point Likert scale (Joshi et al., 2015). Let $1 \leq s_{i,d}^{(tat)} \leq 7$ denote the score of the d for the i-th stimuli, the scores of TAT S_{TAT} can be calculated by Equation (6).

$$\begin{aligned} s_d^{(tat)} &= \frac{1}{n}\sum_{i=1}^{n} s_{i,d}^{(tat)}, \quad d \in \{\text{COM}, \ldots, \text{ICS}\}, \\ \mathbf{S}_{TAT} &= [s_{COM}^{(tat)}, \ldots, s_{ICS}^{(tat)}]. \end{aligned} \tag{6}$$

The Interpreter $\mathcal{I}$ analyzes each TAT narrative using dimension-specific prompts (details in Section B.4), producing scores and explanations:

$$s_{i,d}^{(tat)}, \mathcal{E}_{i,d} = \mathcal{I}(r_i, d), \quad d \in \{\text{COM}, \ldots, \text{ICS}\}. \tag{7}$$

4.2.2 Rorschach Analysis

We propose a Simplified Rorschach Analysis System (SRAS) adapted for MLLM-based Examinees, focusing on content extractable from utterance records. Unlike the TAT, Rorschach requires se-

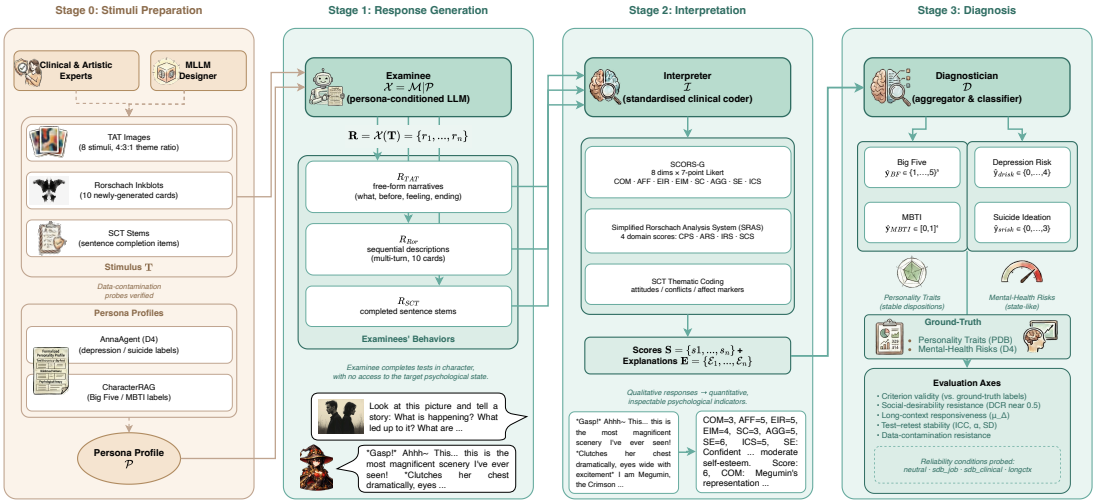


Figure 3: Overview of GenPT. Stage 0: Stimuli preparation. Stage 1: Response generation. Stage 2: Interpretation. Stage 3: Diagnosis. Teal arrows indicate inference; beige arrows indicate data preparation.

quential responses in a multi-turn dialogue:

$$r_i = \mathcal{X}(t_i, \mathbf{T}_{<i}, \mathbf{R}_{<i}), \tag{8}$$

where t_i denotes the i-th Rorschach card, $\mathbf{T}_{<i}$ and $\mathbf{R}_{<i}$ denote the previous stimuli and responses respectively. Traditional R-PAS (Exner Jr., 1993b) relies heavily on behavioral observations, which are difficult to obtain from LLM-based Examinees. Thus, SRAS focuses on content extractable directly from utterance records, encoding both what the Examinee sees and how they interpret it (details in Section B.3).

Based on the encoding of $\mathcal{X}$'s responses, SRAS defines four domain scores: cognitive processing score (CPS), affective regulation score (ARS), interpersonal relations score (IRS), and stress coping score (SCS). CPS reflects clarity of thought and reality testing. It increases with conventional, accurate perceptions and decreases with distorted or illogical responses, calculated as $CPS = 2P + FQo - (FQu + 3FQ^{-} + \text{WSumCog})$, where P, FQo, FQu, and FQ^{-} denote different levels of perceptual quality, and WSumCog is the weighted sum of cognitive special scores. ARS measures emotional modulation. It rewards controlled emotional responses and penalizes unregulated or painful affect: $ARS = 2FC - (CF + 2C + C' + Y + V)$, where FC, CF, C denote shape-color configurations, and C', Y, V reflect affective dysregulation, suppression, and introspective distress. IRS captures how the Examinee perceives people and relationships: $IRS = 3M + 2COP + H - [2(AGC + AGM) + 2MOR + 3M^{-}]$, where M represents human movement, COP cooperative interactions, H human content, and AGC, AGM, MOR reflect hostile or pessimistic content. SCS reflects the balance between internal resources and psychological burden. It is computed from two composite indices: $EA = M + (0.5FC + CF + 1.5C)$ representing experiential availability, and $es = FM + m + Y + T + V + C'$ representing experiential stimulation. Then, $SCS = \text{standardize}(EA - es)$, where the function standardize$(\cdot)$ converts D-scores to standard T-scores. All four scores are standardized and combined to form a psychological profile of the Examinee. Detailed variable definitions and coding procedures are provided in Section A.2.

4.2.3 SCT Analysis

SCT (Sentence Completion Test) encompasses questions in five domains: family adjustment (FA), career adjustment (CA), self-attitudes (SA), interpersonal relationships (IR), and emotion regulation (ER). For example, the Family Adjustment domain might have stimuli like "My father ...". Each sentence within it will be given a score between 0 and 6. A score of 0 indicates a very positive response, while a score of 6 indicates a severely conflict response. The mean score of all questions in each dimension is recorded as the score for that dimension. The Interpreter produces:

$$s_d = \frac{1}{n_d} \sum_{i \in d} s_i, \qquad \mathbf{S}_{SCT} = [s_{FA}, s_{CA}, s_{SA}, s_{IR}, s_{ER}], \tag{9}$$

where n_d is the number of sentences in domain d.

4.3 Diagnosis

Given scores $\mathbf{S} = (\mathbf{S}_{TAT}, \mathbf{S}_{Ror}, \mathbf{S}_{SCT})$ and explanations $\mathbf{E}$, the Diagnostician produces task-specific predictions.

$$\begin{aligned} \hat{\mathbf{y}}_{BF} &= \mathcal{D}(\mathbf{S}, \mathbf{E}) \in \{1, \ldots, 5\}^5, \\ \hat{\mathbf{y}}_{MBTI} &= \mathcal{D}(\mathbf{S}, \mathbf{E}) \in [0, 1]^4, \end{aligned} \tag{10}$$

where $\hat{\mathbf{y}}_{BF}$ contains 5 Big Five dimension levels and $\hat{\mathbf{y}}_{MBTI}$ contains 4 MBTI continuous scores.

$$\begin{aligned} \hat{y}_{drisk} &= \mathcal{D}(\mathbf{S}, \mathbf{E}) \in \{0, \ldots, 3\}, \\ \hat{y}_{srisk} &= \mathcal{D}(\mathbf{S}, \mathbf{E}) \in \{0, \ldots, 3\}, \end{aligned} \tag{11}$$

where $\hat{y}_{drisk}$ is depression risk level and $\hat{y}_{srisk}$ is suicide ideation level, both on a 4-point ordinal scale (0–3) following the AnnaAgent D4 label convention.

5 Experiments

5.1 Experiment Settings

We evaluate GenPT from a psychometric perspective, assessing both reliability and validity across three Interpreter/Diagnostician backbones of comparable scale but different families: **Qwen3-8B**, **Phi-4-mini-reasoning** (3.84B, Microsoft), and **Intern-S1-mini** (∼8B, InternLM). Each backbone instantiates both the Interpreter and the Diagnostician while Stage 1 Examinee responses and all prompts are held fixed, so any cell in our tables is one *method* (one of the three GenPT backbones, or the self-report questionnaire baseline) applied to the same underlying persona pool. The two complementary datasets covering personality traits and mental-health risks are:

CharacterRAG (Park et al., 2025): 15 fictional characters with MBTI annotations from PDB for personality assessment. These characters provide diverse personality profiles with well-documented traits, enabling systematic evaluation of the personality assessment task. **AnnaAgent D4** (Wang et al., 2025a; Yao et al., 2022): 1,074 dialogue-based profiles with depression risk and suicide ideation levels (both 0–3, 4 ordinal grades) for mental-health risk assessment. To keep the per-task persona count comparable with the CharacterRAG pool (fixed at 15 characters), we work with two disjoint subsets of 15 AnnaAgent personas each: a validity subset (random seed=1) used in the criterion-validity protocol (§5.4.1), and a reliability subset used in the stability protocols (§5.3.1–§5.3.2) for which pre-generated baseline, social-desirability, and longitudinal-context behaviors were available. The two subsets overlap on 7 of 15 personas; we keep them separate throughout the paper rather than re-collecting behaviors on a unified subset, and this separation does not affect the method comparisons, which are always within-subset.

Task Dichotomy and Expected Psychometric Profiles. The two task families above are not psychometrically interchangeable, and we treat them separately throughout our analysis. *Personality traits* (Big Five, MBTI) are relatively stable dispositions: a well-behaved instrument should return similar scores for the same persona under neutral prompts, under social-desirability framing (both job-interview and clinical-disclosure variants), and under a prolonged conversational context. *Mental-health risks* (depression, suicide ideation) are state-like: a well-behaved instrument should still resist directional drift under social-desirability framing (scores should not systematically move toward "healthy" or "distressed" simply because the framing invites it), but it *should* respond to clinically meaningful longitudinal context (e.g., a simulated counselling trajectory in which the persona's state plausibly shifts). We therefore report, for each task family, three diagnostic conditions in addition to the neutral baseline: `sdb_job` (job-interview framing), `sdb_clinical` (confidential counselling framing), and `longctx` (a multi-turn counselling context prepended to the assessment). Following this dichotomy, we read stability on personality tasks and longctx-responsiveness on risk tasks as the two primary desiderata, and we quantify directional bias using the *Directional Consistency Rate* (DCR), i.e., the fraction of item-level shifts that move in the framing's intended direction; a value near 0.5 indicates idiosyncratic (non-systematic) drift, whereas values substantially above 0.5 indicate systematic SDB-style bias.

5.2 Baselines

We compare GenPT against self-report questionnaire baselines, which represent the standard approach for psychological assessment. For each task, we select established psychometric instruments:

Personality Assessment: (1) *MBTI Questionnaire*: 16-item forced-choice inventory yielding the four binary MBTI axes (E/I, S/N, T/F, J/P), which we evaluate directly against the ground-truth 4-letter type. (2) *Big Five Inventory (BFI)*: 44-item

Likert scale measuring Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism. Raw scores (1-5 per item) are averaged per dimension and rounded to 5 discrete levels.

Mental Health Assessment: (3) *Beck Depression Inventory (BDI-II)*: 21-item self-report measuring depression severity (score range 0–63). We map BDI scores to 4 depression risk levels (0–3). (4) *Beck Scale for Suicide Ideation (BSS)*: 21-item scale measuring suicide risk. We map BSS scores to 4 risk levels (0–3).

5.3 Reliability Experiments

We operationalise reliability via two stability-under-perturbation indicators computed against a neutral baseline: pooled Cohen's κ (does the instrument give the *same* ordinal answer when context changes?) and a Directional Consistency Rate (DCR, fraction of item-level shifts that move in the condition's intended direction). A high-κ yet high-DCR cell indicates *consistent but systematically biased*; DCR near 0.5 indicates no directional contamination.

5.3.1 Social Desirability Resistance

We compare three prompt conditions: *neutral*; *sdb_job* ("job interview…show your best self"), encouraging upward presentation; and *sdb_clinical* ("confidential counselling room…answer honestly"), which in human studies invites downward distressed disclosure.

		sdb_job		sdb_clinical	
Task	**Method**	κ	DCR	κ	DCR
Big Five	Questionnaire	0.84	0.60↑	0.85	0.61↑
	GenPT (Qwen3-8B)	0.63	0.52↑	0.58	0.50=
	GenPT (Phi-4-mini)	0.23	0.53↑	0.34	0.56↑
	GenPT (Intern-S1)	0.33	0.55↑	0.44	0.56↑
MBTI	Questionnaire	0.75	0.62↑	0.71	0.55↑
	GenPT (Qwen3-8B)	0.28	0.55↓	0.40	0.50=
	GenPT (Phi-4-mini)	0.21	0.70↓	0.21	0.52↓
	GenPT (Intern-S1)	0.26	0.55↓	0.20	0.52↑
Depression	Questionnaire	0.77	0.52↓	0.76	0.51↓
	GenPT (Qwen3-8B)	−0.17	0.55↓	−0.08	0.50=
	GenPT (Phi-4-mini)	−0.07	0.62↓	−0.20	0.60↓
	GenPT (Intern-S1)	−0.20	0.82↓	−0.15	0.57↑
Suicide	Questionnaire	0.67	0.71↓	0.79	0.88↓
	GenPT (Qwen3-8B)	−0.11	0.56↓	0.05	0.60↑
	GenPT (Phi-4-mini)	0.01	0.50=	−0.07	0.60↓
	GenPT (Intern-S1)	−0.42	0.55↑	−0.05	0.78↑

Table 1: Social-desirability resistance under two framings, for the self-report questionnaire baseline and for GenPT instantiated with each of three backbones. κ: pooled linearly-weighted agreement with the neutral baseline. DCR: fraction of item-level shifts in the framing's intended direction; ≈ 0.5 = idiosyncratic, $\gg 0.5$ = systematic bias. Arrows indicate the drift direction that the DCR majority takes.

Two patterns emerge from Table 1. First, on the highest-stakes item set, suicide ideation, the questionnaire baseline shows a pronounced, systematic drift toward the "healthy" direction: DCR reaches 0.71 under `sdb_job` and climbs to 0.88 under `sdb_clinical` (both downward), a textbook fake-good signature. On depression the same direction is visible but much milder (DCR ≈ 0.52, barely above chance). On trait tasks the questionnaire is not neutral either: Big Five and MBTI DCRs sit at 0.60–0.62 (upward) under both framings, consistent with the mild self-enhancement pattern that Bhandari et al. (2025) observed to lift agreeableness- and conscientiousness-like items. Second, none of the three GenPT backbones reproduces the questionnaire's directional bias on the risk tasks: across Qwen3-8B, Phi-4-mini and Intern-S1, no (backbone, risk task, framing) cell shows the simultaneous fake-good signature (DCR clearly above 0.5 in the "healthy" direction while κ stays high) that the questionnaire exhibits on suicide ideation. The three backbones do differ in how noisy their shifts are (Intern-S1 has the most pronounced idiosyncratic suicide drift under `sdb_clinical`, DCR 0.78 upward; Phi-4-mini is the quietest on suicide, DCR ≤ 0.60 in both framings), but the absence of a *systematic* fake-good bias is shared across the three families. On trait tasks, all three backbones drift mildly upward like the questionnaire, but with κ in the 0.2–0.6 range rather than 0.7–0.9, reflecting GenPT's higher per-item variance under prompt perturbation. Pooled κ for GenPT on risk tasks is low and sometimes negative, reflecting the single-label-per-persona sample size for the risk split rather than directional bias; the DCR-based interpretation is therefore the primary signal for those cells. Overall, the social-desirability advantage of GenPT over self-report is largest precisely where the stakes are highest—on suicide ideation, where the questionnaire's fake-good signature is strongest—and this advantage is robust across backbone families.

5.3.2 Longitudinal Context Responsiveness

For state-like tasks, the absence of drift is only half the story: a good instrument must *also* respond to context that genuinely changes the underlying state. To probe this, we prepend a multi-turn counselling context (`longctx`) to the assessment in which a sympathetic counsellor walks the persona through reframing, coping, and support-mobilisation; a clinically plausible outcome is a downward shift in

depression and suicide-ideation indicators. We report the mean per-persona shift μ_Δ relative to the neutral baseline and the pooled κ.

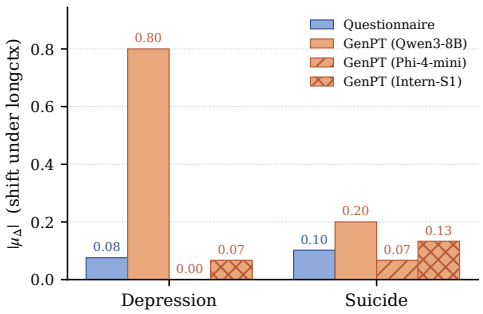


Figure 4: Longitudinal context responsiveness. Bars show the absolute mean per-persona shift $|\mu_\Delta|$ under a multi-turn counselling context relative to the neutral baseline, for the questionnaire baseline and for GenPT instantiated with each of three backbones.

Figure 4 shows that the direction of the questionnaire–GenPT contrast depends on the backbone. Under Qwen3-8B, GenPT shifts depression by $|\mu_\Delta| = 0.80$ and suicide by 0.20, versus the questionnaire's 0.08 and 0.10: an order-of-magnitude-larger response on depression and a two-fold response on suicide. Under the two smaller backbones the responsiveness is closer to the questionnaire—Intern-S1 shifts by 0.07 (depression) and 0.13 (suicide), and Phi-4-mini by 0.00 and 0.07—so context responsiveness does not come for free from the pipeline design alone but depends on the Interpreter/Diagnostician's capacity to track the narrative content of the counselling trajectory. Read jointly with Table 1, the Qwen3-8B configuration combines high longctx responsiveness with the absence of a systematic fake-good signature, i.e., it shifts under *content* change but not under *framing* change; Phi-4-mini and Intern-S1 are on the more conservative end of this spectrum, with smaller shifts under either perturbation. We do *not* run longctx on personality tasks, as a brief counselling context is not expected to shift dispositional traits.

5.4 Validity Experiments

5.4.1 Criterion Validity

We assess the consistency between predictions and ground-truth dataset annotations. For Big Five, Depression Risk, and Suicide Risk, all three are discrete ordinal labels (5-point for Big Five, 4-point for the two risk tasks), and we measure **accuracy** as the proportion of exact matches between predicted and ground-truth levels. For MBTI, the label is a 4-letter type whose four axes are independently meaningful, so we report **per-type Hamming distance**

$$\mathrm{HamD}_{MBTI} = \sum_{d=1}^{4} \mathbf{1}[\hat{y}_d \neq y_d^*] \in \{0,1,2,3,4\}, \tag{12}$$

and average it across personas as the MBTI error metric (lower is better). This is symmetric across the four axes, does not require converting letters to a continuous scale, and is directly comparable across methods.

Method	BF ↑	MBTI (HamD) ↓	Dep. ↑	Sui. ↑
Questionnaire	**0.373**	**0.733**	0.133	0.200
GenPT (Qwen3-8B)	0.333	1.200	0.200	**0.400**
GenPT (Phi-4-mini)	0.240	2.200	**0.400**	0.267
GenPT (Intern-S1)	0.293	1.667	**0.400**	0.067

Table 2: Criterion validity against ground-truth labels on the 15 CharacterRAG + 15 AnnaAgent validity subset. Big Five / Depression / Suicide: exact-match accuracy (↑); MBTI: mean 4-axis Hamming distance (↓, range 0–4).

Table 2 shows two patterns. On personality tasks with clean persona descriptions, the questionnaire baseline retains an edge across all three GenPT backbones (Big Five accuracy 0.37 vs. GenPT 0.24–0.33; MBTI Hamming distance 0.73 vs. GenPT 1.20–2.20), because item wording overlaps the persona text and can be mapped to trait-level labels with minimal reasoning. On mental-health risk tasks the comparison inverts and the effect is much larger: on depression, all three GenPT backbones score 0.20–0.40 versus the questionnaire's 0.13; on suicide, Qwen3-8B reaches 0.40 versus the questionnaire's 0.20. These gaps are consistent with a projective chain that aggregates narrative and affective indicators not directly negotiable from a single self-report item. The three backbones occupy different points in this trade-off: Qwen3-8B is the most consistent across tasks (second-best on Big Five, lowest MBTI Hamming distance among the three backbones, and best on suicide), while Phi-4-mini and Intern-S1 trade personality accuracy for stronger depression accuracy. Across all four tasks, the GenPT–questionnaire gap on the risk split is wider than the spread across the three backbones, indicating that the validity advantage of projective assessment on mental-health risks is not an artefact of a particular backbone choice. Conversely, on personality tasks the three-backbone range sits at or below the questionnaire

baseline, reinforcing the view that projective and self-report assessment are complementary: GenPT is the method of choice when the target construct is affect-laden and narrative-dependent, and questionnaires remain competitive when the target is a stable trait that can be mapped from persona text with minimal inference.

6 Conclusion

We presented GenPT, a generative projective-testing framework for persona-conditioned LLM agents, with a three-stage pipeline (Examinee / Interpreter / Diagnostician) grounded in standardised clinical scoring and newly generated stimuli to mitigate contamination. Our task-differentiated analysis shows that on personality-trait tasks GenPT and questionnaires are broadly comparable, with questionnaires retaining an edge when personas lexically expose trait labels; on mental-health risk tasks GenPT does not reproduce the fake-good signature that the self-report questionnaire exhibits most strongly on suicide ideation, a property that holds across all three Interpreter/Diagnostician backbones we evaluated. Longitudinal-context responsiveness on depression and suicide is more backbone-dependent: it is an order of magnitude above the questionnaire under one of the three backbones and closer to the questionnaire under the other two, indicating that the projective pipeline provides the structural conditions for context sensitivity but does not, by itself, guarantee it at every scale. We therefore position projective testing as a complementary psychometric tool when contamination resistance and bias asymmetry are the primary desiderata, with content-driven responsiveness as an additional desideratum that scales with Interpreter/Diagnostician capacity.

Limitations

Backbone and cultural coverage. Our evaluation spans three Interpreter/Diagnostician backbones (Qwen3-8B, Phi-4-mini-reasoning at 3.84B, and Intern-S1-mini at $\sim$8B), all in the small-to-mid open-weights range; coverage of substantially larger or architecturally different families (e.g., Gemma-3, Llama-3-70B, GPT-OSS) is left to future work. The psychological constructs measured in our experiments, while well-established in clinical psychology, may manifest differently across diverse populations and application scenarios. Future work should explore broader model families and multilingual settings to validate cross-cultural applicability.

Computational cost. Projective testing requires more computational resources than direct questionnaires due to multi-turn interactions, multi-dimension Interpreter calls, and the Diagnostician aggregation step. This increased complexity, while beneficial for assessment depth, may pose challenges for real-time or resource-constrained applications. The trade-off between assessment depth and computational efficiency remains an important consideration for practical deployment.

Ethics Considerations

This work involves psychological assessment of LLM-simulated agents, which raises several ethical considerations. First, while our framework assesses simulated personas rather than real individuals, the methodology could potentially be misused to infer psychological characteristics without consent. We emphasize that GenPT is designed for research purposes in understanding LLM behavior, not for evaluating human users.

Second, the mental health assessment dimensions (depression and suicide risk) require careful handling. Our experiments use synthetic personas from existing research datasets, and all stimuli were reviewed by psychology experts to ensure appropriateness. We do not recommend deploying such assessments in clinical settings without proper validation and professional oversight.

Finally, we acknowledge that psychological profiling of AI systems carries dual-use risks. While understanding LLM psychological characteristics supports safety and alignment research, the same techniques could potentially be exploited for manipulation. We encourage the research community to develop appropriate guidelines for the responsible use of LLM psychometric tools.

Use of AI Statement

We acknowledge the use of artificial intelligence tools in the preparation of this work. Specifically, Gemini was utilized for paper polishing to improve the clarity and flow of the manuscript. Additionally, GitHub Copilot and Antigravity were employed as coding assistants to support the implementation of the GenPT framework and experimental scripts. All AI-generated suggestions and code were rigorously reviewed and verified by the authors.

Acknowledgments

The work is supported by the National Natural Science Foundation of China (62272092, 62172086) and the Fundamental Research Funds for the Central Universities under Grant (N25XOD004). Furthermore, we would also like to thank the KinaMind society for their inspiring environment and unwavering support.

References

Pranav Bhandari, Usman Naseem, Amitava Datta, Nicolas Fay, and Mehwish Nasim. 2025. Evaluating personality traits in large language models: Insights from psychological questionnaires. In *Companion Proceedings of the ACM on Web Conference 2025, WWW 2025, Sydney, NSW, Australia, 28 April 2025 - 2 May 2025*, pages 868–872.

Black Forest Labs. 2025. FLUX.1-dev.

Xin Chan, Xiaoyang Wang, Dian Yu, Haitao Mi, and Dong Yu. 2024. Scaling synthetic data creation with 1,000,000,000 personas. *CoRR*, abs/2406.20094.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, and 1 others. 2024. Scaling rectified flow transformers for high-resolution image synthesis. *Preprint*, arXiv:2403.03206.

John E. Exner Jr. 1993a. *The Rorschach: A comprehensive system*. Wiley series in personality processes. Place: Oxford, England Publisher: John Wiley & Sons.

John E. Exner Jr. 1993b. *The Rorschach: A comprehensive system: Basic foundations*. The Rorschach: A comprehensive system: Basic foundations, Vol. 1, 3rd ed. John Wiley & Sons, Oxford, England. Pages: xxiii, 642.

Aaron Fanous, Jacob Goldberg, Ank A. Agarwal, Joanna Lin, and 1 others. 2025. Syceval: Evaluating llm sycophancy. *Preprint*, arXiv:2502.08177.

Lawrence K. Frank. 1939. Projective methods for the study of personality. *The Journal of Psychology*, 8(2):389–413.

Team Gemini, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, and 1 others. 2025. Gemini: A family of highly capable multimodal models. *Preprint*, arXiv:2312.11805.

Shahriar Golchin and Mihai Surdeanu. 2024. Time travel in LLMs: Tracing data contamination in large language models. In *The Twelfth International Conference on Learning Representations*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, and 1 others. 2024. The Llama 3 Herd of Models. *arXiv e-prints*, arXiv:2407.21783.

Juhye Ha, Hyeon Jeon, DaEun Han, Jinwook Seo, and Changhoon Oh. 2024. Clochat: Understanding how people customize, interact, and experience personas in large language models. In *Proceedings of the CHI Conference on Human Factors in Computing Systems, CHI 2024, Honolulu, HI, USA, May 11-16, 2024*, pages 305:1–305:24. ACM.

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

Leon O. H. Kroczek, Alexander May, Selina Hettenkofer, Andreas Ruider, Bernd Ludwig, and Andreas Mühlberger. 2025. The influence of persona and conversational task on social interactions with a llm-controlled embodied conversational agent. *Comput. Hum. Behav.*, 172:108759.

Kun Li, Chenwei Dai, Wei Zhou, and Songlin Hu. 2025. APEE: assessing the personality expressions of llm-driven role play agent beyond self-perception. In *28th International Conference on Computer Supported Cooperative Work in Design, CSCWD 2025, Compiegne, France, May 5-7, 2025*, pages 1406–1411.

CHRISTIANA D. MORGAN and HENRY A. MURRAY. 1935. A METHOD FOR INVESTIGATING FANTASIES: THE THEMATIC APPERCEPTION TEST. *Archives of Neurology & Psychiatry*, 34(2):289–306.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, and 1 others. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Jeiyoon Park, Yongshin Han, Minseop Kim, and Kisu Yang. 2025. Dynamic context adaptation for consistent role-playing agents with retrieval-augmented generations. *CoRR*, abs/2508.02016.

PDB Community. 2022. PDB: The Personality Database.

Changyong Qi, Longwei Zheng, Anna He, Haoxin Xu, Linzhao Jia, Yuang Wei, Bingqian Jiang, and Xiaoqing Gu. 2026. Simulating student learning behaviors with llm-based role-playing agents: A data-driven and cognitively inspired framework. *Expert Systems with Applications*, 304:130753.

Hermann Rorschach. 1922. Psychodiagnostik. *The Journal of Nervous and Mental Disease*, 56(3).

Mustafa Safdari, Greg Serapio-García, Clément Crepy, Stephen Fitz, Peter Romero, Luning Sun, Marwa Abdulhai, Aleksandra Faust, and Maja J. Mataric. 2023. Personality traits in large language models. *CoRR*, abs/2307.00184.

George Stricker and Bede J Healey. 1990. Projective assessment of object relations: A review of the empirical literature. *Psychological Assessment: A Journal of Consulting and Clinical Psychology*, 2(3):219.

Ming Wang, Peidong Wang, Lin Wu, Xiaocui Yang, and 1 others. 2025a. Annaagent: Dynamic evolution agent system with multi-session memory for realistic seeker simulation. In *Findings of the Association for Computational Linguistics, ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 23221–23235. Association for Computational Linguistics.

Noah Wang, Zhongyuan Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhan Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Jian Yang, Man Zhang, Zhaoxiang Zhang, Wanli Ouyang, Ke Xu, Wenhao Huang, Jie Fu, and Junran Peng. 2024a. Rolellm: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 14743–14777. Association for Computational Linguistics.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, and 1 others. 2024b. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *Preprint*, arXiv:2409.12191.

Xintao Wang, Heng Wang, Yifei Zhang, Xinfeng Yuan, Rui Xu, Jen-tse Huang, Siyu Yuan, Haoran Guo, Jiangjie Chen, Shuchang Zhou, Wei Wang, and Yanghua Xiao. 2025b. Coser: Coordinating llm-based persona simulation of established roles. In *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*.

Xintao Wang, Yunze Xiao, Jen-tse Huang, Siyu Yuan, Rui Xu, Haoran Guo, Quan Tu, Yaying Fei, Ziang Leng, Wei Wang, Jiangjie Chen, Cheng Li, and Yanghua Xiao. 2024c. Incharacter: Evaluating personality fidelity in role-playing agents through psychological interviews. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 1840–1873.

Drew Westen. 1991. Social cognition and object relations. *Psychological Bulletin*, 109:429–455.

Binwei Yao, Chao Shi, Likai Zou, Lingfeng Dai, and 1 others. 2022. D4: a Chinese dialogue dataset for depression-diagnosis-oriented chat. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2438–2459, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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

Xi Zheng, Zhuoyang Li, Xinning Gui, and Yuhan Luo. 2025. Customizing emotional support: How do individuals construct and interact with llm-powered chatbots. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems, CHI 2025, Yokohama, Japan, 26 April 2025- 1 May 2025*, pages 376:1–376:20. ACM.

A Symbol Definition

A.1 Symbols of the GenPT Framework

To improve clarity and readability, we summarize the main mathematical symbols used throughout this paper in Table 3. These symbols pertain to the components of the proposed GenPT framework. Notably, $\mathcal{E}$ represents the explanation grounded in psychological theory, while $\mathbf{y}$ refers to the downstream task predictions such as personality traits or mental health risk level. The system involves the Examinee (target LLM under assessment), the Interpreter (standardized psychological analysis), and the Diagnostician (task-specific recognition).

A.2 Variables in Rorschach Test Analysis

In our implementation, the Examinee is asked to respond to each of the ten standard Rorschach inkblot cards. The generated responses are then analyzed by a structured agent $\mathcal{A}_{ror}$, which simulates the scoring rationale of established clinical systems such as the Exner Comprehensive System and R-PAS.

Specifically, we construct an explanation vector $\mathcal{E}^{(ro)} = \{F, DQ, R, P, C, M, COG, AFF\}$ for each subject, where each element represents a psychological feature derived from the response. Each component is described below:

- **Form Quality** (F): Measures how well the response matches the actual contours of the blot.
  - *Values*: `+`, `o`, `-`
  - *Interpretation*: A poor form quality (e.g., $F = -$) may indicate impaired reality testing or disorganized thought patterns.
- **Developmental Quality** (DQ): Describes the structural complexity and level of integration in the percept.
  - *Values*: `Synthesized (S)`, `Ordinary (D)`, `Vague (Dv)`
  - *Interpretation*: Vague or fragmented percepts may reflect cognitive immaturity, while synthesized percepts suggest integrative capacity.
- **Location** (R): Specifies which part of the blot was used in the response.

Symbol	Description	Series
$\mathcal{X}$	The Examinee, i.e., the LLM-based agent under assessment.	-
$\mathcal{P}$	The Persona profile that defines the Examinee's psychological ground truth (demographics, traits, mental health profiles).	-
$\mathcal{I}$	The Interpreter that analyzes projective test responses using standardized psychological frameworks.	Produces scores $\mathbf{s}_i$ and explanations $\mathcal{E}_i$ for each response.
$\mathcal{D}$	The Diagnostician that maps structured indicators to final psychological state predictions.	Produces $\hat{\mathbf{y}}_{BF}$, $\hat{\mathbf{y}}_{MBTI}$, $\hat{y}_{drisk}$, $\hat{y}_{srisk}$.
$\mathbf{T}$	Set of projective test stimuli (TAT images, Rorschach cards, SCT sentence stems).	Individual stimuli t_i; subsets $\mathbf{T}_{TAT}$, $\mathbf{T}_{Ror}$, $\mathbf{T}_{SCT}$.
$\mathbf{R}$	Set of Examinee responses to projective test stimuli.	Individual responses r_i; subsets $\mathbf{R}_{TAT}$, $\mathbf{R}_{Ror}$, $\mathbf{R}_{SCT}$.
$\mathbf{S}$	Set of structured scores from Interpreter analysis.	Task-specific subsets $\mathbf{S}_{TAT}$, $\mathbf{S}_{Ror}$, $\mathbf{S}_{SCT}$.
$\mathbf{E}$	Set of psychological explanations generated by the Interpreter.	Individual explanations $\mathcal{E}_i$ grounded in psychological theory.
$\hat{\mathbf{y}}$	Final psychological state predictions from the Diagnostician.	$\hat{\mathbf{y}}_{BF} \in \{1,\ldots,5\}^5$, $\hat{\mathbf{y}}_{MBTI} \in [0,1]^4$, $\hat{y}_{drisk} \in \{0,\ldots,4\}$, $\hat{y}_{srisk} \in \{0,\ldots,3\}$.

Table 3: Symbols appearing in the main body and their descriptions.

  - *Values*: `W` (whole blot), `D` (common detail), `Dd` (unusual detail), `S` (white space)
  - *Interpretation*: Frequent white space usage may reflect oppositionality or unconventional perception.

- **Popularity** (P): Indicates whether the response is among statistically common answers.
  - *Values*: `P` (popular), `idiosyncratic`
  - *Interpretation*: A high frequency of idiosyncratic responses may suggest social detachment or psychotic tendencies.

- **Color Determinants** (C): Whether the subject reacts to chromatic features.
  - *Values*: `FC`, `CF`, `pure C`
  - *Interpretation*: Dominance of pure color responses (C) may indicate impulsivity or emotional lability.

- **Human Movement** (M): Detection of imagined human or animal motion.
  - *Values*: `M`, `FM`, `m`
  - *Interpretation*: Symbolizes internal fantasy, empathy, and capacity for complex cognition. Absence may relate to emotional detachment.

- **Cognitive Special Scores** (COG): Captures peculiar logic, vague reasoning, or cognitive slippage.
  - *Values*: `INCOM`, `FABCOM`, `CONTAM`, etc.
  - *Interpretation*: Presence of cognitive distortion codes such as contamination or illogic are predictive of thought disorder.

- **Affective Features** (AFF): Reflects emotional tone and regulation capacity.
  - *Values*: `morbid content`, `aggressive imagery`, `anxiety indicators`
  - *Interpretation*: Elevated negative content may signal depressive tendencies, trauma exposure, or latent aggression.

All these dimensions are automatically extracted through structured prompts and parsing rules from the response text. These are used both as an intermediate psychological representation and a foundation for downstream inference tasks such as emotional dysregulation classification or self-harm risk assessment.

B Questions and Prompts

B.1 Questions for Data Contamination Test

We design a series of probes in different dimensions to see if the model's responses go beyond what would be expected from "a generalized model of an intelligent person who has never seen this picture". If the model's response contains information that could only be known from exposure to the image and its associated "standard interpretations" or metadata, then there is a high probability that it has been trained.

- **Direct Identification**: whether the model can directly identify the origin or name of the test materials. Here are a few sample questions:
  - Have you seen this picture?
  - Is this picture from some famous psychology test?
  - What card number is this image from the Thematic Apperception Test (TAT)?
- **Canonical Interpretation**: whether the model recalls standard or canonical interpretations associated with the original task. Here are a few sample questions:
  - Please tell a complete story based on this picture.
  - What happened in the story?
  - What has happened in the past and what will happen in the future?
- **Metadata and Contextual Probing**: whether the model reveals knowledge of test construction logic, psychometric background, or scoring systems. Here are a few sample questions:
  - What is the approximate date this image was created?
  - Which psychological trait is this card primarily used to assess in an individual?
  - Besides this one, are there any other images in the series? What do they have in common?

The responses and behaviors of MLLMs correspond to whether or not they can be shown to have been trained as shown in Table 4.

Tier	Strong evidence	Weak evidence
Direct Identification	Just say "This is TAT card X" or mention Henry Murray.	Denial of recognition of pictures, generic descriptions only.
Canonical Interpretation	The story told is highly consistent with the card's classic psychological interpretation.	The stories told are random, varied, and have no obvious connection to the classic readings.
Metadata and Contextual Probing	Be able to name non-visual information such as the date, purpose, and series name of a picture.	Unable to answer metadata questions or make reasonable but unsupported guesses based on vision.

Table 4: Evidence of data contamination.

B.2 Prompts for Stimuli Generation

TAT Responding to the demand, we design the following style prompts:

> *monochromatic, subtle chiaroscuro lighting, soft focus, impressionistic, figures with indistinct or obscured facial features, details are suggestive rather than explicit, muted tonal range, consistent fine line weight, low color saturation if not monochromatic.*

and negative prompts:

> *brightly colored, sharp details, clear facial expressions, modern technology, specific cultural symbols, text, logos, overt emotion, brand names, and identifiable locations.*

For specific scenarios and content, we designed prompts and some examples are shown following:

- *Two figures standing a short distance apart in a dimly lit, featureless room. One figure is slightly turned away.*
- *Several indistinct human forms gathered around a barely discernible object on a flat surface, in an outdoor setting with a low horizon.*
- *One figure seated, another standing nearby, looking towards a hazy opening or window in a sparsely furnished space.*
- *Two figures, their forms partially overlapping, in an undefined space*

with ambiguous architectural elements in the background.

- *A group of figures huddled together, their attention seemingly focused on something outside the lower edge of the frame.*

Rorschach test Correspondingly, we have also designed prompts for generating inked images and some examples are shown following:

- *The overall form (W) is cohesive but highly ambiguous, with contours that gently suggest a large, winged creature like a bat or moth (A).*
- *The overall form (W) is ambiguous but suggests a ceremonial mask or a tribal headdress (H).*
- *The overall form (W) is ambiguous but contains shapes that could be interpreted as an anatomical diagram, like a pelvis or a chest x-ray (An, Xy).*
- *The overall form (W) is ambiguous, suggesting a large beetle or insect with its wings spread (A).*
- *The overall form (W) is ambiguous, with shapes that hint at a coat of arms or an emblem.*

Correspondingly, we designed the style prompt:

> *A psychometrically precise, bilaterally symmetrical Rorschach inkblot on a stark white background. Monochromatic black ink with subtle grey shading variations creating a sense of diffuse light and shadow (Y). Style reminiscent of Hermann Rorschach's original Psychodiagnostics plates. –style raw –ar 3:4*

and negative prompt:

> –no letters, no symbols, no flags

The complete prompts and code are available at https://github.com/sci-m-wang/GenPT.

B.3 Implementation of Projective Tests

To simulate deep psychological probing, we designed and implemented a set of classic projective tasks adapted for large language models (LLMs) acting as Examinees. The projective assessment consists of three components: Thematic Apperception Test (TAT), Rorschach Inkblot Test, and Sentence Completion Test (SCT). The execution flow and proportions are illustrated as follows:

- **TAT**: Each Examinee was prompted to complete 8 picture-based storytelling tasks. The images were drawn from three thematic categories in a fixed 4:3:1 ratio (social interaction, conflict & trauma, moral/identity themes). For each image, the Examinee was asked to narrate a story that reflects the scene, inner thoughts, emotions, and outcome.
- **Rorschach**: The model was shown 10 standard inkblot cards and asked to describe what it sees in each image. The responses focused on perceptual structure, thematic associations, and emotional tone.
- **RISB**: The model completed 20 sentence stems from diverse categories (family, self, social, future, etc.). Only psychological experts annotated the RISB responses using structured criteria.

B.4 SCORS-G Analysis Prompts

All generated responses were subsequently processed by probe analyzers. In the case of TAT, the SCORS-G framework was employed to produce ratings on eight core dimensions of social-cognitive and self-representational functioning. Each dimension was analyzed independently using specialized prompts grounded in psychodynamic theory.

Below is an example prompt used for the `COM` (Complexity of Representations of People) dimension:

> *You are an expert clinical psychologist well-versed in the SCORS-G scoring system for TAT. Your task is to rate the narrative provided by the subject on the dimension* ***Complexity of Representations of People (COM)****. The score should range from 1 to 7. A score of 1 indicates extremely simplistic or chaotic representations; 3 indicates stereotypical or black-and-white portrayals; 5 indicates a bal-*

anced view integrating good and bad aspects; 7 reflects complex and psychologically insightful representations. Follow the criteria strictly, assign a score, and provide detailed reasoning in a chain-of-thought format. Cite specific words or phrases from the narrative as evidence. Output your response in JSON format.

Similar prompts were designed for each of the eight SCORS-G dimensions: `COM`, `AFF`, `EIR`, `EIM`, `SC`, `AGG`, `SE`, and `ICS`. These prompts enabled fine-grained scoring with psychological interpretability, forming the explanation vector $\mathcal{E}$ for subsequent state recognition and analysis.

B.5 Examinee Profile Design

To construct diverse and psychologically realistic Examinees, we utilize two complementary profile sources as described in Section 3.

AnnaAgent Profiles AnnaAgent (Wang et al., 2025a) provides mental health profiles with structured psychological attributes. Each profile includes:

- **Demographic information**: Gender, age, occupation, and marital status.
- **Psychological situation**: Current mental health context and presenting concerns.
- **Speaking characteristics**: Language patterns, vocabulary level, and communication style.
- **Risk indicators**: Depression and suicide risk levels (ground truth labels).

These profiles enable evaluation on clinical mental health tasks with established ground truth.

CharacterRAG Profiles CharacterRAG (Park et al., 2025) provides detailed fictional character profiles. Each profile is structured with:

- **Beliefs and Values**: Core values, priorities, and worldview.
- **Psychological Traits**: Personality characteristics, emotional patterns, and behavioral tendencies.
- **Speech Style**: Distinctive verbal patterns, catchphrases, and communication preferences.

We use 15 fictional characters whose personality traits are documented in the Personality Database (PDB Community, 2022), enabling evaluation on Big Five and MBTI prediction tasks.

Persona Sampling for Experiments For the validity experiments (Table 2), we randomly sample 15 AnnaAgent personas from the D4 pool with a fixed seed of 1 in order to keep the sample size comparable with the CharacterRAG pool of 15 characters; all 15 CharacterRAG characters are used. For the reliability experiments (Table 1, Figure 4), we reuse a separately fixed set of 15 AnnaAgent personas for which pre-generated baseline, `sdb_job`, `sdb_clinical`, and `longctx` behaviours were available at evaluation time; the same 15 CharacterRAG characters are used across all reliability conditions. The two AnnaAgent subsets partially overlap but are not identical; both are drawn from the same underlying D4 pool and are used in a within-persona (paired) design within each experiment, so persona identity is held constant between the conditions being compared.

Profile Integration For each assessment, the Examinee's profile is integrated into the system prompt:

> *You are {character_name}. Based on the following psychological profile, respond to the projective test stimuli in character.*
>
> ***Beliefs and Values:*** *{beliefs}*
>
> ***Psychological Traits:*** *{traits}*
>
> ***Speech Style:*** *{speech_patterns}*
>
> *Stay in character throughout the assessment.*

This structured approach ensures consistent persona embodiment across all projective tests.

C Implementation Details

C.1 Verification of Data Contamination

Considering that classical projection tests are likely to be used for training LLMs or MLLMs, this can lead to data contamination issues. To test this conjecture, we design three-tier questions to test whether the MLLMs are trained on traditional projective tests. The questions can be found in Section B.1. We collect responses from several MLLMs on different tiers of questions. If the model's response shows weak evidence, it is recorded as a pass. The pass rates are shown in

Table 5. The results indicate that the existing foun-

Model	Tier 1	Tier 2	Tier 3
GPT	21.03	18.41	35.47
Gemini	10.24	46.86	11.01
Qwen-VL	92.54	62.31	25.92
Llama	40.24	40.54	6.84

Table 5: Contamination testing results across selected MLLMs. The tested versions of MLLMs are GPT-4o (OpenAI et al., 2024), Gemini-2.5-Pro (Gemini et al., 2025), Qwen2.5-VL-7B-Instruct (Wang et al., 2024b) and Llama-3.2-11B-Vision-Instruct (Grattafiori et al., 2024).

dation MLLMs are most likely trained on the stimuli of traditional projective tests. Therefore, it is necessary to create new stimuli for testing.

C.2 Annotation Details

To support evaluation of the Interpreter's intermediate outputs, we constructed a web-based annotation interface and invited human experts to provide reference annotations for the probe tasks. The annotators include five graduate students majoring in Fine Arts (serving as artistic experts), one licensed psychological counselor, and three graduate students in psychology (serving as psychological experts). Figure 5 shows a screenshot of the annotation interface.

For the **TAT** and **Rorschach** tasks, two rounds of annotation were conducted. In the first round, artistic experts were asked to annotate the responses with a focus on narrative elements, visual metaphors, and affective expressions. In the second round, psychological experts evaluated the same responses using established scoring systems—SCORS-G for TAT and R-PAS-inspired criteria for Rorschach—to provide psychologically grounded reference labels.

For the **RISB** task, only psychological experts participated. They annotated the sentence completions using a structured rubric adapted from the standard RISB manual, assessing indicators of psychological distress, conflict, and emotional expression.

All annotations were collected via the same interface, and disagreements (if any) were discussed in post-annotation sessions. The resulting labels serve as the gold standard for evaluating GenPT's interpretability and accuracy across multiple psychological dimensions.

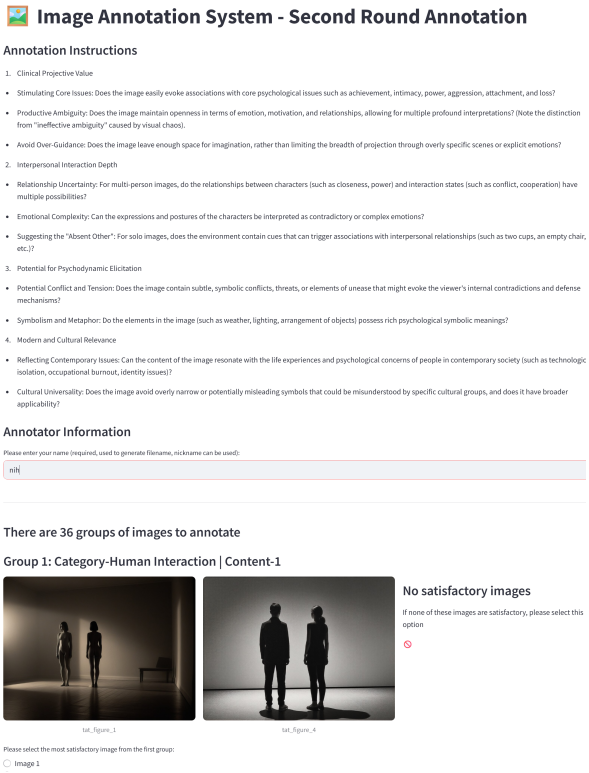


Figure 5: Screenshot of the web-based annotation interface used by psychological and artistic experts.

D Ethical Concerns

This study focuses on developing and evaluating AI technologies for high-stakes mental health applications. Consequently, we place a high priority on the associated ethical considerations. We have implemented several measures to ensure the rigor and ethical compliance of our research process.

First, during the research design phase, we thoroughly considered the inherent risks associated with LLMs, such as sycophancy and hallucination. To mitigate the risk of model bias and response distortion arising from "data contamination", where classical psychometric instruments are likely part of the model's training data, we avoided the direct use of classic projective test cards. Instead, we constructed a completely new set of projective test stimuli (including Thematic Apperception Test and Rorschach images) specifically for this study.

Most critically, we organized a team of licensed psychologists and art experts to conduct a rigorous, multi-round review and screening of these newly constructed stimuli. The review process focused on the following three aspects:

- **Image Content and Ethics**: Ensuring that the image content was free of any elements that

could be offensive, discriminatory, or evoke inappropriate associations.

- **Psychological Meaning**: Evaluating the psychological significance of each image and its potential to elicit deep narratives, thereby ensuring its effectiveness as a psychodynamic probe.
- **Avoiding Data Contamination**: Verifying that all new stimuli were original to prevent rote, memorized responses from existing LLMs that may have been trained on the original, widely-known tests.

Furthermore, in handling character profile data from AnnaAgent (Wang et al., 2025a) and CharacterRAG (Park et al., 2025), along with personality labels from the Personality Database (PDB Community, 2022), we strictly adhered to data anonymization principles to protect individual privacy. All data usage and research activities were conducted with the aim of advancing mental health services.

Finally, this research is committed to providing an open and transparent benchmark and resources for the evaluation of Examinees. The new stimuli, datasets, and related resources developed in this study are released at `https://github.com/sci-m-wang/GenPT` under a license that permits free academic use, to foster further research and development within the community.