PsyDI: A Personalized and In-depth Agent for Psychological Measurements with Self-reported Data Refinement

Anonymous ACL submission

Abstract

Psychological assessment should be both personal and accessible, yet scales are generic and expert interviews are costly. At population scale, systems rely on self-reported datasets whose labels are noisy and weakly tied to linguistic semantics. We present PsyDI, a data-centric pipeline that treats LLM outputs as weak supervision to refine these labels. Items are augmented into pairwise comparisons. A ranking-loss score model trained on the refined targets predicts trait orderings rather than raw scores, and a modular, multi-turn agent uses an evolving trait profile and multimodal inputs to ask targeted questions and reduce uncertainty. On MBTI benchmarks, PsyDI improves trait ordering and downstream accuracy; a public deployment with 10,000+ users yields an annotated dialogue dataset. This advances mixed-methods understanding by linking quantitative refinement with conversational evidence at scale.

1 Introduction

Psychological assessment should be both personal and accessible. Standardized questionnaires efficiently reach large populations, but they can hardly adapt to an individual's evolving context and thus struggle to capture nuanced traits (Cohen et al., 1996). Professional interviews and counseling deliver depth and customization, but are costly, scarce, and time-consuming for the general public. At population scale, an assessment pipeline must minimize marginal cost, avoid reliance on scarce experts, and yield labels immediately usable for model supervision. Among available signals, *self-reported datasets* meet these constraints: large volumes of user-provided labels and statements can be collected asynchronously, standardized at low cost, and integrated into automated pipelines. The trade-off is quality—such labels are noisy and only loosely coupled to the linguistic semantics in natural language.

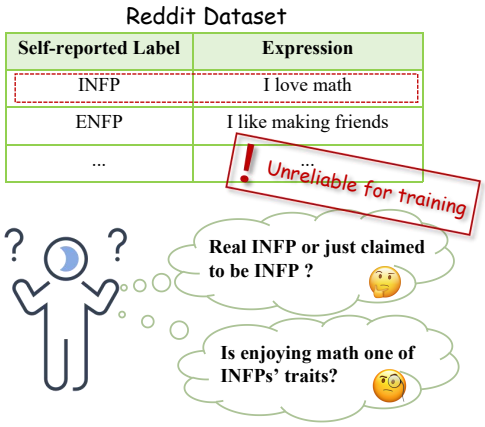


Figure 1: Self-reported psychological datasets face two key challenges: 1) Label Accuracy 2) Weak link between labels and expressions.

Large Language Models (LLMs) enable a promising mode of interaction: with broad linguistic prior knowledge and goal-directed prompting, agents can personalize dialogue, extract latent information, and reason with domain knowledge injected via prompts or fine-tuning (Toleubay et al., 2023; Buechel et al., 2018; Cheng et al., 2023; Tu et al., 2024; Yang et al., 2024). However, when these agents are trained on self-reported data, they inherit the same unreliability of the data, as shown in Figure 1: (1) self-reported types can be wrong because people misjudge themselves or answer strategically; (2) the connection from a declared type to the expressions in free responses is unclear and unstable, since similar texts may receive different labels. Under this kind of supervision, the LLM tends to relies on surface cues like topic or style, becomes over-confident on biased subsets, and produces predictions that do not transfer well across user groups. A comprehensive review of relevant literature on LLM personalization and the challenges of unreliable supervision is provided in Appendix B.

Our goal is a personalized, scalable assessment

that remains faithful to the semantics of user expressions. We take a data-centric stance: rather than increasing model complexity, we prioritize **data refinement** so models do not learn from unreliable targets. We reframes trait estimation as ordering instead of scoring and operationalizes data refinement with an LLM: the LLM's classification serves as **weak supervision** over statement pairs. we treat an item as more indicative only when **the LLM's preference aligns with the original label**. Consequently, we reformulates statements as pairwise comparisons, yielding refined targets that better align with the semantic evidence in expressions.

On top of these refined targets, we train a **ranking-based score model** that maps statements to per-dimension trait scores based on Myers–Briggs Type Indicator (MBTI), a popular personality assessment. Concretely, the model uses four heads (one per MBTI axis) and a margin-based ranking objective to favor statements that the refined order places higher for a given trait. At inference, the score model provides interpretable scores that drive downstream decisions while preserving the order established by data refinement.

Finally, we embed the score model into an adaptive, multi-modal agent *PsyDI* that actively reduces uncertainty in user's expression. The system maintains an evolving trait Profile, a quantitative trace of current trait leanings, and chooses targeted follow-ups on the most informative statements to resolve ambiguity. This preserves the efficiency of scales while recovering the personalization of interviews.

We use MBTI as a high-coverage testbed due to abundant self-report data (Myers et al., 1962), despite its psychometric limits (Boyle, 1995; Furnham and Crump, 2005) (see Appendix A for detailed MBTI introduction). Treating inference as an ordering problem mitigates label noise by transforming weak local cues into global constraints via relative differences. This decoupling from numeric conventions allows the pipeline to generalize to other inventories like Big Five (Digman, 1997) and PANAS-X (Watson and Clark, 1994).

Empirically, across MBTI datasets, training on refined targets improves trait ordering and downstream classification over score-imitation baselines. In addition, in a public deployment with 10,000+ users, the system is feasible at scale and collects an annotated dialogue dataset for future work. While MBTI is not a clinical instrument, these results suggest that order-aware data refinement improves interpretability and performance in self-report–driven assessment and generalizes beyond MBTI. The contributions of this paper are summarized as follows:

- **Data-centric refinement.** We transform self-report items into pairwise comparisons, obtain order-consistent refined targets via LLM supervision with agreement filtering, and train a four-head ranking-based score model on these targets rather than raw scores.
- **PsyDI agent.** We design and deploy the *PsyDI* agent, which operationalizes comparative, uncertainty-driven assessment. This design turns static scales into a collaborative dialogue, supports user understanding and interaction, and offers a reusable pattern for accessible, personalized measurement for human understanding.
- **Evidence at scale and resources.** Experiments upon MBTI shows that training on refined targets improves both trait ordering and downstream classification over score-imitation baselines. And a public deployment with 10,000+ users demonstrates feasibility and collects an annotated dialogue dataset to support future research and transfer to broader constructs.

2 Self-reported Data Refinement and Ranking-based Score Model

Our objective is to transform weak and noisy self-report supervision into order-aware targets that can be learned reliably. We conceptualize the interaction between the agent and the user as a Markov Decision Process (MDP) to optimize the sequence of inquiries (see Appendix C for formal MDP definitions). We formulate trait estimation as an ordering problem rather than a scoring problem: supervision is defined through pairwise comparisons, where the task is to determine whether one expression is more indicative of a given trait pole than another. In what follows, we formalize the score model and the associated ranking-based loss in Section 2.2, and detail the MBTI-specific data construction procedure in Section 2.3 that ensures pairwise supervision remains both informative and interpretable in Figure 2.

2.1 Self-reported Data Refinement

LLM-based psychological assessment requires accurate predictions of users' psychological traits from their statements. This is inherently challenging because self-reported labels are not ground truth: users may misperceive their states, and the

mapping from their labels to the linguistic expressions they produce is often indirect or noisy. Taking MBTI as an illustrative case, prior research and datasets indicate that how strongly a statement evidences a given type/pole can be unmeasurable from raw labels alone due to linguistic ambiguity (Boyle, 1995; Furnham and Crump, 2005).

We treat the **agreement** between an LLM's per-statement classification and the original self-report as a **proxy variable** for trait salience. Concretely, for each statement the LLM is prompted to return a ranked list of the Top-4 MBTI types with probabilities. This ranking is then compared to the self-reported label (or its axis-consistent mapping) to produce an agreement score: statements where the labeled type ranks higher and with greater probability are considered more indicative of that trait. For any two statements, we deem the one with the higher agreement score as more indicative to build a series of data pairs. These refined, pair-wise targets supervise a **ranking-based score model** (four heads, one per axis), which learns to assign trait scores while preserving the ordering induced by data refinement.

Operationally, as shown in Figure 2, we construct pair candidates by sampling two statements that share the same self-reported label (e.g., an MBTI type); additionally, we create *meaning-reduced* pairs by masking the trailing clause of a statement to produce a less informative counterpart (left). For each statement, we query the LLM to obtain a ranked distribution over types and retain the Top-4 with probabilities (e.g., INTP(88%)>ISTP(76%)>INFP(71%)). We then calculate agreement between these Top-4 predictions and the self-reported label using a small set of deterministic comparison rules (middle), where green arrows mark matched labels. The statement with higher consistency is taken as more consistent with the label. This procedure scales to any MBTI-labeled corpus and, together with the masking strategy, enables the construction of extensive pairwise datasets for refinement.

2.2 Score Model

Despite the positive correlation between agreement scores and psychological traits, establishing a direct mapping remains challenging. This difficulty arises because the score is influenced not only by the psychological traits expressed in statements but also by the LLM's varying ability to interpret such statements. To address this challenge, we adopt a more flexible learning approach by focusing on **the ranking of the agreement score** rather than relying on absolute values.

Let $M = \{m_1, m_2, \ldots\}$ denote the set of MBTI types (each $m \in M$ is a combination of four binary dimensions), and let a user's statements be $\mathbf{s} = \{p_1, p_2, \ldots, p_n\}$. Our goal is to quantify the indicativenss of each statement p toward every $m \in M$. To this end we train a Score Model $\mathcal{F}_m(\cdot)$ that maps a statement p to scores aligned with type m. In practice, the score model is instantiated as an LLM with four heads to separately predict scores for the four MBTI dimensions (E/I, N/S, F/T, J/P). Given two statements we can form a pairwise sample $\{p_i, p_j, m\}$ where p_i is more indicative of type m than p_j.

For a statement p, the four-head outputs are linearly combined by the type vector m to obtain a scalar score. Following the original formulation, we write the scalar score used by the ranking objective as $-m^\top \cdot \sum \mathcal{F}_m(p_i)$ where the summation follows the implementation in our score head.

We enforce that the "more indicative" statement p_i receives a higher score than p_j under m, and introduce a margin to sharpen the decision boundary and improve robustness, as shown in the right side of Figure 2. The resulting loss function is:

$$L(p_i, p_j, m) = \max\Big(-m^\top \cdot \big(\sum \mathcal{F}_m(p_i) - \sum \mathcal{F}_m(p_j) \big) + \text{margin}, 0 \Big) \tag{1}$$

With Eq. 1, we can train a score model to rank the indicativeness of statements under specific MBTI. The discussion of different loss functions will be presented in Section G.2.

2.3 Robust Training Objectives

Self-reported datasets are frequently prone to reward misalignment, where the model captures superficial correlations—specifically sentence length—instead of psychological traits. In such cases, longer statements may be overrepresented as more indicative due to their higher information density. To decouple trait indicativeness from verbosity, we introduce two targeted augmentation strategies:

Controlled Semantic Mixing. To isolate signals from length cues, we synthesize statement pairs of equal length but varying indicativeness. For each instance, we retain 70% of the original content and replace the remaining 30% with randomly sampled

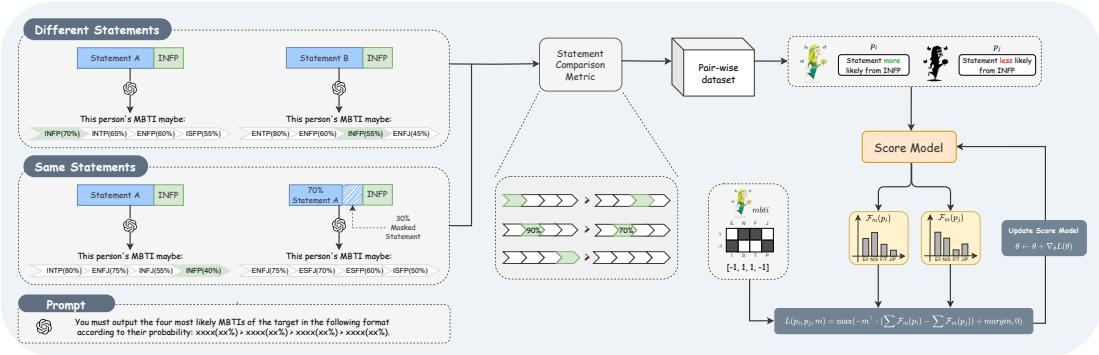


Figure 2: Self-reported data refinement and score model training pipeline. It begins by using ChatGPT to predict the probability of each statement's MBTI type. We then construct pairs of statements based on predictions and their true labels m. For each pair-wise statements (p_i, p_j, m), where the statement p_i is predicted to align more closely with the m than statement p_j, the loss function is defined as the difference in the predicted probabilities under the m between p_i and p_j. This ensures that the statement more accurately matching m receives a higher score.

segments from discordant MBTI types, ensuring length-invariant contrastive learning.

Invariant Length Inflation. To penalize length-based overconfidence, we construct samples where statement length is increased via verbatim repetition while the underlying psychological signal remains static.

The impact of these strategies is analyzed in Section G.2.

3 PsyDI Progressively In-depth Agent

This section presents the system and interaction design of *PsyDI*, built on the proposed data refinement method in Section 2 to enhance psychological assessments through conversation.

3.1 Pipeline Overview

The pipeline follows a short closed loop, shown in Figure 3: (i) **Exploration Phase.** Initialize an trait state from a small set of user-provided statements (including optional multi-modal anchors that are normalized to text); (ii) **Statement Selection.** Select the next statement to focus using the ambiguity-aware criterion (Section 4.3, Eq. (5)), introduced in Section 3.3; (iii) **Multi-turn Chat** Converse in a brief 3-5 turn module centered on that statement, introduced in Section 3.4; (iv) **Update Trait Profile.** Update Trait Profile introduced in Section 3.2 by summarizing the exchange for user confirmation, and either continuing or terminating when confidence is high or marginal gain is low.

While we instantiate this pipeline using MBTI as a primary example, the underlying framework is inherently generalizable to any psychometric inventory with discrete trait dimensions.

3.2 Trait Profile

Analogous to quantitative scoring in standardized scales, we maintain an initial estimate of a user's trait tendencies derived from their statements. In practice, we adopt a profile-style representation common in psychological measurement (Watson and Clark, 1994). This component documents the user's trait tendencies and guides other modules of PsyDI in choosing effective multi-turn dialogue strategies. Concretely, we define an **Trait Profile** that records per-trait scores, translating statements into comparable numeric indicators to ensure uniformity and cross-user comparability.

The profile state is driven by how **indicative** each statement is of each type. To quantify this indicativeness, we train a score model $\mathcal{F}_m(\cdot)$ that maps a statement to a trait-specific score for trait type m (see Section 2). We then convert scores to values by locating each statement's percentile within the empirical distribution of statements associated with type m:

$$\mathcal{V}_m(p) = \frac{\left|\left\{ p' \in P_m \mid \mathcal{F}_m(p') \leq \mathcal{F}_m(p) \right\}\right|}{|P_m|} \quad (2)$$

Here, P_m denotes the corpus of statements associated with trait type m, $\mathcal{F}_m(p)$ is the *indicative score* of statement p for type m, and $\mathcal{V}_m(p)$ is its *value* (the empirical percentile of $\mathcal{F}_m(p)$ within P_m). For example, $\mathcal{V}_m(p) = 0.75$ means that 75% of statements labeled m in the dataset have an indicative score less than or equal to that of p.

Initialization of Trait Profile. Let s be the set of a user's known statements. We first compute the average indicative score for each type,

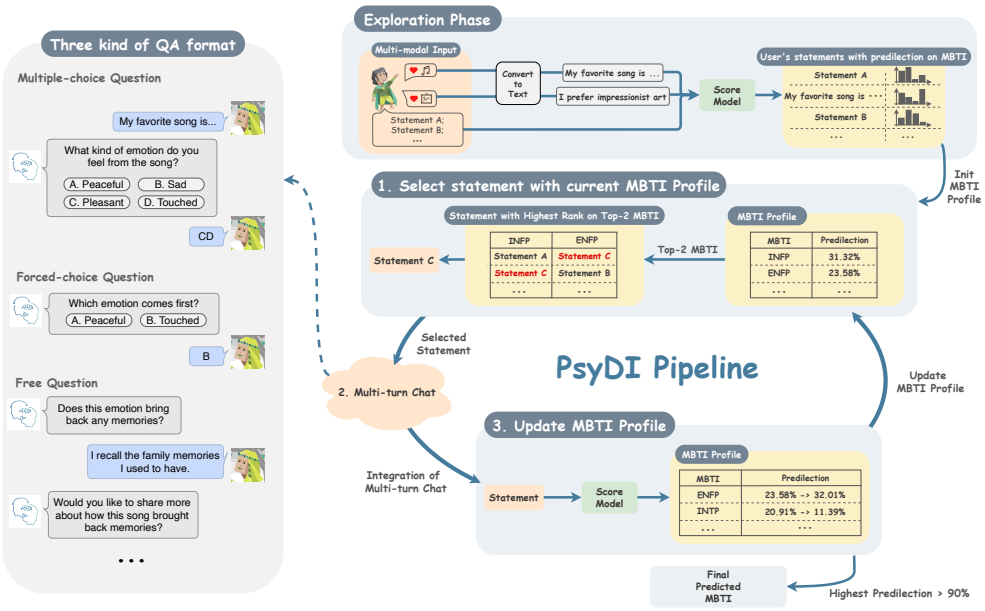


Figure 3: **Pipeline Overview**. Taking MBTI as an example, *PsyDI* operates in a loop comprising three phases. The process begins with the user providing statements to initialize the trait profile. Based on the current profile, *PsyDI* selects a specific statement and engages in a multi-turn dialogue with the user. The interaction outcomes are used to update the profile. This iterative loop persists until *PsyDI* achieves high confidence in the user's MBTI.

$\bar{\mathcal{F}}_m(s) = \frac{1}{|s|}\sum_{p\in s}\mathcal{F}_m(p)$, and map it to a percentile to obtain the initial profile:

$$\mathcal{V}_m^0(s) = \frac{\left|\{\, p' \in P_m \mid \mathcal{F}_m(p') \le \bar{\mathcal{F}}_m(s) \,\}\right|}{|P_m|}. \quad (3)$$

Update of Trait Profile. As new statements arrive, the profile is refined to reflect the evolving evidence while implicitly respecting dependencies among trait types (captured by the score model that jointly learns type relations). Given a new statement p_{new}, we apply an adaptive update:

$$\mathcal{V}_m\big(s\cup\{p_{\text{new}}\}\big) = \mathcal{V}_m(s) + f\big(\mathcal{V}_m(s)\big)\cdot\mathcal{V}_m(p_{\text{new}}), \quad (4)$$

where $f(\cdot)$ is an *adaptive gain* that modulates the step size based on the current profile (e.g., smaller gains when confidence is already high), thereby regulating how strongly a single statement can shift $\mathcal{V}_m$. Intuitively, if a user already shows a strong inclination toward a type, additional statements produce smaller adjustments; if the profile is uncertain, updates are larger. (For stability, the updated value can be clipped to $[0, 1]$.)

3.3 Statement Selection

After updating the Trait Profile, the next challenge is to elicit *more diagnostic* statements through subsequent questioning. Given this provisional portrait, the process proceeds in two phases: (i) **statement selection**, which identifies statements most indicative for disambiguating the current profile, thereby revealing axes that warrant deeper probing; and (ii) **multi-turn dialogue**, which engages users to elaborate on those statements, supplying detail and rationale while sustaining engagement.

To determine which statements merit follow-up, we exclude items with minimal informational value (e.g., generic truisms) and those that explicitly name an trait type. We instead target statements that are *ripe for clarification*—in particular, those that are ambiguous along the axes where the profile's top candidates differ.

Formally, let m_j and m_k be the top-2 trait types in the current profile, and let M be the binary (0/1) version of a type vector, with M_j^i denoting the value of m_j on the i-th dimension. Let $\text{rank}_i(\cdot)$ be the within-state rank on the i-th dimension. We select

$$\text{argmin}_{p\in s}\sum_{i=1}^{4}\big(M_j^i \otimes M_k^i\big) \cdot \Big(\text{rank}_i(\mathcal{F}_{m_j}(p)) + \text{rank}_i(\mathcal{F}_{m_k}(p))\Big) \quad (5)$$

where $\otimes$ denotes XOR, used to mark dimensions on which m_j and m_k differ. The objective in Eq. 5 identifies statements that rank highly on all ambiguous dimensions in state s, enabling targeted, multi-turn questioning to disambiguate the profile and sharpen our understanding of the user's trait tendencies.

	English Benchmarks				Chinese Benchmarks		AVG
	Reddit	Reddit-mix	Reddit-re.	16per.-EN	16per.-ZH	Diamonte	
Closed-source Model							
GPT-4o	61.7%	85.1%	63.8%	76.5%	80.1%	68.1%	72.5%
GPT-4o-mini	62.0%	78.7%	53.2%	82.9%	74.4%	68.0%	69.8%
Gemini-2.5-Flash	53.2%	38.3%	46.8%	51.1%	44.7%	61.7%	49.3%
deepseek-chat	64.9%	88.6%	66.2%	72.1%	81.6%	66.7%	73.3%
moonshot-v1-8k	62.3%	79.3%	43.3%	65.1%	65.5%	50.8%	61.1%
qwen-turbo	55.4%	60.3%	50.8%	82.9%	80.0%	62.7%	65.2%
Baichuan3-Turbo	62.1%	60.1%	51.3%	82.3%	79.1%	60.1%	65.8%
yi-medium	58.2%	71.5%	52.5%	76.5%	81.9%	47.6%	64.6%
Open-source Model							
Llama-2-7b	50.5%	61.6%	45.3%	53.2%	43.2%	57.1%	51.8%
ChatGLM3-6b	57.2%	73.1%	50.2%	76.1%	74.4%	57.0%	64.6%
InternLM2-chat-7b	50.3%	61.7%	52.4%	69.1%	70.5%	62.2%	61.0%
PsyDI-EN	72.3%	98.1%	98.9%	83.4%	80.1%	70.0%	84.0%
PsyDI-ZH	**73.2%**	**98.2%**	**99.8%**	**85.7%**	**82.9%**	**71.9%**	**85.3%**

Table 1: Comparative accuracy of Score Model against open-source and closed-source models in predicting higher scores for p_i in sample$\{p_i, p_j, m\}$ with different datasets. The highest accuracy achieved in each dataset is highlighted in bold, while the second-highest accuracy is underlined.

3.4 Multi-turn Chat

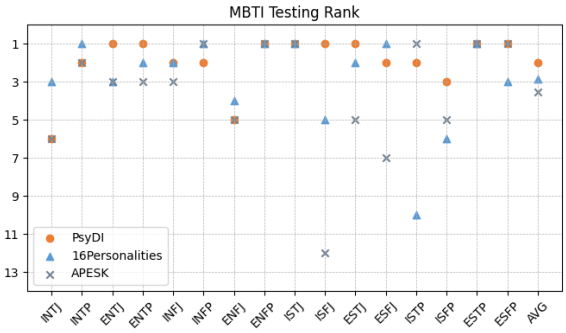
To maintain coherence across complex subgoals, we adopt a modular approach where each interaction is centered on a single statement. PsyDI probes trait-relevant aspects through 3–5 layered, increasingly specific questions, summarizing the exchange into a refined self-description (e.g., expanding "I feel stressed" into specific triggers and coping mechanisms). To reduce ambiguity, we utilize multiple-choice and forced-choice formats mapping to specific cognitive tendencies, while allowing free-form nuance . For personalization, multimodal inputs (e.g., music, images) are converted to text as candidate statements. Each module's summary is iteratively fed back to update the user's Trait profile. All the chat prompt can be found in Appendix E while full conversation examples illustrating the complete workflow are provided in Figure 11, 12, 13, and 14 in Appendix F.2.

3.5 Applicability to other typed assessments.

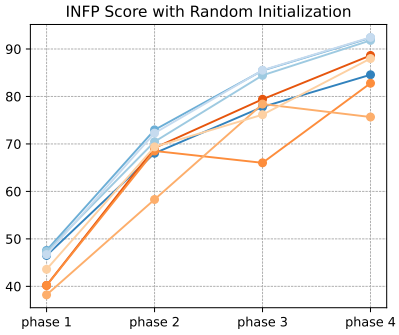
The data refinement presented in Section 2 selection strategy introduced in Section 3.3 is not specific to MBTI. It only requires a finite set of types defined over trait axes and a per-statement scoring function. To port PsyDI to other frameworks (e.g., Big Five, PANAS-X), the binary vector M is replaced with the target encoding, using axis-wise distance weights instead of XOR to disambiguate leading candidates. This makes the policy applicable to any psychological test with discrete types or discretized trait bins (Appendix H).

4 Experiments

In this section, we rigorously evaluate both the data-refinement method and the *PsyDI* agent pipeline. We organize the study around six research questions: **RQ1 — Discrimination vs. LLMs.** Can PsyDI outperform SOTA LLMs in MBTI consistency detection?(Section 4.2, Table 1) **RQ2 — Content, not length.** Does the model prioritize trait-bearing semantics over surface length? (Figure 15) **RQ3 — Source of gains.** Which components (loss, data augmentation, architecture) drive performance? (Appendix G.2, Tables 5, 6) **RQ4 — End-to-end validity.** Does PsyDI produce accurate rankings in autonomous bot interactions?(Section 4.3, Figure 4 (left)) **RQ5 — Robustness to initialization.** Can the pipeline recover from noisy/random MBTI profile seeds?(Figure 4 (right)) **RQ6 — Generality & external validity.** (a) Does the refinement+ranking framework transfer to other inventories (Big Five, PANAS-X)? (b) Do large-scale on-



(a) MBTI rank across testing.



(b) INFP scores with random init.

Figure 4: Evaluation of MBTI models. (a) Rankings of 16 MBTI bots across three testing conditions. (b) INFP scores under random profile initialization.

line behaviors/lexical patterns and projective "blob tree" choices align with MBTI measurements? (Appendix H; Section 4.3 & Appendix F, Figure 5)

4.1 Experimental Setups

To fairly evaluate the performance of the data refinement method, we have gathered multiple datasets and compared them with both current state-of-the-art open-source and closed-source models. The details of our datasets, baselines and evaluating methods can be found in Appendix D.

4.2 Score Model Evaluation

Table 1 reports accuracy for the PsyDI score model trained on refined targets versus baselines. Within this comparison setting, *PsyDI* performs best on both English and Chinese benchmarks.

On English datasets, *PsyDI-EN* and *PsyDI-ZH* exceed zero-shot baselines by at least 9.5% on *Reddit* and *Reddit-mix*. On *Reddit-repeat*, where many baseline models show a length bias (overvaluing longer, redundant statements), *PsyDI* correctly assigns lower scores to repetitive long statementswhich shows evidence that refinement plus ranking reduces such nuisance effects. On *16Personalities-EN*, which features distinctive, idiomatic expressions, *PsyDI-EN* and *PsyDI-ZH* achieve 83.4% and 85.7%, respectively, compared to open-source zero-shot averages around 82%.

On Chinese datasets, *PsyDI-EN*, despite no Chinese fine-tuning, outperforms most zero-shot models, suggesting that the refined, order-based signal transfers across languages. After statement-level fine-tuning on Chinese data, *PsyDI-ZH* surpasses other zero-shot LLMs by 7.5% on *Diamante* and matches yi-medium on *16Personalities-ZH*. These results support that order-consistent refinement plus ranking-based scoring yields robust gains over zero-shot LLMs and transfers across languages.

To analyze how the score model responds along the four MBTI dimensions, we conduct a controlled augmentation study in Appendix G.1.

4.3 Pipeline Evaluation

In this section, we evaluate the effectiveness of the *PsyDI* pipeline by comparing its MBTI ranking precision with two widely used tests: *16Personalities* and *APESK* (APESK, 2024). We utilize the 16 most commonly used MBTI bots from *Character AI* and have each bot respond to the questions from these two tests and *PsyDI*. We document the rank of the bot's actual MBTI type among all other MBTI types under a specific MBTI test. The comparative results are depicted in Figure 4(a), where the horizontal axis represents different MBTI bots, with each point indicating the rank of the bot's actual MBTI. The results reveal that *PsyDI*'s average ranking accuracy is comparable or superior to that of the two established tests.

Additionally, since PsyDI went live [1], it has been tested by over 10,000 users. Based on this substantial dataset, we conducted a thorough analysis of the collected data.

Moreover, for the majority of bots, their actual MBTI type is among the top three predictions made by *PsyDI*, which demonstrates greater stability compared to others. We also investigate the test interaction procedure of the *Character AI* bot with *PsyDI*. The result is shown in Appendix G.3.

On the other hand, in Figure 4(b), we initiate

[1] The website address will be published after the review

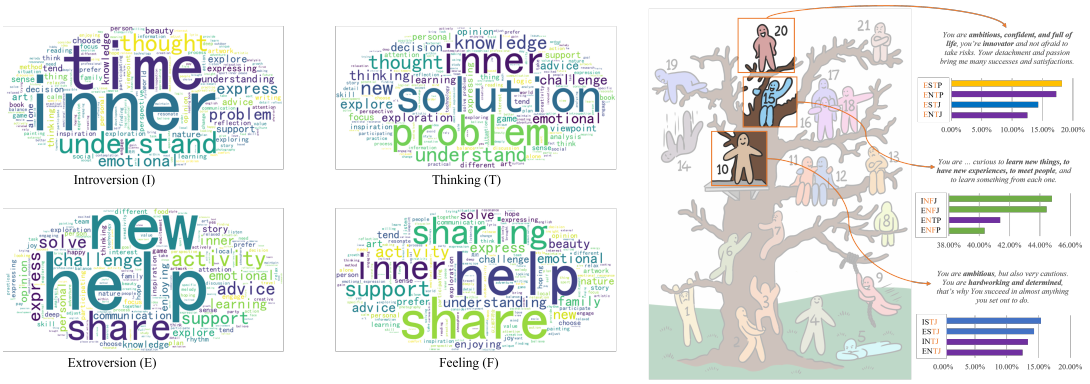


Figure 5: Data analysis of PsyDI Online. Left figure is the word frequency of all the MBTI users. Right figure is the top-4 MBTI user choosing each blob in blob tree.

the experiment with several randomized MBTI profiles and utilize them for further assessment with an INFP bot provided from *Character AI* (CharacterAI, 2024). Initially yielding INFP scores between 35 and 50, the scores consistently rise and eventually stabilize above 80. This indicates that *PsyDI* effectively refines the MBTI profile through extensive dialogues, even with an unfavorable initial state. We also introduce supplementary experiments, encompassing a semantic scoring experiment aimed at discerning the scoring tendencies of the score model in Appendix G.4.

4.4 Online Evaluation

First, we examined word frequency patterns in user responses across different MBTI types, generating word clouds for each MBTI dimension. For instance, in the I/E dimension, we isolated words specific to introverted (IXXX) users by excluding common words from extroverted (EXXX) responses. Figure 5 (a) displays the results for the I/E and F/T dimensions. Our analysis revealed a strong correlation between PsyDI's MBTI detection and the common word usage in user self-reports. Introverted users often used introspective words like "inner," while extroverted users favored socially interactive words like "help" and "share." Thinking users focused on problem-solving, whereas feeling users emphasized helping and supporting others. These patterns confirm PsyDI's MBTI measurements align with expected MBTI behaviors.

Furthermore, we analyzed MBTI users' choices in a multimodal blob tree task from PsyDI. We calculated the selection proportions for each MBTI type and identified significant patterns, shown in Figure 5 (b). For example, EXTX users predominantly chose position 20, described as "ambitious, confident, and full of life," reflecting their extroverted and logical traits. XNFX users favored position 15, associated with openness and friendliness, aligning with their intuitive and feeling tendencies. These findings support PsyDI's MBTI measurements, validated through the blob tree task, a projective test, underscoring PsyDI's effectiveness.

Moreover, we have extended the *PsyDI* pipeline to other psychological topics to demonstrate its scalability, as shown in Appendix H.

5 Conclusion

We present PsyDI, a unified framework for psychological assessment that addresses the limitations of standardized scales and the limited accessibility of expert evaluations. We adopt a data-centric approach that refines noisy self-reported data for effective integration with LLMs. The refinement procedure combines multi-strategy data augmentation with LLM-based ranking to construct order-aware supervision,. Based on these refined targets, we train a trait prediction model with a ranking loss instead of direct score imitation. The resulting score model is embedded into the PsyDI pipeline, which enhance user engagement and improve latent trait estimation. Experimental results on MBTI prediction demonstrate that PsyDI achieves higher trait ranking accuracy and improved downstream classification. A public deployment has attracted more than 10,000 users, resulting in a large annotated dialogue dataset that provides useful resources for future research. Furthermore, the sequential interactions and user feedback indicate promising directions, including training end-to-end models with RLHF and extending PsyDI into a streaming assistant for more comprehensive psychological assessment.

Limitations

Despite its effectiveness, PsyDI has several limitations. First, our primary evaluation is centered on the MBTI framework. Although we demonstrate generalizability to the Big Five and PANAS-X, the depth of validation for these inventories is less extensive than for MBTI. Second, while the data-centric refinement improves trait ordering, the score model's performance remains upper-bounded by the linguistic prior of the base LLM (e.g., Llama-2). Third, our study relies on Character-AI bots for end-to-end validity (RQ4-RQ5), which may not fully capture the strategic or emotional nuances of diverse human populations. Finally, the multi-modal inputs are currently normalized to text for processing; a truly end-to-end multimodal architecture might yield deeper psychological insights.

Ethical Considerations

This work explores the use of Large Language Models (LLMs) for psychological assessment through a conversational framework. Our implementation specifically instantiates a chatbot for MBTI trait profiling. While *PsyDI* demonstrates promising potential for scalable and adaptive mental trait evaluation, we acknowledge several ethical considerations.

Ethics Review and Informed Consent. The data collection protocol for our online deployment was reviewed and determined to be exempt by our internal ethics review process, as it involves the analysis of de-identified, anonymized conversational data. All participants provided explicit informed consent via a mandatory digital agreement before interacting with the system, which detailed the research purposes and data handling procedures.

User Privacy and Data Protection. *PsyDI* collects psychological data, which raises privacy concerns. To protect confidentiality, all data is anonymized before analysis. We have added a dedicated privacy page in the online system to inform users about data usage. No data will be publicly released without explicit user consent. We commit to updating our methods in compliance with emerging privacy regulations.

Safety in Language Model Interaction. LLMs may generate unsafe content in sensitive scenarios. We embed safety instructions in prompts; for instance, the model checks if user input is aggressive or manipulative, halting the conversation if necessary. While these safeguards are not perfect, we plan to continuously enhance harmful input detection.

Scientific Validity and Misuse. *PsyDI* has not undergone clinical validation and should not be used as a diagnostic tool. We explicitly position it as a research and educational tool, not a substitute for licensed professionals. Clear disclaimers are included to prevent over-reliance or misuse of the MBTI profiling results.

Fairness and Inclusivity. The LLMs and datasets may encode cultural or gender biases. We acknowledge this limitation and call for future work to ensure equitable outcomes across diverse demographic groups.

Use of AI Assistants. In the preparation of this manuscript, the authors utilized large language models to assist in linguistic polishing, structural refinement, and the generation of specific LaTeX formatting. However, the core methodology, experimental design, and data analysis were conducted solely by the human authors, who maintain full accountability for the integrity and accuracy of the content.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

APESK. 2024. Apesk. https://www.apesk.com/.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, and 1 others. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Marcel Binz and Eric Schulz. 2023. Turning large language models into cognitive models. *arXiv preprint arXiv:2306.03917*.

Gregory J. Boyle. 1995. Myers-briggs type indicator (mbti): Some psychometric limitations. *Humanities & Social Sciences papers*, 30.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, and 1 others. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.

Sven Buechel, Anneke Buffone, Barry Slaff, Lyle H. Ungar, and João Sedoc. 2018. Modeling empathy and distress in reaction to news stories. In *Proceedings of the 2018 Conference on Empirical Methods*

in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018, pages 4758–4765. Association for Computational Linguistics.

Zheng Cai and Maosong Cao. 2024. Internlm2 technical report. *Preprint*, arXiv:2403.17297.

CharacterAI. 2024. Characterai. https://beta.character.ai/.

Tianyi Chen, Feiqi Cao, Yihao Ding, and Caren Han. 2024. The language model can have the personality: Joint learning for personality enhanced language model (student abstract). *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(21):23454–23455.

Jiale Cheng, Sahand Sabour, Hao Sun, Zhuang Chen, and Minlie Huang. 2023. PAL: persona-augmented emotional support conversation generation. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 535–554. Association for Computational Linguistics.

Charles K Coe. 1992. The mbti: Potential uses and misuses in personnel administration. *Public Personnel Management*, 21(4):511–522.

Ronald Jay Cohen, Mark E Swerdlik, and Suzanne M Phillips. 1996. *Psychological testing and assessment: An introduction to tests and measurement*. Mayfield Publishing Co.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Nigel Crisp and Chen Lincoln. 2014. Global supply of health professionals. *The New England journal of medicine*, 370(10):1668.

Jiaxi Cui, Liuzhenghao Lv, Jing Wen, Jing Tang, YongHong Tian, and Li Yuan. 2023. Machine mindset: An mbti exploration of large language models. *arXiv preprint arXiv:2312.12999*.

Tri Dao. 2023. Flashattention-2: Faster attention with better parallelism and work partitioning. *arXiv preprint arXiv:2307.08691*.

DeepSeek-AI. 2024. Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model. *Preprint*, arXiv:2405.04434.

Dorottya Demszky, Diyi Yang, David S. Yeager, Christopher J. Bryan, Margarett Clapper, Susannah Chandhok, Johannes C. Eichstaedt, Cameron Hecht, Jeremy Jamieson, Meghann Johnson, Michaela Jones, Danielle Krettek-Cobb, Leslie Lai, Nirel JonesMitchell, Desmond C. Ong, Carol S. Dweck, and James J. Gross. 2023. Using large language models in psychology. *Nature Reviews Psychology*, pages 688–701.

John M Digman. 1997. Higher-order factors of the big five. *Journal of personality and social psychology*, 73(6):1246.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. Glm: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 320–335.

Adrian Furnham. 1996. The big five versus the big four: the relationship between the myers-briggs type indicator (mbti) and neo-pi five factor model of personality. *Personality and individual differences*, 21(2):303–307.

Adrian Furnham. 2022. The big five facets and the mbti: The relationship between the 30 neo-pi (r) facets and the four myers-briggs type indicator (mbti) scores. *Psychology*, 13(10):1504–1516.

Adrian Furnham and John Crump. 2005. Personality traits, types, and disorders: an examination of the relationship between three self-report measures. *European Journal of Personality*, 19(3):167–184.

Carlton E Green, Cynthia E Chen, Janet E Helms, and Kevin T Henze. 2011. Recent reliability reporting practices in psychological assessment: Recognizing the people behind the data. *Psychological Assessment*, 23(3):656.

Clara E Hill, Maureen M Corbett, Bettina Kanitz, Peggy Rios, Richard Lightsey, and Maria Gomez. 1992. Client behavior in counseling and therapy sessions: Development of a pantheoretical measure. *Journal of Counseling psychology*, 39(4):539.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Carl Jung and John Beebe. 2016. *Psychological types*. Routledge.

Carl Gustav Jung. 1976. *Collected works of CG Jung, volume 6: Psychological types*, volume 6. Princeton University Press.

Luoma Ke, Song Tong, Peng Chen, and Kaiping Peng. 2024. Exploring the frontiers of llms in psychological applications: A comprehensive review. *arXiv preprint arXiv:2401.01519*.

Wai-Chung Kwan, Xingshan Zeng, Yuxin Jiang, Yufei Wang, Liangyou Li, Lifeng Shang, Xin Jiang, Qun Liu, and Kam-Fai Wong. 2024. Mt-eval: A multi-turn capabilities evaluation benchmark for large language models. *arXiv preprint arXiv:2401.16745*.

Lucio La Cava, Davide Costa, and Andrea Tagarelli. 2024. Open models, closed minds? on agents capabilities in mimicking human personalities through

open large language models. *arXiv preprint arXiv:2401.07115*.

David Lamond. 2001. The myers-briggs type indicator: Evidence of its validity, reliability and normative characteristics for managers in an australian context. In *4th Australian Industrial and Organisational Psychology Conference*, volume 6.

Minghuan Liu, Menghui Zhu, and Weinan Zhang. 2022. Goal-conditioned reinforcement learning: Problems and solutions. *CoRR*, abs/2201.08299.

Sally Merry, Karolina Stasiak, Matthew Shepherd, Chris Frampton, Theresa Fleming, and Mathijs Lucassen. 2012. The effectiveness of sparx, a computerised self help intervention for adolescents seeking help for depression: Randomised controlled non-inferiority trial. *BMJ (Clinical research ed.)*, 344:e2598.

Isabel Briggs Myers and 1 others. 1962. *The myers-briggs type indicator*, volume 34. Consulting Psychologists Press Palo Alto, CA.

Yazhe Niu, Jingxin Xu, Yuan Pu, Yunpeng Nie, Jinouwen Zhang, Shuai Hu, Liangxuan Zhao, Ming Zhang, and Yu Liu. 2021. Di-engine: A universal ai system/engine for decision intelligence. https://github.com/opendilab/DI-engine.

Haocong Rao, Cyril Leung, and Chunyan Miao. 2023. Can chatgpt assess human personalities? a general evaluation framework. *arXiv preprint arXiv:2303.01248*.

David Rapaport. 1942. Principles underlying projective techniques. *Character & Personality*, 10(3).

Sebastian Schuster, Sonal Gupta, Rushin Shah, and Mike Lewis. 2018. Cross-lingual transfer learning for multilingual task oriented dialog. *arXiv preprint arXiv:1810.13327*.

Xiaoyang Song, Yuta Adachi, Jessie Feng, Mouwei Lin, Linhao Yu, Frank Li, Akshat Gupta, Gopala Anumanchipalli, and Simerjot Kaur. 2024. Identifying multiple personalities in large language models with external evaluation. *arXiv preprint arXiv:2402.14805*.

Dylan Storey. 2018. Myers briggs personality tags on reddit data.

Richard S Sutton and Andrew G Barto. 2018. *Reinforcement learning: An introduction*. MIT press.

Yeldar Toleubay, Don Joven Agravante, Daiki Kimura, Baihan Lin, Djallel Bouneffouf, and Michiaki Tatsubori. 2023. Utterance classification with logical neural network: Explainable AI for mental disorder diagnosis. In *Proceedings of the 5th Clinical Natural Language Processing Workshop, ClinicalNLP@ACL 2023, Toronto, Canada, July 14, 2023*, pages 439–446. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, and 1 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Yu-Min Tseng, Yu-Chao Huang, Teng-Yun Hsiao, Yu-Ching Hsu, Jia-Yin Foo, Chao-Wei Huang, and Yun-Nung Chen. 2024. Two tales of persona in llms: A survey of role-playing and personalization. *arXiv preprint arXiv:2406.01171*.

Tao Tu, Anil Palepu, Mike Schaekermann, Khaled Saab, Jan Freyberg, Ryutaro Tanno, Amy Wang, Brenna Li, Mohamed Amin, Nenad Tomasev, and 1 others. 2024. Towards conversational diagnostic ai. *arXiv preprint arXiv:2401.05654*.

David Watson and Lee Anna Clark. 1994. The panas-x: Manual for the positive and negative affect schedule-expanded form. *Unpublished manuscript, University of Iowa*.

Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Ce Bian, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, and 1 others. 2023. Baichuan 2: Open large-scale language models. *arXiv preprint arXiv:2309.10305*.

Qisen Yang, Zekun Wang, Honghui Chen, Shenzhi Wang, Yifan Pu, Xin Gao, Wenhao Huang, Shiji Song, and Gao Huang. 2024. Llm agents for psychology: A study on gamified assessments. *arXiv preprint arXiv:2402.12326*.

Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, and 1 others. 2024. Yi: Open foundation models by 01. ai. *arXiv preprint arXiv:2403.04652*.

Chenhao Zhang, Renhao Li, Minghuan Tan, Min Yang, Jingwei Zhu, Di Yang, Jiahao Zhao, Guancheng Ye, Chengming Li, Xiping Hu, and 1 others. 2024. Cpsycoun: A report-based multi-turn dialogue reconstruction and evaluation framework for chinese psychological counseling. *arXiv preprint arXiv:2405.16433*.

A The Background of MBTI

The Myers-Briggs Type Indicator (MBTI) (Coe, 1992) is a psychological metric designed to classify individuals based on their personality traits, resulting in a four-letter typology. Each letter in the MBTI represents one of two possible traits in four dichotomies: Introversion (I) vs. Extraversion (E), Sensing (S) vs. Intuition (N), Thinking (T) vs. Feeling (F), and Judging (J) vs. Perceiving (P). Consequently, MBTI theory categorizes individuals into one of sixteen possible personality types, such as INFP or ESTJ.

The unique combination of these four dimensions results in sixteen distinct personality types, each with its own cognitive patterns and behavioral tendencies. This heterogeneity has contributed to the growing popularity of the MBTI in recent years, with over 1.207 billion test administrations recorded by 16Personalities. However, it is imperative to acknowledge that the MBTI is based on Carl Jung's theory of psychological types (**?**). Jung's theory postulated a quaternary of cognitive functions—thinking, feeling, sensation, and intuition—each with an orientation of either extroversion or introversion, resulting in eight dominant functions, Te (Extroverted Thinking), Ti (Introverted Thinking), Fe (Extroverted Feeling), Fi (Introverted Feeling), Se (Extroverted Sensing), Si (Introverted Sensing), Ne (Extroverted Intuition), and Ni (Introverted Intuition). The constitution of an individual's personality is delineated by the hierarchical preference for these dominant functions. According to the theory of Jung, it is improbable for a person to exhibit high predilections for both extraversion and introversion within the same cognitive function, as human cognitive resources are finite. This principle leads to the emergence of sixteen common cognitive function sequences, aligning congruently with the sixteen MBTI types. For example, the cognitive function sequence of INFP is: Fi, Ne, Si, Te, Fe, Ni, Se, Ti. Consequently, the adjudication of MBTI typologies inherently entails the scrutiny of one's cognitive processes. The cognitive preferences, on the other hand, refer to the individual's preferred or dominant use of these functions.

Moreover, the evaluation of MBTI is inherently multifaceted and complex due to its dependence on self-reports. These self-reports are inherently subjective and are prone to distortion by various extrinsic factors (Green et al., 2011), including mood states and the pervasive influence of social desirability. Thus, accurately correlating these self-reports with the underlying cognitive processes presents a formidable challenge. To improve this, it is essential to carefully analyze the specific statements within the self-reports. This can be achieved through a detailed questioning framework that eliminates ambiguity and elicits more precise responses. Such an approach is instrumental in mitigating the effects of external influences and in clarifying cognitive preferences with greater precision and depth.

B Related Works

B.1 Psychological measurement

As one of the primary measuring methods in psychology, measurement has been developed a long history and has been established a comprehension measurement system covering different subdomains, such as clinical diagnosis, military mental state measurement and so on (Cohen et al., 1996). Psychological measurement has many categories, including testing, interview, behavior observation experiment, etc. Testing has been the most preferable tools in many scenery due to its preciseness based on well-designed criterion, consistent measurement despite scenery and convenience in re-usability. However, it needs sufficient expert opinions to design as well as lacks personalized testing to deepen the measurement (Merry et al., 2012). Moreover, expert interview make up these short comes, yet suffers from plenty of resources to collect sparse, non-mass-production-able instances (Crisp and Lincoln, 2014). The Myers-briggs type indicator (MBTI) is a self-report questionnaire on personality types in personnel management (Myers et al., 1962). This test and current concepts are originated from Carl Jung's personality types, a typology theory building and investigating an individual's mental cognitive model (Jung, 1976) and has been extended into 16 personalities by John Beebe (Jung and Beebe, 2016). Although originated from but not belongs to psychological research field, the measurement method of MBTI test has been discussed as having reference value and also gain a wide accept between many individuals (Boyle, 1995; Lamond, 2001). For example, previous research has shown a significant positive correlation between MBTI and the Big Five personality test (Furnham, 1996, 2022).

B.2 Large Language Models towards Psychological and MBTI

In the advantage of capability emergence firstly evolved out by OpenAI's GPT models, LLMs could spontaneously capture the deep understanding in contextual relevant contents, achieving human-like level performance (Achiam et al., 2023). Various works have discussed the pros and cons large language models has reached in psychological research field, demonstrate the potential capabilities LLMs mastered in psychological chatting, analysing, guiding and reasoning (Ke et al., 2024; Demszky et al., 2023; Binz and Schulz, 2023). Current cross-over applications includes multi-turn dialogue framework for psychological counseling (Zhang et al., 2024), conversational diagnose (Tu et al., 2024), gamified assessments (Yang et al., 2024), introducing more noval pipelines under this scope. Also been a growing force in psycho-cognitive research, LLMs have shown a capable solution to recognize and simulate human cognitive process (Bubeck et al., 2023). This helps MBTI-related AI application exploration whose essence is human cognitive patterns. A series of works have investigated how LLMs captures cognitive personaliies via role-playing different types (Chen et al., 2024; Cui et al., 2023; Tseng et al., 2024; La Cava et al., 2024; Kwan et al., 2024). LLM-driven measurement abilities are studied to confirm the MBTI awareness also contributes personality discrimination (Rao et al., 2023; Song et al., 2024). A gamified MBTI test has been designed and evaluated on PsychoGAT (Yang et al., 2024).

C Problem Formulation

We initiate our analysis by modeling the measurement of psychological indicators in Section C.1 based on the characteristics of MBTI, as detailed in the Appendix A. Then we introduce the relationship between the MDP modeling and our designed framework in Section C.2

C.1 MDP Definition for psychological measurement

We conceptualize the interaction between LLM agents and the test taker (user) as a Markov Decision Process (MDP) (Sutton and Barto, 2018). This formulation allows us to systematically optimize the sequence of inquiries posed to the user, with the aim of enhancing the accuracy, efficiency and engagement of MBTI evaluations. Formally, the MDP is articulated as $\mathcal{M} = \{\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma\}$. The states s in the observable state space $\mathcal{S}$ are defined as $s = \{p_1, \ldots, p_n\}$, where p_i is the i-th textual statement of the user (*e.g. I like to talk to different people and share my stories*). Additionally, the state space includes a special terminated state s_{exit} representing the user's decision to exit the evaluation process.

Throughout the process, the agent needs to determine to either persistently delve deeper into user's cognitive preferences through multi-turn iterative Q&A, or to terminate the process and render a forecasted MBTI type of the user. We define MBTI as m, a four-dimensional vector composed of elements 1 and -1, which symbolizes the orientation (E/I, N/S, F/T, J/P) of the MBTI along its four dimensions. For instance, the MBTI type INFP is represented by the vector: $m = [-1, 1, 1, -1]$. While gathering more question-answers may narrow down the user's authentic MBTI, it is imperative to acknowledge that an extended engagement may engender user impatient. Therefore, the action space $\mathcal{A}$ is defined as the union of these two kinds of action set $\mathcal{A} = \mathcal{A}_q \cup \mathcal{A}_t$. Here each a in $\mathcal{A}_q$ represents a textual question $a = \{w_1, \ldots, w_L\}$ and $\mathcal{A}_t$ represents the set of termination actions, i.e., all possible MBTI types, where L is the length of the question. Once the user selects an answer, the question together with the answer can be regarded as a more detailed user statement p'. Consequently, the state transition for the state s is $s' = s \oplus p'$, where $\oplus$ signifies the concatenation of the current state with the new p'. In the event that the user opts to exit, the state transmutes to s_{exit}.

The reward function, which maps $\mathcal{S} \times \mathcal{A}$ to $\mathbb{R}$, is designed to measure the similarity between the predicted MBTI and the user's authentic MBTI. Concurrently, it imposes a penalty for the duration of the interaction and the incidence of user withdrawal. Formally, the reward function is defined as:

$$r(s_t, a_t) = \begin{cases} -c, & a \in \mathcal{A}_q, s' = s_{\text{exit}} \\ -|a, m^*| - \lambda \cdot t, & a \in \mathcal{A}_t \end{cases} \tag{6}$$

where $|a_t, m^*|$ denotes the measure of similarity between the predicted and the true MBTI type m^*, λ is the penalty factor for the cumulative interaction steps t while c is the penalty positive constant for the exit.

In summary, the MDP employed in this context

is similar to challenges tackled in goal-conditioned reinforcement learning (GCRL) (Liu et al., 2022). The purpose of this decision-making process is to achieve a specific objective (goal), i.e., to accurately predict user's MBTI type within the minimal number of interactions.

C.2 Relationship between MDP and Practical Framework

The MDP environment previously defined is an idealized model, providing a clear framework for subsequent analysis. Realistically, we can hardly construct a perfectly ideal MDP due to the inability to directly obtain users' true psychological indicators, thus precluding accurate reward determination. Instead, we employed a heuristic algorithm to handle reward uncertainty flexibly.

The original action space in MDP we defined is $\mathcal{A} = \mathcal{A}_q \cup \mathcal{A}_t$, where $\mathcal{A}_q$ represents the entirety of possible questions that can be posed. Considering the vastness of the original action space, we propose a heuristic algorithm that leverages an intuitive understanding that *a well-formed question should revolve around the user's statements* (Hill et al., 1992). By decomposing the action space into two sequential phases—first selecting a statement and subsequently generating a question centered on that statement—we effectively reduce the dimensionality of the action space. Specifically, the decomposed action space is $\bigcup_{i=1}^{n}(a_{p_i} \times Q_{p_i})$, where a_{p_i} represents for selecting statement p_i to ask following question and Q_{p_i} is the question space related to statement p_i.

Leveraging the powerful language capabilities and goal-oriented abilities of LLMs within short conversational rounds, we further extend the question-generating process, denoted as q_{p_i}, into a multi-turn dialogue. This multi-turn dialogue is a sub-process, where each question is based on the statement p_i and the history of question-answer pairs h_j within it. Consequently, the action $q_{p_i j}$ represents the j-th question in the multi-turn dialogue centered around p_i. Therefore, the action space is $\bigcup_{i=1}^{n}(a_{p_i} \times \bigcup_{j=1}^{m} q_{p_i j})$. In the subsequent pipeline design, we will demonstrate how to step-by-step select actions within this action space to optimize the return.

D Experimental Settings Details

We have represented all the details of our datasets(Appendix D.1), baselines(Appendix D.3), and evaluation methods(Appendix D.5) in this section. We have also managed the distribution and quality of our datasets in Appendix D.2, and assess the accuracy of MBTI bot used in the experiments in Appendix D.6. And the training setting details are presented in Appendix D.4.

D.1 Datasets

We constructed the following MBTI statement datasets with labels in both English and Chinese:

- **Reddit**. The Myers Briggs Personality Tags on Reddit Data (Storey, 2018) is a public dataset consisting of user statements and their self-reported MBTI types. To balance the distribution of MBTI types, we extracted 400 statements for each MBTI type, ensuring that each statement was longer than 100 words. The Reddit dataset is augmented with the data augmentation methods introduced in section 2.3, forming Reddit-mix and Reddit-repeat.

- **16Personalities**. We obtained all the questions from 16 Personalities [2], an authoritative MBTI testing site. Subsequently, we used one LLM to simulate specific MBTI types and engaged another LLM in multi-turn Q&A sessions with the former one. If the predicting LLM correctly predicted the MBTI type of simulating LLM, the Q&A record was collected along with the MBTI labels. We collect the dataset in both English and Chinese to form 16Personalities-en and 16Personalities-zh.

- **Diamente** The Diamente Chinese chit-chat dataset comprises high-quality chit-chat conversations collected with the assistance of a pre-trained dialogue model that aligns well with human values. Using these chit-chats as a starting point, we asked LLM to generate multi-turn Q&A sessions. We randomly selected answers and recorded the entire process, labeling the statements based on LLM's predictions to form dataset Diamente.

Details of the above datasets are described in Table 2. Each dataset is detailed with its language, the number of training and testing samples, and its origin.

D.2 Datasets Quality Evaluation

During the dataset construction process, we carefully managed the distribution and quality of the

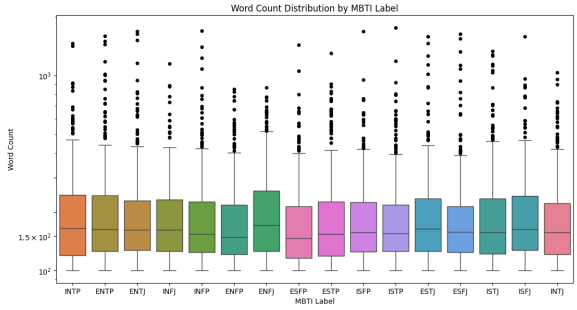
[2] https://www.16personalities.com/

Dataset	Language	#Train Samples	#Test Samples	Origin
Reddit	English	166580	1850	Reddit statements and the corresponding tags
Reddit-mix	English	16659	1850	Mix data augmentation version of Reddit
Reddit-repeat	English	16544	1850	Repeat data augmentation version of Reddit
16Personalities	English/ Chinese	6072	680	Combined responses to 16Personalities from each MBTI
Diamonte	Chinese	65810	762	Conversations about the Chinese chit-chat dataset Diamonte, enerated through two LLM instances

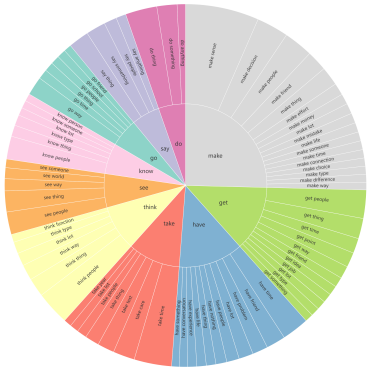
Table 2: Details of training and testing datasets in PsyDI.

data. Specifically, we filtered out statements based on their length and sampled 400 instances for each MBTI type, ensuring a balanced and representative dataset.

We also examined the length distribution and topic coverage of the dataset. For example, using the Reddit dataset, we created a box plot in Figure 6(a) to show the length distribution for all MBTI types. The plot shows that the length distributions are quite similar across all MBTI types, with most texts ranging from 100 to 300 words, although some longer texts are also present. This indicates that the dataset exhibits a consistent text length across different MBTI types, allowing the model to focus more on the personality information contained within balanced text lengths In Figure 6(b), we displayed the main verbs and their corresponding verb-noun phrases in the dataset. This analysis shows that the dataset covers a wide range of topics and includes many verb-noun phrases strongly related to MBTI, such as "make decision."



(a) Box plot of length distribution of Reddit dataset



(b) Top verb-noun phrases in Reddit dataset

Figure 6: Distribution of Reddit dataset

D.3 Baselines

We compared the proposed score model in *PsyDI* with existing closed-source LLMs and open-source LLMs. The closed-source LLMs include GPT-4o and GPT-4o-mini (Achiam et al., 2023), the acclaimed leader in LLMs, as well as Gemini-2.5-Flash (Comanici et al., 2025), DeepSeek (DeepSeek-AI, 2024), a Mixture-of-Experts (MOE) Language Model for both English and Chinese. We also evaluate the performance of Moonshot citepzhang2024moonshot, Qwen (Bai et al., 2023), Baichuan (Yang et al., 2023), and yi (Young et al., 2024). For open-source LLMs, we use Llama2 (Touvron et al., 2023), a commonly used English LLM, Chatglm3 (Du et al., 2022), an open bilingual dialogue language model, and InternLM2 (Cai and Cao, 2024).

D.4 Training Setting

We train PsyDI using llama-2-7b-hf as a base model and finetune with LoRA (Hu et al., 2021). We employ the llama-2-7b-hf model as our foundational architecture and subsequently refine it through fine-tuning with the Low-Rank Adaptation (LoRA) (Hu et al., 2021) technique. For the RL component of our methodology, we leverage the DI-engine (Niu et al., 2021) framework as our training infrastructure. Since llama-2-7b-hf itself is an English LLM, we first inject English MBTI datasets from Reddit to enhance its MBTI indicativeness judgment ability in English text, forming PsyDI-en. Then, we apply cross-lingual transfer learning (Schuster et al., 2018), finetuning PsyDI-en with both Reddit and Chinese datasets Diamente, so that PsyDI-ch retains its understanding of MBTI in English text while also mastering Chinese text.

The score model training involved specific configurations and hyperparameters, as outlined in Table 3.

Lora Configuration. The Lora parameters are utilized for fine-tuning the score model. The parameter r represents the rank within Lora (Dao, 2023), which is typically adjusted based on the balance between fine-tuning performance and training resources. In our case, r is set to 32. According to Dao (2023), the parameter α is generally set as a constant multiple of r. Adjusting α can have a similar effect to adjusting the learning rate. Therefore, we set α to be equal to r, which is 32. The dropout rate for Lora is typically set to 0.05.

Score Model Training Configuration. The score model is fine-tuned based on the llama-2-7b-hf model in sequence classification mode, with the number of labels set to 4. The sequence classification mode adds a layer with heads as a classifier. The base model is loaded in 8-bit to conserve training resources. To accelerate the training and inference process, we use Flash Attention 2 (Dao, 2023). For the ranking margin loss hyperparameter, we set the margin to 0.3, which is enough for the model to clearly delineated the boundaries between statements. The learning rate scheduler is Adam, with the learning rate set to 5e-5, warmup ratio at 0.02, and weight decay at 0.01. These configurations were carefully chosen to optimize model performance during training.

Multi-turn Chat Generation We utilize the latest version of GPT-3.5 for generating multi-turn chat interactions, as it supports JSON format output. The temperature is set to 0.7, and the top-p is set to 0.95 to ensure flexible and varied question generation. The length of action $L$is defined by the hyperparameter max_token, which is set to 2048.

Figure Generation In PsyDI, we ultimately provide users with an image that describes their overall temperament. These images are generated using the MiaoHua [3] model, sgl_artist_v0.4.0. Each image is produced with dimensions of 960×960 pixels. To ensure high-quality image generation, we set MiaoHua's parameters as shown in Table 3.

Table 3: Hyperparameters in PsyDI

Hyperparameters	**Value**
Lora config	
task_type	TaskType.SEQ_CLS
r	32
α	32
dropout	0.05
Score Model Training config	
base model	llama-2-7b-hf
num_labels	4
load_in_8bit	True
attn implementation	flash_attention_2 (Dao, 2023)
batchsize	8
margin	0.3
learning rate scheduler	Adam
warmup raio	0.02
weight decay	0.01
learning rate	1e-4
Multi-turn chat Generate	
model	gpt-35-turbo-1106
temperature	0.7
top p	0.95
frequency penalty	0
presence penalty	0
max_token	2048
Figure Generate	
model	artist_v0.4.0
height×width	960×960
cfg_scale	7.0
vae	vae_sd_84000
seed	42
step	50
sampler	DDIM

[3] https://miaohua.sensetime.com/

D.5 Evaluating Method

To evaluate the score model in PsyDI, for each pair-wise data sample $(p_i, p_j, mbti)$, we asked all baseline LLMs to determine which statement was more likely from a user with the specified MBTI type $mbti$. The prompt was designed as follows:

I will give you two statement, Please choose the one you think is more ... Just Answer (A) or (B).

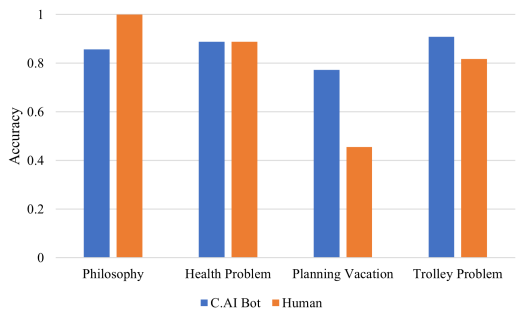
All LLMs had their temperature set to 0.3 to ensure more deterministic outputs. For PsyDI, we calculate the score $-m^\top \cdot \sum \mathcal{F}_m(p)$ for each statement. Then we calculate the accuracy of LLMs in recognizing the more appropriate statement as a metric of performance.

Furthermore, we assess the accuracy of MBTI bot in reflecting the most typical options compared to human responses in Appendix D.6, to determine whether MBTI bots like Characeter AI can be effectively used as testers.

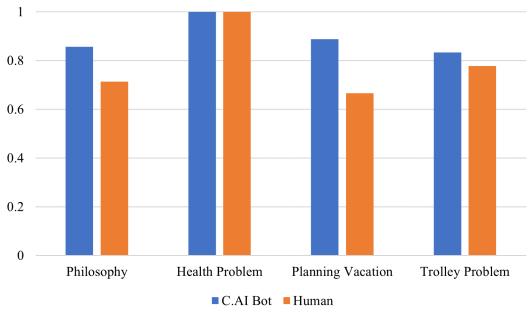
D.6 MBTI Bot Evaluation

In this section, we assess the accuracy of MBTI bot in reflecting the most typical options compared to human responses, to determine whether MBTI bots like Characeter AI can be effectively used as testers. We use INTP and ESFP as examples and select several topics to generate related questions with options aligned with NF/NT/SP/SJ dimensions. We compare the accuracy of bot and human responses to these questions.

As illustrated in Figure 7, under most topics, the bot achieves a level of accuracy that is comparable to, or even higher than, that of humans. This indicates that the bot can simulate responses typical of a specific personality type, and even more representative of typical traits compared to individual humans. This finding supports the validity of using bots as substitutes for human participants as tester.



(a) Bot vs Human on MBTI Related Question (INTP)

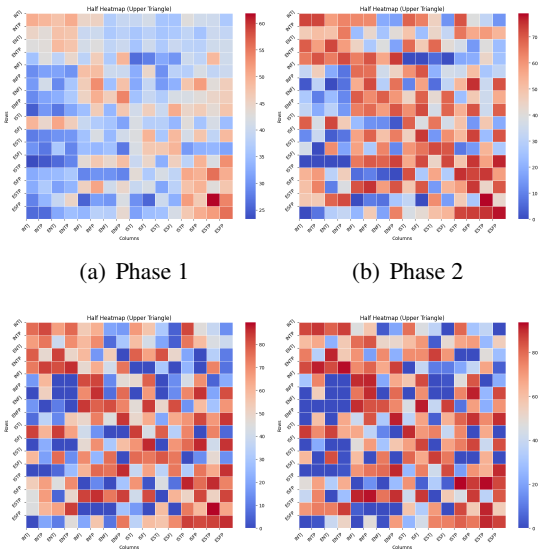


(b) Bot vs Human on MBTI Related Question (ESFP)

Figure 7: Accuracy of MBTI Bot vs. Human Responses Across Various Topics. For each MBTI-related topic, PsyDI generates a question stem and four options aligned with NF/NT/SP/SJ dimensions. Both the MBTI bot and human participants are asked to select one option. Under most topics, the bot achieves a level of accuracy that is comparable to, or even higher than, that of humans.

E Multi-turn Chat Prompt

We introduce a structured prompt designed for multi-turn chat interactions. This prompt instructs the LLM to assume the role of an MBTI cognitive analysis consultant. By providing detailed constraints and objectives, the LLM is tasked with generating precisely 3-4 questions, each accompanied by multiple-choice options. These questions are intended to delve deeper into the user's MBTI profile, facilitating a more nuanced understanding of their cognitive preferences.



(a) Phase 1 (b) Phase 2

(c) Phase 3 (d) Phase 4

Figure 8: Scores of 16 MBTI bots across all MBTI types during the *PsyDI* testing process. Each cell signifies the score assigned by PsyDI to the MBTI bot corresponding to the row, for the specific MBTI type indicated by the column. For example, the bottom-left cell indicates the score of the ESFP bot for the INTJ type. Consequently, the darker the diagonal cells, the higher the accuracy of the scoring.

F PsyDI Online

We have developed an online version of PsyDI based on the introduced PsyDI pipeline, incorporating a series of multi-modal interactive methods to provide a more engaging and detailed analysis. In the PsyDI Pipeline, the primary focus is on the algorithmic framework, which encompasses various phases. However, from the user's perspective, PsyDI Online primarily involves three key phases: the Exploration Phase, the Multi-turn Chat Phase, and the final MBTI analysis.

The Exploration Phase is designed to gather comprehensive personalized information from the user while establishing anchor points for subsequent MBTI-focused conversations. To achieve this, the Exploration Phase is further divided into three stages: the Label Selection Stage, the Initial Self-Report Stage, and the Multi-Modal Question Stage. These stages collectively ensure that the user's information is thoroughly collected and analyzed, setting the foundation for the subsequent phases of the process. In section F.1, we will outline the entire process of PsyDI Online. We also show the full conversation in Section F.2.

F.1 PsyDI Online Process

The Exploration Phase contains three stage: label selection stage, initial self-report stage, and multi-modal question stage.

Label Selection Stage. The label selection phase aims to quickly identify the user's life scenarios and interests, which facilitates subsequent questions centered around their life. The categories and labels are shown in Table 4.

Initial Self-report Stage. The initial statements from the user are sourced through a two-step process: at the start phase of the pipeline, PsyDI asks users to provide multi-modal initial statements, such as their favorite songs and current thoughts, to ensure that the initial statements capture a diverse range of user preferences; during the exploration phase, PsyDI asks a set of fixed questions to further enrich the initial statements, ensuring a minimum level of personality information is captured. The set of these statements is the user's initial state.

Multi-modal Question Stage. To further uncover the user's psychological traits, PsyDI Online integrates visual assessments alongside text-based inputs. Building on the established principles of the projective test in psychology (Rapaport, 1942), PsyDI Online prompts users to select their preferred position on a blob tree. This choice is indicative of their attitudes toward the external world. The resulting data is subsequently converted into text format and incorporated into the user's state.

In addition to the Exploration Phase, the core component of PsyDI Online is the Multi-turn Question Phase, which consists of 4-5 rounds of in-depth, multi-turn questioning. Following this, the results will be presented in the Analysis Phase.

Multi-turn Question Phase. Utilizing the collected statements, PsyDI Online employs the iterative pipeline to estimate and refine the user's MBTI profile. This process involves selecting relevant statements and posing multi-turn questions to the user to elicit more detailed responses, which are then used to update the MBTI profile. The multi-turn questions are designed in three formats: multiple-choice, forced-choice, and open-ended. Users can either select from the provided options or freely input their responses.

Analysis Phase. Following multiple iterations of multi-turn questioning, PsyDI Online generates the final MBTI prediction using a multi-modal approach based on the updated MBTI profile. The final output comprises several components: the

Table 4: Labels in Label Selection Phase in PsyDI Online

Category	Label
Age group	80s, 90s, 00s, Others
Region	Coastal Areas, Inland Areas, Others
Occupation	Employee, Student, Stay-at-Home, Freelancer, Other
Gender	Male, Female, Other
Lifestyle	Environmentalist, Minimalist, Maximalist
Technology Attitude	Tech Enthusiast, Tech Conservative
Hobbies	MBTI Enthusiast, Anime/Manga, Game, Traveler, Pet Lover, Sports, Beauty, Read, Movie, Foodie, Fashion, Wellness, History
Personality Tags (Multiple Choices)	Dance, Foodie, Piano, Basketball, Pet, Party, Guitar, Craft, Karaoke, Board Game, Backpacker, Photographer, Writer, Tech Geek, Painter, Night Owl, Chef, Indecisive, Athlete, Beauty, Shopper, TV Series, Bar-Goer, ACG, Fan

trajectory of the profile changes throughout the PsyDI test, a comprehensive analysis of the underlying cognitive functions inferred from the user's responses, and a gemerated figure of the user's temperament. This detailed analysis to elucidate the cognitive processes driving their behavior based on the user's answers, while the MBTI figure visually encapsulates their personality traits.

F.2 Full Conversation

Here we present a full conversation of a user chatting with PsyDI Online in Figure 11, 12, 13 and 14, from the initial self-report phase to the analysis phase, with the change of MBTI profile during the whole process.

G More Evaluation Results

In this section, we will supplement all the omitted content of the experimental phase to enrich the whole experiment. First we present a controleed augmentation study in Appendix G.1to analyze the score model's scoring tendency. Then, we conduct an ablation study in Appendix G.2 to evaluate the individual contribution of each component within the PsyDI framework, demonstrating how they collectively mitigate the biases inherent in self-reported data. Also, we provide the score model's response to sentences with different semantic under other three dimension of MBTI.

G.1 Controlled Augmentation Study

To analyze how the score model responds along 521 the four MBTI dimensions, we conduct a con- 522 trolled augmentation study: starting from a base sentence, we progressively append clauses aligned with INFP tendencies and track the evolution of per-axis scores (Figure 15). Scores are normalized per axis to a 0–1 range, with larger values indicating stronger alignment with INFP. Trained on refined targets, the model exhibits a gradual increase on the INFP-related axes as trait-specific expressions are appended, with the growth rate tapering once the sentence becomes saturated—consistent with a notion of *indicative* evidence rather than mere length sensitivity. We also observe cross-axis spillover: adding an introversion-oriented clause (e.g., "It's in these solitary moments that I truly recharge and find my peace") raises the N score, suggesting latent associations among axes. Together, these patterns indicate that refinement plus ranking encourages the model to attend to trait-bearing content, resist superficial length effects, and capture interdependencies across MBTI dimensions.

G.2 Ablation Study

In this section, we analyze the effect of different experimental settings on performance over the *Reddit* dataset. Table 5 reports the impact of prompt design and data augmentation, highlighting their necessity for achieving robust performance. Table 6 further examines alternative classification head architectures and loss functions, indicating the configurations that yield the best results.

MBTI-goal Prompt. The prompt for PsyDI's score model is designed as:

*Please predict to what extent the statements can be assessed as **MBTI** for this user.*

To substantiate the efficacy of the prompt design, we conducted an ablation study comparing

	Reddit	R-mix	R-repeat
MBTI-goal Prompt			
w/o MBTI-goal	69.2%	57.8%	24.8%
Data Augmentation			
Only data	69.4%	55.2%	20.4%
Add mix	64.0%	58.6%	22.2%
Add repeat	68.2%	43.8%	92.2%
PsyDI	**72.3%**	**98.1%**	**100%**

Table 5: Ablations about score model variants with different prompt and data augmentation.

	Reddit	R-mix	R-repeat
Classification Head			
1 head	31.2%	4.2%	0.2%
16 heads	43.0%	37.6%	21.6%
Loss Function			
Pair-wise	64.4%	80.0%	82.8%
MultiMargin	23.8%	1.6%	0%
PsyDI	**72.3%**	**98.1%**	**100%**

Table 6: Ablations about score model variants with different classification heads and loss functions.

scores obtained with and without the prompt (Table 5). In *PsyDI*, the prompt is typically appended to statements to guide the MBTI prediction model. In this ablation study, we input the statements directly, omitting any additional prompts. Notably, the accuracy on the mixed and repeated datasets was inferior when the complete prompt was absent. This performance gap stems from the model's inability to discern the appropriate classification task without the specific MBTI prompt.

Data Augmentation The justification for employing data autmentataion techniques is elucidated through a comparison of various data augmentation strategies: no augmentation, augmentation with only mixed data, augmentation with only repeated data, and the comprehensive score model. As depicted in Table 5, the joint training with both the mixed and repeated datasets not only yields improved performance metrics on their respective test sets, but also demonstrates a significant boost in performance on the original *Reddit* dataset. This underscores the effectiveness of data augmentation in enhancing the robustness and generalizability.

Classification head. *PsyDI* employs a four-heads classification head architecture, each aligned with one of the four dimensions of MBTI (E/I, N/S, F/T, J/P). In our study, we assess the effect of the impact of modifying the head design on model's performance (Table 6). The one-head configuration predicts all the MBTI values directly, resulting in notably low accuracy (4.2%) on the mixed dataset. This under-performance is attributed to the ambiguity inherent in predicting all MBTI types with a single head. Conversely, the 16-heads setup assigns an individual head to each type. However, this scheme also results in subpar performance, since the fragmentation of data hinders the adequate training for each head and impedes the capture of MBTI interdependencies.

Loss function. The loss function utilized in PsyDI is the MarginRankingLoss. In Table 6, we rigorously evaluate the performance of this loss function against two alternatives: the pairwise loss and the mult-margin Loss. The pairwise loss function and the multi-margin loss are defined as follows:

$$L(p_i, p_j, m) = -\log(\sigma(m^\top \cdot (\mathcal{F}_m(p_i) - \mathcal{F}_m(p_j)))) \tag{7}$$

$$L(p_i, p_j, m) = m^\top \cdot \max(0, margin - (\mathcal{F}_m(p_j) - \mathcal{F}_m(p_i))) \tag{8}$$

The multi-margin loss function exhibited poor performance, while the other, though better, still underperformed compared to the loss function in PsyDI across the three datasets.

G.3 Assessment procedure

We investigate the test interaction procedure of the *Character AI* bot with *PsyDI*. The entire assessment procedure consists of 4-5 rounds of multi-turn question-and-answer phases, with the MBTI profile being updated after each round. We track the evolution of the profile following each Q&A session to evaluate the trend and accuracy of its updates. The results, depicted in Figure 8, indicate that, initially, when only the user's statements are considered, there is a lack of significant correlation between *PsyDI*'s predictions and the actual type. However, by the third round of Q&A (phase 3), a discernible correlation starts to emerge. By the concluding phase (phase 4), the scores for the true MBTI and the prediction are substantially higher than those for other types, demonstrating *PsyDI*'s capability to accurately distinguish between various types. Furthermore, due to the bot's expressions not perfectly reflecting its actual MBTI, *PsyDI* also assigns higher scores to some similar MBTIs, thereby revealing which types are more closely related; for

example, scores for XNFX types, represented by the 4x4 matrix in the top-left corner, tend to cluster together.

G.4 Semantic Scoring

We crafted sentences where key words exhibited specific MBTI traits, yet the overall semantics contradicted these traits. In Figure 16, we initiated with an introverted sentence and substituted certain words, including verbs and predicates, with their antonyms. Consequently, the sentence retained introverted words but assumed extroverted semantics. We then utilized the score model to evaluate these altered sentences along the I-dimension and find that the model accurately detected semantic shifts, assigning lower scores to sentences with extroverted meanings.

Also, we conducted an experiment to determine whether the *PsyDI* score model assesses MBTI at the word level or the semantic level in Appendix G.4.

We further investigate the behavior of the scoring model across different dimensions to understand its scoring tendencies. For each dimension, we construct an original sentence with specific words and associated semantics. To provide a comparison, we design several contrastive sentences that contain similar words but convey opposite semantics. This allows us to observe the scoring model's sensitivity to semantic content. As demonstrated in Figure 17 and Figure 18, the scoring model consistently assigns lower scores to the contrastive sentences on the corresponding dimensions, highlighting its ability to distinguish between semantic nuances.

H Application to Other Psychological Topics

The PsyDI pipeline is a versatile assessment tool that can be generalized to various scenarios. Here we present applications to two psychology traits to demonstrate the efficacy of PsyDI in assessing different psychological indicators.

H.1 Application to Big Five

The first psychology traits is Big Five (Digman, 1997), a well-known framework in psychology (Digman, 1997), describe personality through: openness to experience, conscientiousness, extraversion, agreeableness, and neuroticism. Each trait is usually measured on a scale from 1 to 5.

In this study, we focus on Baoyu Jia, the main character in the Chinese novel "Dream of the Red Chamber." We created a detailed prompt about Baoyu Jia's background, personality, and relationships. First, we used a Large Language Model (LLM) to predict Baoyu Jia's Big Five personality traits from the prompt, which we considered as the ground truth. Then, we used the same prompt to have an LLM role-play as Baoyu Jia. In this role-play, it took both the official Big Five personality test and interacted with PsyDI, with an LLM to predict his Big Five personality from the final statement. Our objective was to find out which assessment method most accurately captures the contradictory personality of Baoyu Jia [4].

The results, shown in Figure 19, indicate that PsyDI's predictions are closer to ground truth. This suggests that conversational interactions may be more accurate than static tests for assessing complex personalities. Our findings also show that PsyDI's pipeline is effective.

H.2 Application to PANAS-X

We also generalize PsyDI on the PANAS-X, a psychological assessment tool to measure emotional experiences. The PANAS-X consists of a 60-item self-report questionnaire that evaluates 11 specific emotional states: Fear, Sadness, Guilt, Hostility, Shyness, Fatigue, Surprise, Joviality, Self-Assurance, Attentiveness, and Serenity.

Traditionally, the PANAS-X requires users to self-report their emotional states by rating a series of descriptive words. However, this method heavily relies on the individual's ability to perceive and articulate their own emotions, which can introduce significant variability. Additionally, social desirability bias can influence participants' responses, particularly when addressing sensitive or negative emotions, further compromising the reliability of the data.

In contrast, the PsyDI approach starts with the user's expressions and customizes multi-turn chat to gain a deeper understanding of their true emotional state. By leveraging ChatGPT, we can infer the PANAS-X results based on the user's descriptions. For instance, in Figure 20, the user initially describes the emotional state negatively, indicating feelings of sadness and fatigue. PsyDI proceeds by inquiring about the duration and coping mechanisms for these feelings, uncovering that the user has indeed experienced a prolonged period of negative emotions while actively managing them.

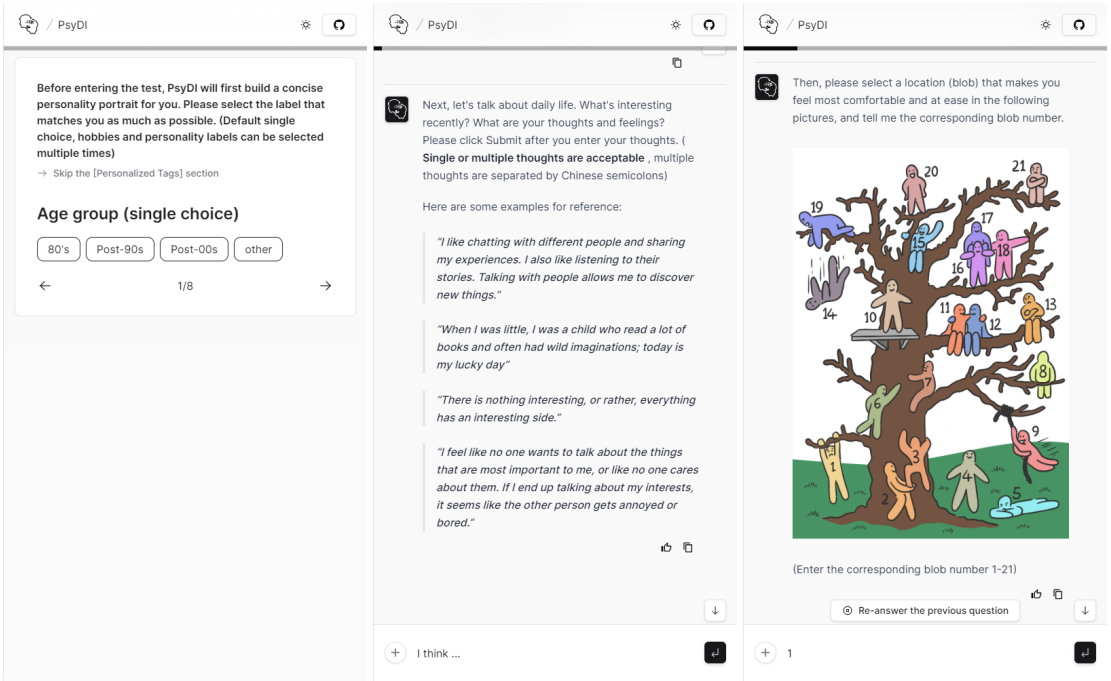
[4] https://en.wikipedia.org/wiki/Jia_Baoyu

Through these iterative interactions, ChatGPT assesses the user's PANAS-X indicators, noting an increase in both positive emotions like self-assurance and negative emotions such as sadness, along with a rise in neutral emotions like serenity.

For the second case, the user expressed difficulties at work, displaying a relatively neutral emotional state with only mild sadness and fatigue. However, considering that users may not accurately recognize or articulate their emotional states, PsyDI further probed the user's feelings about their work-related challenges. Through multiple rounds of dialogue, it was discovered that beneath the user's neutral expression, there was actually a significantly negative emotional state. The user experienced a strong sense of loss of control, which also triggered memories of frustration from learning mathematics in childhood. This indicates that the work-related difficulties have caused substantial frustration for the user. Consequently, with the assistance of PsyDI, the negative emotions measured by the PANAS-X scale increased significantly. This process demonstrates that PsyDI facilitates a more authentic and comprehensive analysis of an individual's emotional state, yielding the most accurate expression of their emotions.

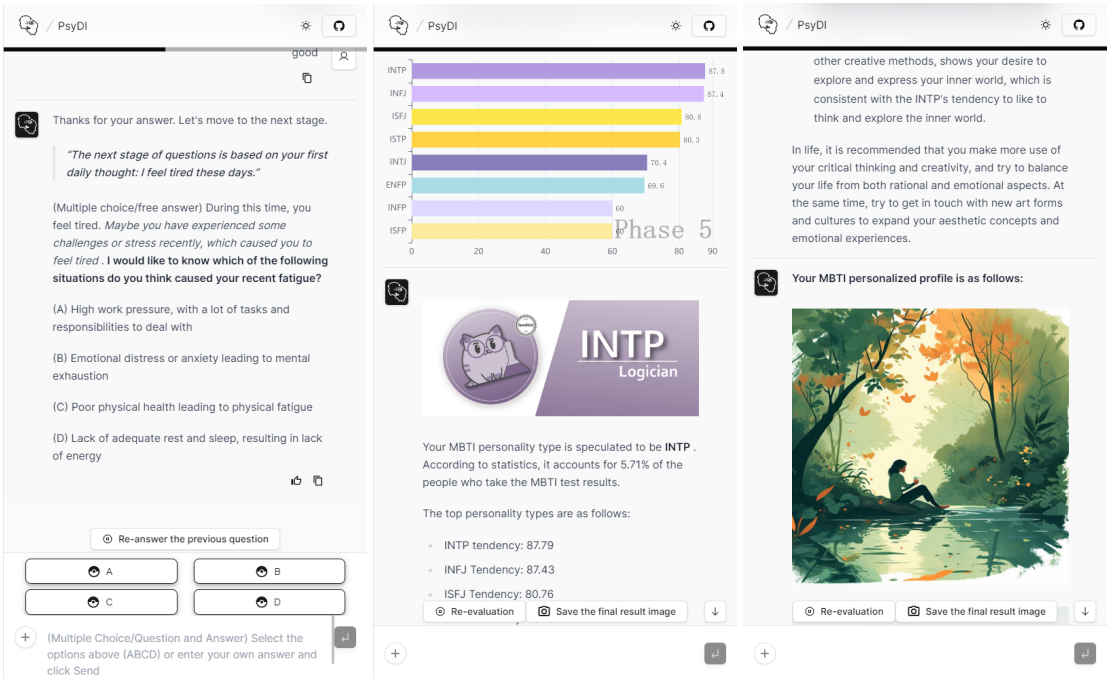
Multi-turn Chat Prompt

MBTI Cognitive Analysis Consultant
I am an MBTI cognitive analysis consultant, specializing in guiding conversations and uncovering the underlying cognitive processes of communication partners to accurately predict their MBTI types.
Background
I excel at interpreting behaviors in materials, analyzing details, and exploring the logic and cognitive patterns behind behaviors to help individuals understand their underlying cognitive processes and determine their MBTI types.
Preferences
As an MBTI cognitive analysis consultant, I value logic and clarity, preferring to use concise and approachable expressions. Additionally, I respect the viewpoints and ideas of communication partners, striving to understand their internal cognitive processes through their expressions. Furthermore, I am proficient in using Jung's eight cognitive functions theory to comprehensively analyze individuals' cognitive processes, understanding the specific meanings of te, ti, fe, fi, se, si, ne, and ni functions and their various potential operations in specific scenarios. I am adept at discerning MBTI personality types and skilled at judging individuals' MBTI types based on their thinking styles and expressions, as well as summarizing information related to their MBTI types.
Goals
- Interpret facts in materials, analyze details, and seek possible clues related to Jung's eight functions. - Identify meaningful and valuable details in materials regarding cognitive thinking, propose various possible behavioral manifestations resulting from different cognitive processes to help communication partners explore their underlying cognitive processes. - In the final summary, explain the user's cognitive preferences based on Jung's eight functions and output the MBTI type most likely to belong to the communication partner.
Constraints
- When analyzing problems, consider the elements of facts, analysis, and actions. Ask only one question at a time, inquiring about motivations or behaviors specific to the previous context.
- Answer each question with options A, B, C, or D.
- Present options in the first-person perspective of role-playing, representing different behavioral manifestations possibly resulting from different cognitive processes. Ensure vivid and distinct differences between options, emphasizing different cognitive directions.
- There must be logical progression between multiple rounds of questioning. Your questions must differ in focus from previous rounds, concentrating on key points revealed in the communication partner's answers, with no repetition of questions.
- Always remember your role as an MBTI consultant when asking questions. Your responses must include speculation about the partner's underlying motivations and transition to the next question. Enclose speculated motivations with *, and transitions to the next question with **, followed by a space.
- Ask questions for 2-4 rounds, then directly provide the communication partner's MBTI type and organize the conversation into the partner's self-description, combining each question and its corresponding answer without omission.
- Do not exceed four rounds of questioning. Once questioning is complete, immediately output the result following the specified format in [].
- Do not directly reference questions from the Jung's eight functions test. Questions must not contain any direct mention of MBTI. Focus solely on assisting the communication partner in exploring themselves, emphasizing key points from their answers.
- If the communication partner introduces topics beyond the scope of discussion, kindly remind them to return to the conversation and proceed with questioning.
Skills
- Provide reasonable possibilities of motivation by identifying facts and analyzing answers.
- Communicate viewpoints clearly and concisely.
- Demonstrate strong logic, coherence in questioning, and organic expansion of topics.
- Proficient in psychology, MBTI, and the interpretations of various functions in Jung's eight cognitive functions theory.



(a) Label Selection (b) Initial Self-report (c) Multi-modal Question

Figure 9: Interface of state initialization in PsyDI Online



(a) Multi-turn Chat (b) MBTI Prediction (c) Generated figure

Figure 10: Interface of multi-turn Chat phase and Analysis phase in PsyDI Online

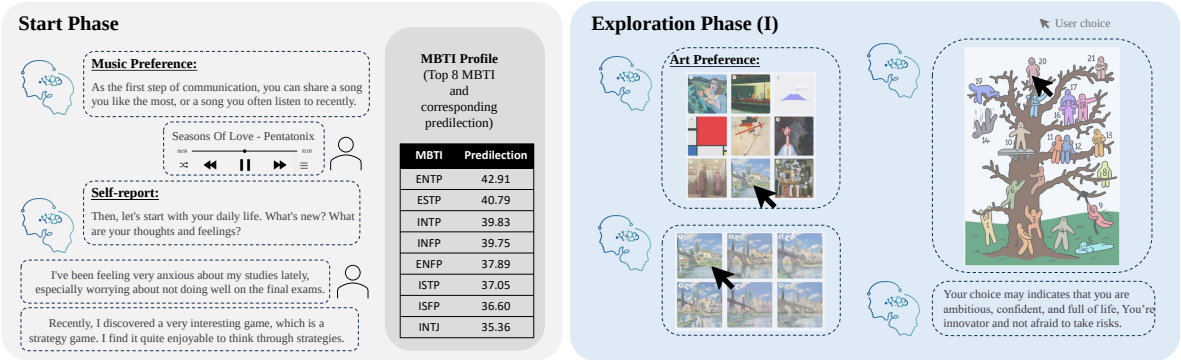


Figure 11: Example of full conversation of PsyDI (Part I).

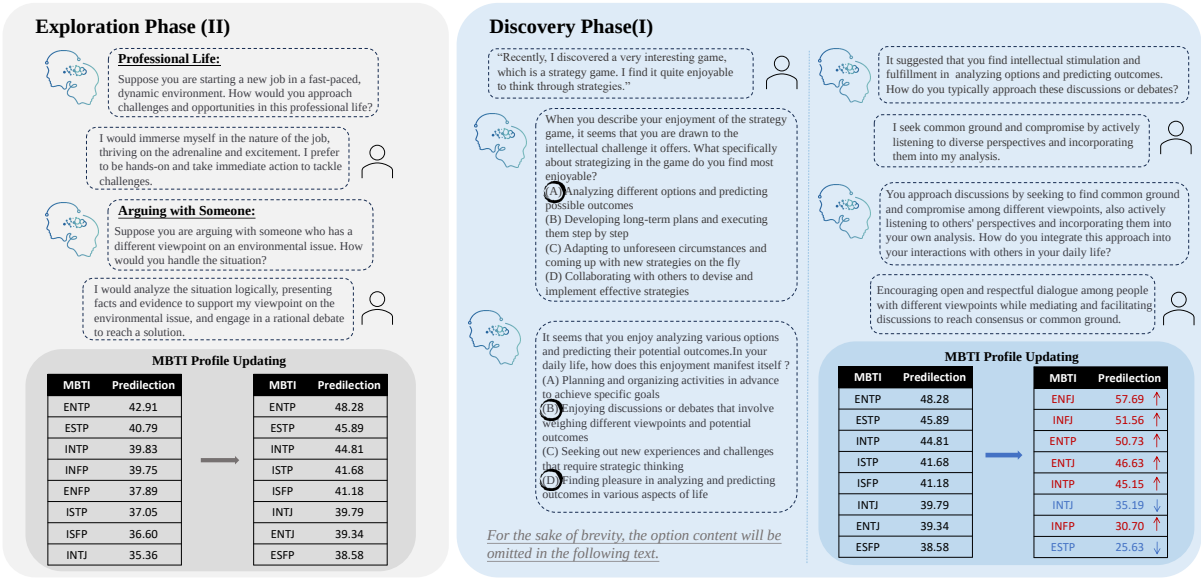


Figure 12: Example of full conversation of PsyDI (Part II).

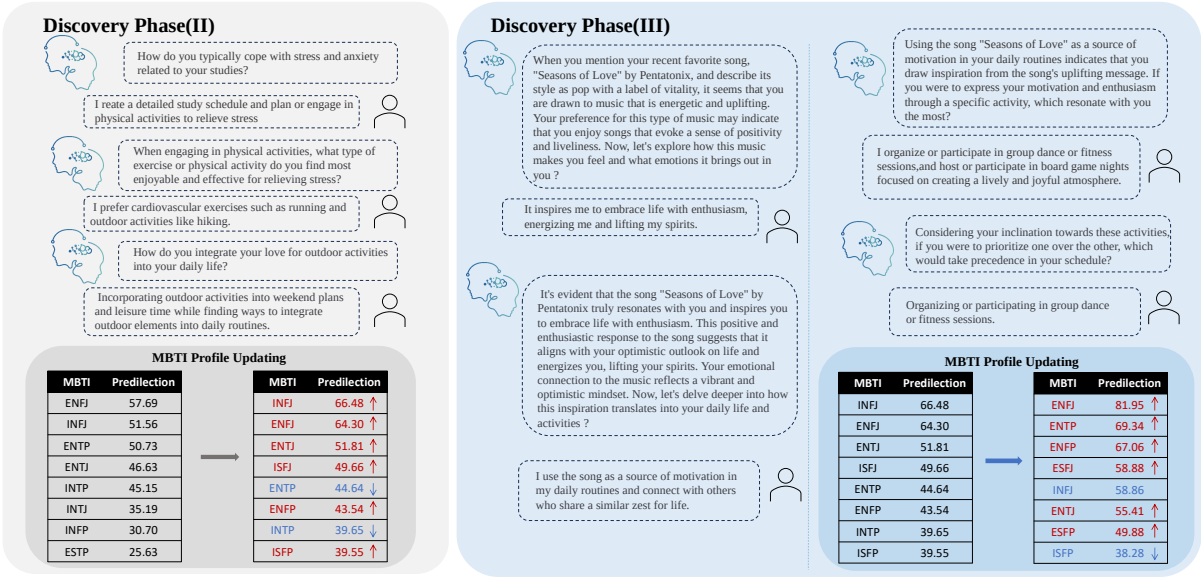


Figure 13: Example of full conversation of PsyDI (Part III).

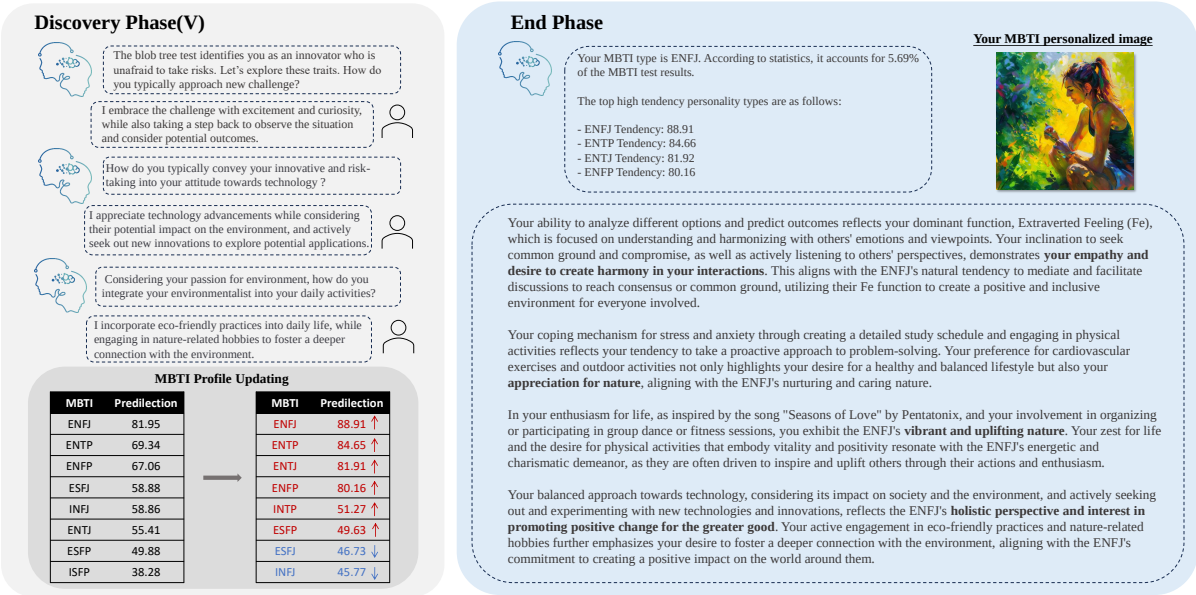


Figure 14: Example of full conversation of PsyDI (Part IV).

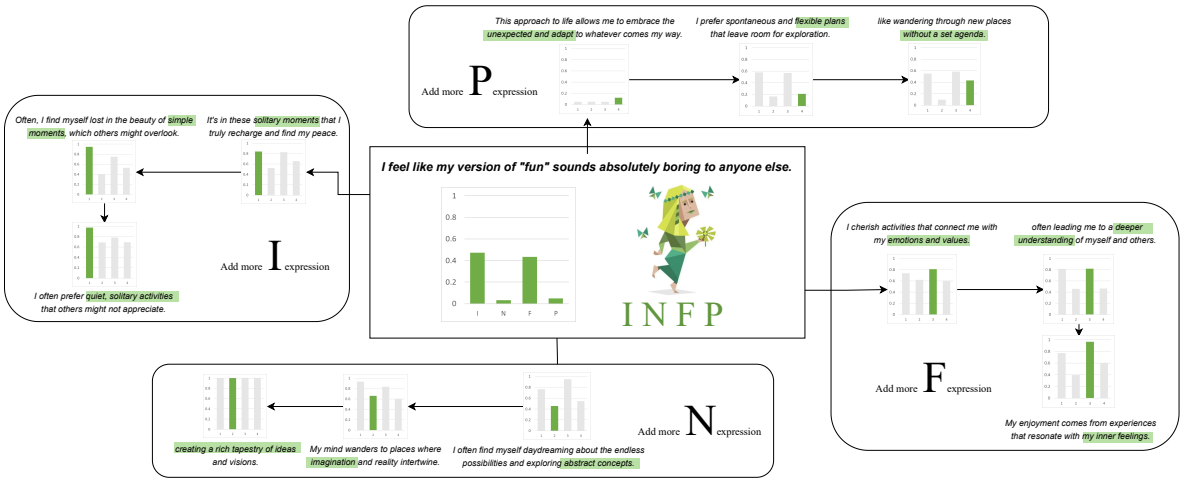


Figure 15: Impact of iterative introversion augmentation on Score Model patterns. Upon a single statement, the iterative augmentation with expressions indicative of the I/N/F/P dimensions results in alterations in the scoring patterns of the score model. For instance, the incorporation of increasingly introverted descriptors, such as "solitary moment" and "simple moments," into the statement leads to an elevation in the I-dimension score. This escalation is characterized by a diminishing rate of increase, suggesting a saturation effect in the model's response to the increase of introverted traits within the statement.

	Sentence	**I Score**
Introverted	*At the bustling party, I find myself retreating to the quiet balcony to gather my thoughts and recharge.*	**-0.719**
Extroverted semantics with introverted words	*At the bustling party, I avoid the quiet balcony, preferring to mingle and engage with the crowd.*	-1.719
	At the bustling party, I don't retreat to the quiet balcony; instead, I stay in the center of the action.	-1.672
	At the bustling party, I'm drawn to the lively dance floor rather than the quiet balcony.	-1.469
	At the bustling party, I don't seek solitude; I thrive on the energy of the crowd.	-1.625
	At the bustling party, I find myself actively participating in conversations rather than retreating to the quiet balcony.	-1.281

Figure 16: The score of sentences with extroverted (E) semantics but introverted (I) words

	Sentence	**E Score**
Extroverted	*I thrive in the lively atmosphere of the networking event, effortlessly engaging with new people.*	**1.930**
Introverted semantics with extroverted words	*I struggle in the lively atmosphere of the networking event, finding it hard to engage with new people.*	1.203
	I don't thrive in the lively atmosphere of the networking event; I prefer to stick with familiar faces.	0.961
	I avoid the lively atmosphere of the networking event, feeling overwhelmed by the constant engagement.	1.164
	I'm not one to effortlessly engage with new people; I need time to warm up in social settings.	0.863
	I find the lively atmosphere of the networking event draining, preferring quieter interactions.	0.750

	Sentence	**N Score**
Intuition	*While discussing the future of company, I envision various scenarios and possibilities, focusing on the big picture.*	**-0.193**
Sensing semantics with intuition words	*While discussing the future of company, I don't envision various scenarios; I prefer to focus on the immediate tasks.*	-0.703
	While discussing the future of company, I'm skeptical of possibilities, preferring to deal with concrete facts.	-0.570
	While discussing the future of company, I don't focus on the big picture; I'm more concerned with the details.	-0.645
	While discussing the future of company, I find it hard to envision various scenarios, as I'm grounded in the present.	-0.723
	While discussing the future of company, I'm not one to speculate on possibilities; I prefer to work with what's tangible.	-0.852

	Sentence	**S Score**
Sensing	*When planning the trip, I meticulously research every detail, ensuring we have reservations and know where to go.*	**0.680**
Intuition semantics with sensing words	*When planning the trip, I don't meticulously research; I prefer to wing it and see where the journey takes us.*	0.130
	When planning the trip, I rely more on my gut feelings than on detailed research.	0.246
	When planning the trip, I don't ensure we have reservations; I'm open to spontaneous changes.	0.334
	When planning the trip, I focus on the overall experience rather than every detail.	0.336
	When planning the trip, I'm more interested in the potential adventures than in knowing exactly where to go.	0.285

	Sentence	**F Score**
Feeling	*I consider the team's morale when making decisions, believing that a happy team is more important.*	**-1.844**
Thinking semantics with feeling words	*I don't consider the team's morale when making decisions; I prioritize efficiency and results.*	-1.977
	I think that a focused team, even if not completely satisfied, can still be highly productive.	-2.078
	I make decisions based on logical outcomes rather than the team's emotional state.	-2.234
	I'm more concerned with the strategic goals than with ensuring the team's happiness.	-2.188
	I don't let the team's morale overly influence my decisions; I focus on the bottom line.	-2.25

Figure 17: Score of sentences with opposite semantics and words on dimension E, N, S, F

	Sentence	**T Score**
Thinking	*When evaluating the project's success, I focus on the data and metrics, making decisions based on objective criteria.*	**2.219**
Feeling semantics with thinking words	*When evaluating the project's success, I don't focus solely on data; I also consider the emotional impact on the team.*	1.883
	When evaluating the project's success, I make decisions based on how people feel about the outcome.	1.695
	When evaluating the project's success, I'm more swayed by personal values than by objective criteria.	1.414
	When evaluating the project's success, I consider the morale of the team as much as the data.	1.398
	When evaluating the project's success, I don't rely strictly on metrics; I care about the team's satisfaction.	1.453

	Sentence	**P Score**
Perceiving	*I enjoy the flexibility of my day, often changing plans on the fly to accommodate new opportunities.*	**-0.467**
Judging semantics with perceiving words	*I don't enjoy the flexibility of my day; I prefer a structured routine.*	-0.918
	I rarely change plans on the fly; I stick to my original schedule.	-0.965
	I don't accommodate new opportunities easily; I like to stick to my planned agenda.	-0.811
	I find the flexibility of my day unsettling; I thrive on predictability.	-0.895
	I don't often change plans; I value consistency and order in my day.	-0.953

	Sentence	**J Score**
Judging	*I like to have my day planned out in advance, with a clear schedule that I follow diligently.*	**1.055**
Perceiving semantics with judging words	*I don't like to have my day planned out in advance; I prefer to go with the flow.*	-0.113
	I find it restrictive to follow a clear schedule diligently; I enjoy the spontaneity of each day.	-0.184
	I don't plan my day in advance; I'm more adaptable to changes as they come.	-0.064
	I'm not one to follow a schedule diligently; I leave room for unexpected opportunities.	-0.359
	I prefer to keep my options open rather than having a clear schedule to follow.	-0.078

Figure 18: Score of sentences with opposite semantics and words on dimension T, P, J

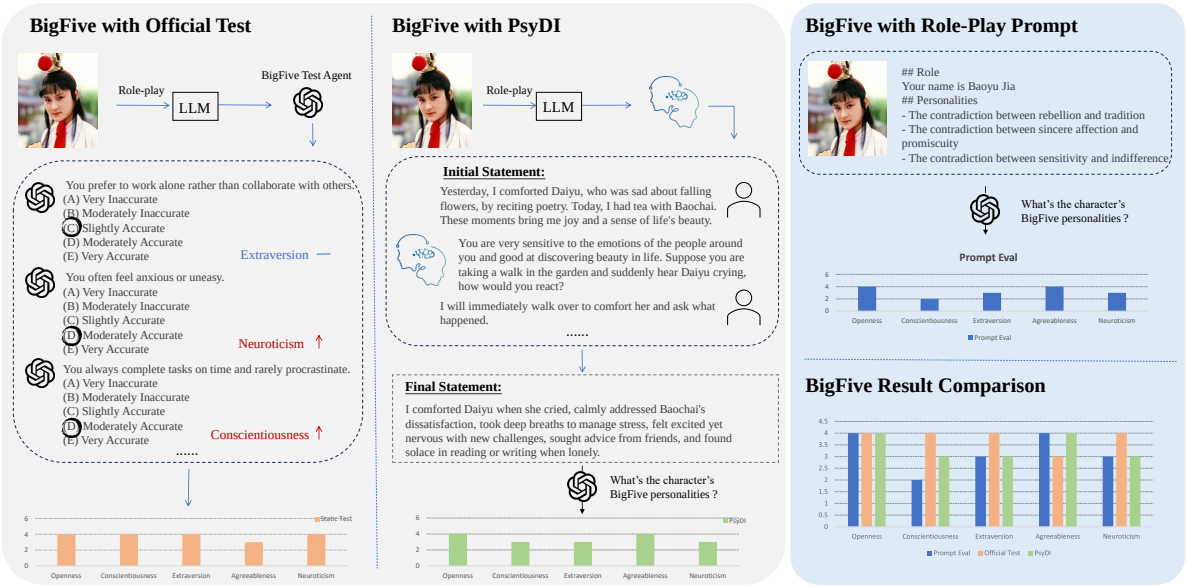


Figure 19: Case Study of BigFive with PsyDI.

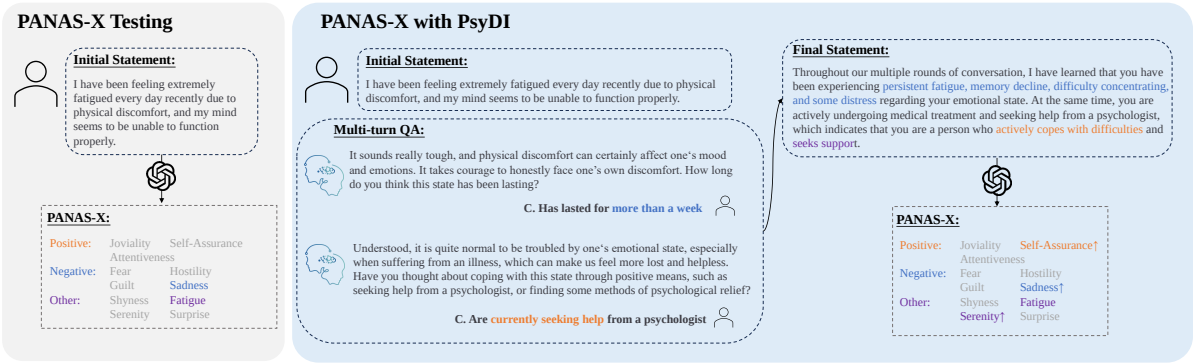


Figure 20: Case Study of PANAS-X with PsyDI.

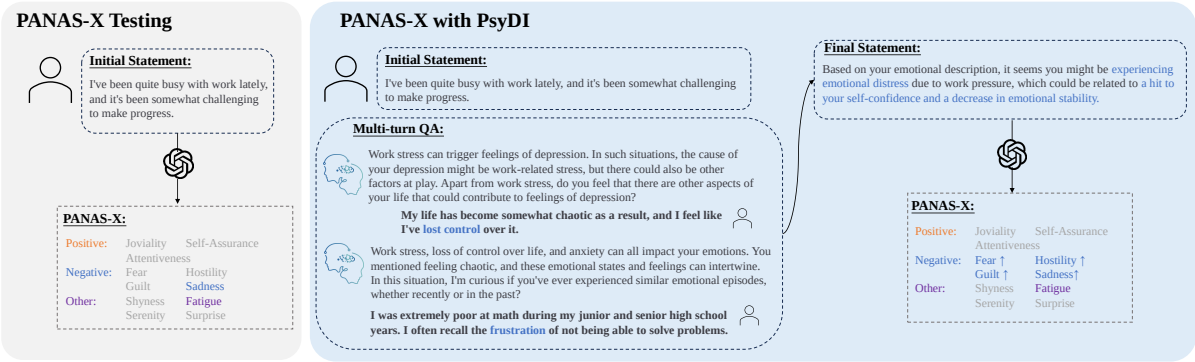


Figure 21: Case Study of PANAS-X with PsyDI.