PsyPath: Psychologically-guided Self-Exploration for Personality Detection

Anonymous ACL submission

Abstract

Personality detection aims to label an individual's traits via identifying linguistic cues from his or her written text. Previous approaches typically perform a direct mapping between text and trait labels or apply static reasoning to this task. In this paper, we argue that dynamic reasoning, underpinned by psychological theory, is essential for personality trait inference. To address this, we propose PsyPath, a novel framework that models personality detection as a process of psychologically-guided self-exploration. By enabling large language models (LLMs) to dynamically generate and answer psychologically meaningful questions, our method creates a dynamic reasoning path to explore the underlying dimensions of personality traits. This mechanism not only makes the reasoning process transparent, but also helps the model understand personality nuances in a way that mirrors expert psychological reasoning. For the "guided self-exploration", we propose a novel hybrid scoring mechanism to step-by-step evaluate the generated nodes in the reasoning paths that balances psychological coherence (black-box scoring) and model output dynamics (white-box scoring). This reasoning-based formulation inherently reflects how psychologists assess personality, as they rely on iterative, diagnostic reasoning. Experiments on two benchmark datasets demonstrate that PsyPath consistently outperforms strong baselines, yielding improvements in predictive accuracy and model interpretability. Moreover, the generated reasoning paths provide psychologically meaningful training data, significantly improving performance and psychologically grounded interpretability in downstream tasks.

1 Introduction

Personality, defined as enduring patterns in how individuals behave, think, and feel, is a core construct in psychological science (John et al., 2008). Personality detection from text is a fundamental challenge at the intersection of psychology and natural language processing, increasing application value and profound impact across various domains, including personalized recommendation (Lu and Kan, 2023), healthcare (Ramachandran et al., 2020), marketing (Meskelis and Whittington, 2020), political targeting (Zarouali et al., 2024), and human-computer interaction (Sharan and Romano, 2020). Due to its wide impact and inherent challenges, personality detection has recently seen growing attention, especially in identifying MBTI and OCEAN traits.

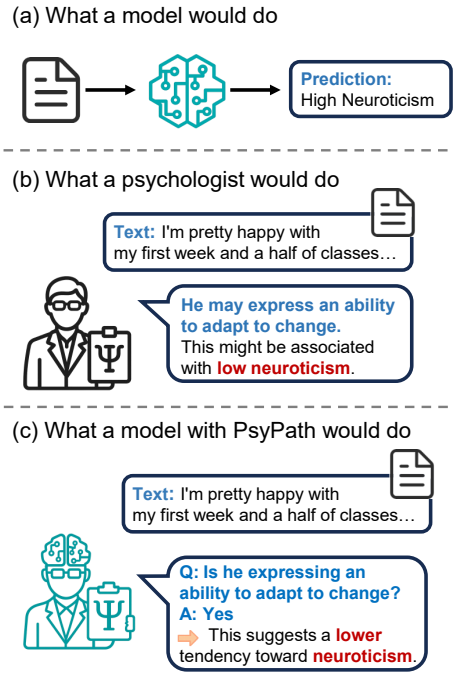

Figure 1: Comparison of personality detection paradigms. (a) Traditional models predict traits directly from text, (b) human psychologists reason via constructs like adaptability. (c) PsyPath mimics human psychologists by prompting reasoning through step-by-step QAs.

Traditional approaches typically frame personality detection as a classification problem, attempting to perform a direct mapping between text and trait labels, as illustrated in Figure 1(a). This paradigm

spans a wide spectrum of methodologies, from early psycholinguistic feature-based models (Francis and Booth, 1993; Cui and Qi, 2017) to sophisticated pre-trained language models (Devlin et al., 2019; Liu et al., 2019). While recent methods (Yang et al., 2023; Hu et al., 2024; Li et al., 2024) have attempted to leverage predefined reasoning templates (e.g., fixed questionnaires), their reasoning logic remains largely simple and static, failing to capture the dynamic nature of psychological assessment. As illustrated in Figure 1(b), human psychologists apply their explicit and implicit psychological knowledge in intermediate analysis and multi-step reasoning to derive conclusions and reach final judgments (Vertue and Haig, 2008; Meehl, 1954). However, existing methods rarely incorporate such a psychological reasoning process into the inference of personality traits. To simulate human psychologists, there exist two key problems: (1) how to effectively model the psychological reasoning process; (2) how to embed the psychological reasoning into personality detection.

For the first problem, drawing on insights from clinical psychology (Vertue and Haig, 2008; Meehl, 1954), we treat psychological reasoning as a dynamic diagnostic process, in which reasoning evolves through iterative question-answering (Q&A) guided by psychological knowledge. This mechanism enables the model to construct step-by-step reasoning trajectories composed of diagnostic Q&A steps, aligning with how human psychologists analyze personality traits. Technically, to avoid the expense of human labeling, we implement this process via Monte Carlo Tree Search (MCTS), which guides the model through a process of self-exploration to generate diagnostic paths by optimizing a hybrid reward that balances psychological coherence (black-box scoring) and model output dynamics (white-box scoring).

It is worth highlighting that the black-box and white-box scores form a complementary mechanism: the former provides a psychologically grounded structural evaluation, while the latter captures probabilistic shifts in model prediction, together ensuring both psychological plausibility and computational effectiveness. The black-box score offers an external assessment rooted in psychological theory, evaluating whether the reasoning path aligns with core clinical constructs in personality assessment. It incorporates three key dimensions: (1) behavioral consistency, (2) cognitive processing styles, and (3) underlying motivations and self-concept—dimensions that mirror how human psychologists interpret trait expression. Serving as a complement, the white-box score captures internal model signals by comparing the prediction probability of the correct label before and after applying the reasoning path, helping to identify which questions or paths most effectively improve model performance. Together, these two scores ensure that the generated reasoning paths are both theoretically grounded and practically impactful.

For the second problem, we adopt Direct Preference Optimization (DPO) to embed psychologically guided reasoning. Since Monte Carlo Tree Search (MCTS) has already explored the search space, DPO can efficiently distill the outcomes of this exploration, thereby avoiding the costly online sampling processes required by RL methods such as GRPO. Crucially, unlike tasks in mathematics or programming, psychological reasoning lacks deterministic verification rules. Although sampling data in such open-ended domains is susceptible to high variance, MCTS mitigates this issue by aggregating evaluation results from multiple simulations. Therefore, leveraging static, high-confidence preference pairs generated by MCTS for DPO optimization is more robust than relying on online sampled data. This approach ensures the stable internalization of the reasoning structure.

Our experiments show that PsyPath significantly outperforms prior approaches in predictive performance. On the Kaggle MBTI dataset, PsyPath outperforms previous works, improving average accuracy (Acc.) by **12%** and the F1-score (F1) by **15.6%**. Similarly, on the Essays dataset, it achieves gains of **3.6%** in accuracy and **7.2%** in F1-score. These results fully highlight the remarkable value of our method's structured, psychology-driven reasoning mechanism for enhancing accuracy and interpretability in computational personality detection. To summarize, this work makes the following three key contributions:

- We introduce a novel paradigm of leveraging psychological principles to enable interpretive reasoning in LLMs for personality detection.
- We instantiate this paradigm through a technical framework that employs iterative self-exploration guided by a hybrid reward, balancing psychological coherence (black-box) and model dynamics (white-box) to generate diagnostic paths.

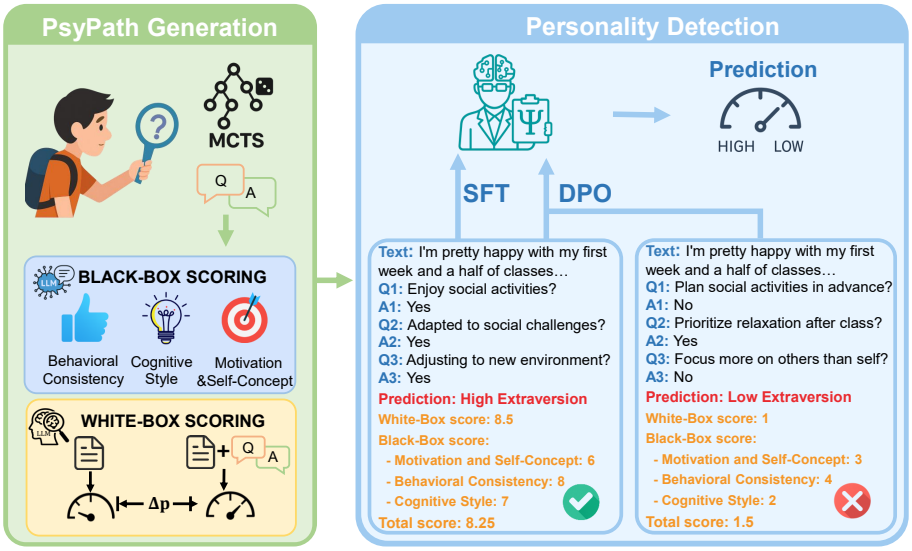

Figure 2: Overview of the PsyPath framework. PsyPath is mainly composed of two phases: the Generation phase uses Monte Carlo Tree Search (MCTS) with a hybrid scoring mechanism to produce psychologically guided Q&A paths. The Personality Detection phase uses the high-quality paths as training data to fine-tune an LLM.

- Under a unified evaluation protocol, our approach achieved competitiveness with state-of-the-art general-purpose LLMs while also significantly surpassing traditional baselines.

2 Model Overview

We formalize text-based personality detection as a psychologically guided reasoning process. Given a collection of user-authored texts $\mathbf{x} = \{s_1, s_2, \ldots, s_n\}$, where s_i is a textual utterance, the goal is to infer a personality label $y \in \mathcal{Y}$ (e.g., Big Five: Extraversion, Openness, Conscientiousness, Agreeableness, Neuroticism). Grounded in psychological theory (Fleeson, 2001; Epstein, 1994; Rogers et al., 1959), we model it as a diagnostic reasoning process involving behavioral patterns, cognitive styles, and motivational factors.

As shown in Figure 2, we design PsyPath as a two-stage framework inspired by how psychologists infer personality through structured reasoning. In the first stage, the framework applies Monte Carlo Tree Search (MCTS) to iteratively generate Q&A paths, guiding a language model to explore key behavioral, cognitive, and motivational signals from the text. These paths are designed to be interpretable and psychologically grounded. In the second stage, the generated paths are used as supervision to fine-tune a language model. By learning from these structured examples, the model acquires the ability to perform multi-step reasoning and produce transparent trait predictions.

3 MCTS-based Self-Exploration

We describe the first stage of our framework which applies MCTS to self-explore the diagnostic reasoning process of psychologists. Given a text and a target trait, MCTS guides a generative model to iteratively pose and answer relevant questions, forming a structured Q&A path. The search is constrained by a fixed depth and number of simulations, and proceeds in four steps: Selection, Expansion, Simulation, and Backpropagation.

3.1 Selection

The MCTS process starts from the root node (the input text) and selects child nodes (candidate next-step Q&A pairs) based on the Upper Confidence Bound for Trees (UCT) score:

$$U(v) = \frac{Q(v)}{N(v)} + C\sqrt{\frac{\ln N(p)}{N(v)}} \quad (1)$$

Here $Q(v)$ is the total score of node v, $N(v)$ and $N(p)$ are the visit counts of v and its parent, and C controls the trade-off between exploration and exploitation. The selection continues until reaching an unexpanded leaf or the maximum depth.

3.2 Expansion

In the Expansion phase, once a leaf node is selected, the LLM generates yes/no questions related to the target trait. To maintain diversity, the previously used questions are excluded using a dynamic

question history. The LLM also answers each question based on the user's text. Here we design the prompts to elicit questions which can cover the three key dimensions including: (1) Behavioral Consistency: stable patterns across contexts and time (Fleeson, 2001); (2) Cognitive Style: typical modes of processing and judgment (Epstein, 1994); and (3) Motivation and Self-Concept Values: inner drives, and beliefs shaping trait expression (Rogers et al., 1959; Schwartz, 1992). These factors align with earlier psychological principles and the black-box scoring criteria (Figure 2). If the LLM fails to generate enough valid questions, a fallback mechanism completes the path.

3.3 Simulation

The simulation step scores each root-to-leaf Q&A path with a Hybrid Scoring Mechanism (black-box plus white-box) to guide the MCTS search.

The Black-box scoring operationalizes psychological validity with an LLM rater conditioned on the author's texts and the candidate Q&A path. Using a scoring criterion aligned with the three dimensions introduced in Section 3.2, the rater outputs a score from 1 to 10 for each dimension. The black-box score takes the average of these three dimension scores (see Appendix C for the scoring prompt).

The White-Box scoring measures the internal prediction confidence of the LLM. We quantify confidence shift as the probability difference without/with Q&A Path. That is, $\Delta P = P_{\text{aug}} - P_{\text{base}}$, where $P_{\text{base}} = P(y \mid \text{Text})$ and $P_{\text{aug}} = P(y \mid \text{Text}, \text{Q\&A Path})$. Both probabilities are computed with a fixed generator Llama-3-8B-Instruct as the softmax probability of the verbalized ground-truth label token (e.g., `high`/`low`). Further, we map ΔP to a bounded white-box score via a scaled sigmoid:

$$S_{\text{white}} = \text{clip}\Big(\lambda_1\big(2\sigma(k\Delta P) - 1\big) + \lambda_2 + \beta \cdot \mathbb{I}(P_{\text{aug}} \geq \tau),\ s_{\min}, s_{\max}\Big), \tag{2}$$

Here $\sigma(\cdot)$ is the sigmoid, k controls sensitivity, λ sets the score range, β rewards high posterior confidence, and the result is clipped to $[s_{\min}, s_{\max}]$ to align with the black-box score for the weighted combination. Using Llama-3 as a fixed external evaluator keeps white-box scoring separate from the downstream detector Qwen2-7B, preventing evaluator-detector coupling and ensuring stable search rewards. The final reward $S_{\text{final}} = \alpha S_{\text{white}} + (1-\alpha) S_{\text{black}}$ balances confidence shift with psychological validity, see ablations in Sec. 5.5.

3.4 Backpropagation

After scoring a path, we backpropagate the score from the evaluated leaf to the root. For each visited node, we increment its $visit\_count$ and add the path's S_{final} to its $total\_score$. This update strengthens promising regions of the tree and steers subsequent MCTS selection toward optimizing high-quality Q&A sequences.

After the preset number of simulations, MCTS selects the explored path with the highest S_{final} as the optimal PsyPath for the input text.

4 Personality Inference and Detection

This section aims to train a downstream model for interpretable personality trait detection, using the self-explored high-quality Q&A paths generated by the PsyPath generation module.

4.1 Data Preparation and Training

For each text $\mathbf{x}$, we employ MCTS to explore the reasoning space and identify the optimal Q&A path. This yields a high-quality dataset of triplets $(\mathbf{x}, P_{\text{opt}}, y_{\text{true}})$, where P_{opt} is the path with the highest hybrid score.

We then fine-tune the target LLM using a two-stage process: Supervised Fine-Tuning (SFT) establishes the Q&A format, while Direct Preference Optimization (DPO) aligns the model's generation with the psychological evaluation criteria.

DPO minimizes the loss on preference pairs (P^+, P^-), where P^+ corresponds to the MCTS-selected optimal path, and P^- is a rejected candidate with a lower hybrid score:

$$\mathcal{L}_{\text{DPO}} = -\log \sigma\Big(\beta\big[\log \frac{\pi_\theta(P^+|x)}{\pi_{\text{ref}}(P^+|x)} - \log \frac{\pi_\theta(P^-|x)}{\pi_{\text{ref}}(P^-|x)}\big]\Big) \tag{3}$$

where π_θ is the optimized model, π_{ref} is the SFT reference, and β controls the preference strength. This approach effectively enables the model to internalize the diagnostic logic derived from the MCTS exploration.

4.2 Multi-step Inference

During inference, the fine-tuned model performs multi-step inference, which generates a structured sequence of diagnostic Q&A for the target trait and

follows this path to a final prediction (e.g., "high" or "low"). Trained with DPO on paired paths, it prioritizes psychologically coherent, diagnostically effective reasoning and produces transparent step-by-step outputs that clearly explain the judgment, and the final trait label is extracted from the structured output for evaluation. This procedure improves interpretability and mirrors psychological assessment.

5 Experiments

5.1 Data Preparation

We evaluated PsyPath on two standard personality detection benchmarks. The first is the Kaggle MBTI dataset[1], which has 8,674 entries from the personality forum PersonalityCafe[2], each labeled with one of the 16 Myers-Briggs Type Indicator (MBTI) types (Myers-Briggs, 1991). The second is the Essays dataset (Pennebaker and King, 1999), containing 2,468 texts annotated with the Big Five personality traits. For data preparation, we used 80% of each dataset to generate the training data and reserved a random 10% as the final test set.

5.2 Baselines

We compare PsyPath against a set of existing methods in personality detection.

These include neural network models like W2V+CNN (Rahman et al., 2019) combining word2vec embeddings, AttRCNN (Xue et al., 2018) integrating a hierarchical Inception variant; and pre-trained language models such as BERT (Devlin et al., 2019), DDGCN (Yang et al., 2022) employing domain-adapted BERT with a dynamic deep graph network, and RoBERTa (Liu et al., 2019) fine-tuned for the task.

For LLM baselines, we evaluate five widely used off-the-shelf general-purpose LLMs under a unified prompting/evaluation pipeline without task-specific fine-tuning: Kimi-K2 (Moonshot AI, 2025), a Mixture-of-Experts model from Moonshot AI aimed at agentic/computer-use scenarios; Gemini 2.5 Pro (Google DeepMind, 2025; Kavukcuoglu, 2025), a "thinking" model family from Google DeepMind designed for complex reasoning and coding; Qwen3-8B (Qwen Team, 2025b,a), Alibaba's next-generation 8B model emphasizing deeper reasoning and faster tool-use; LLaMA 3 8B Instruct (Meta AI, 2024), Meta's open-weight instruction-tuned 8B baseline; and Claude Sonnet 4 (Anthropic, 2025), a mid-size Claude 4 routing variant accessed through a provider-specific API. To ensure fair comparison, these baselines utilize the same prediction prompt template as the final stage of our method, which is provided in Appendix E. In addition, we adopt various strategies for performance improvement, including **TAE** (Hu et al., 2024) which improves performance through text augmentation and contrastive learning, and **PsyCoT** (Yang et al., 2023) which uses psychological questionnaires in a Chain-of-Thought (CoT) process for dialogue ratings.

To comprehensively evaluate the effectiveness of our method's Q&A path generation and utilization strategies, we introduce four specific Qwen2-based variants:

Direct represents the performance of the Qwen2 model directly fine-tuned on raw data, serving as a reference for conventional LLM fine-tuning.

FixedQ conducts MCTS by selecting questions from a fixed, predefined set instead of dynamic generation by the LLM, to evaluate the value of dynamic question generation.

RandPath uses randomly selected Q&A paths in the training phase that lead to correct predictions. By fine-tuning on these instead of relying on MCTS optimization, aiming to explicitly quantify the potential performance gain from the MCTS search optimization.

BoN employs a "Best-of-N" strategy ($N = 8$), generating multiple candidate reasoning paths and selecting the optimal one via our Hybrid Scoring. This baseline strictly isolates the specific contribution of the structured MCTS lookahead search relative to a simple "generate-and-rerank" strategy under an identical reward function.

5.3 Implementation Details

We utilize Meta-Llama-3-8B-Instruct for PsyPath generation and scoring, and Qwen2-7B for downstream detection. MCTS is configured with 50 iterations and a maximum search depth of 3. For scoring (Eq. 2), we set the balancing weight $\alpha = 0.5$, with white-box parameters $\lambda_1 = 10, \lambda_2 = 0.5$, sensitivity $k = 5$, and a bonus $\beta = 2$ for high confidence ($P_{\text{aug}} \geq 0.8$). Qwen2 is fine-tuned using LoRA (rank 16, alpha 32). SFT training employs a learning rate of 3.0×10^{-5} (10 epochs, batch 16),

[1] https://www.kaggle.com/datasnaek/mbti-type
[2] https://www.personalitycafe.com/forum

Methods	I/E		S/N		T/F		J/P		Average	
	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
AttRCNN	-	59.74	-	64.08	-	78.77	-	66.44	-	67.26
DDGCN	78.10	70.26	84.40	60.66	79.30	78.91	73.30	71.73	78.78	70.39
BERT	77.30	62.50	84.90	54.04	78.30	77.93	69.50	68.80	77.50	65.82
RoBERTa	77.10	61.89	86.50	57.59	79.60	78.69	70.60	70.07	78.45	67.06
Kimi-K2	84.00	71.43	86.17	89.09	89.50	88.27	84.25	76.96	85.98	81.44
Gemini-2.5-Pro	82.83	69.07	89.00	93.45	85.00	84.15	83.00	75.47	84.96	80.54
Qwen3-8B	73.50	50.47	74.50	84.31	66.50	68.25	63.00	70.40	69.38	68.36
LLaMA3-8B	65.00	44.44	74.00	83.95	63.00	61.86	63.00	68.38	66.25	64.66
Claude-Sonnet-4	80.50	70.68	91.00	94.77	89.30	88.77	73.00	75.62		
TAE	-	70.90	-	66.21	-	81.17	-	70.20	-	72.07
PsyCoT	79.00	66.56	85.00	61.70	75.00	74.80	57.00	57.83	74.00	65.22
Direct	73.43	72.00	81.57	81.83	73.29	72.23	70.71	72.50	74.75	74.64
FixedQ	78.00	**82.00**	73.00	72.16	74.00	75.00	73.67	73.93	74.67	75.77
RandPath	85.50	71.84	86.67	70.59	83.30	83.30	79.00	63.16	83.62	72.22
BoN	87.40	71.10	89.18	85.03	84.82	85.13	83.09	69.90	86.12	77.79
PsyPath (our)	**91.00**	81.63	**95.13**	**97.24**	**90.25**	**89.57**	**86.62**	**82.14**	**90.75**	**87.65**

Table 1: Overall results of PsyPath and baselines on the Kaggle MBTI dataset.

while DPO uses a rank of 8 and a rate of 8.0×10^{-6} (3 epochs, batch 8). Experiments were conducted on 3 NVIDIA A800 GPUs using vLLM with 50 parallel workers. Results are reported as averages over 5 random seeds. See Appendices A–E for prompts and Appendix H for cost analysis.

5.4 Main Experiments

To evaluate our method's effectiveness, we performed quantitative assessments for each personality trait dimension. We measured performance using Accuracy and F1-score, standard binary classification metrics. Our main experiment results, summarized in Table 1 (Kaggle MBTI dataset) and Table 2 (Essays dataset), consistently demonstrate the superior performance of PsyPath across all evaluated personality trait dimensions. As the results clearly show, PsyPath achieves the highest Accuracy and F1-score across both datasets, outperforming all baseline models.

Specifically, PsyPath substantially and consistently surpasses various baseline models, including traditional machine learning approaches and BERT, RoBERTa, etc., as well as strong LLM-based baselines such as TAE and PsyCoT. For instance, on the MBTI dataset, PsyPath's average score improved by approximately **15%-20%** compared to traditional methods and outperformed LLMs by at least **5%**; on the Essays dataset, the average score also significantly increased by about **7%-8%** points over traditional methods and at least **2%** points over LLMs. Furthermore, comparing Qwen2-based variants isolates the contribution of each component. PsyPath consistently outperforms **Direct** (Qwen2 standard fine-tuning) and **FixedQ**, confirming the necessity of our specialized framework and dynamic questioning capability. Crucially, regarding path optimization, we observe a clear performance hierarchy: **RandPath < BoN < PsyPath**. While **BoN** improves over random selection (validating the effectiveness of our hybrid scorer), PsyPath's superior performance demonstrates that *structured MCTS lookahead* captures deep diagnostic logic beyond what simple "generate-and-rerank" strategies can achieve.

5.5 Ablation Study

Impact of LLM Dynamic Question Generation The importance of LLM dynamic question generation is highlighted by comparing PsyPath with **Qwen2-FixedQ**. Qwen2-FixedQ conducts MCTS by using a fixed set of predefined questions instead of dynamically generated ones. As shown in Table 1 and Table 2, Qwen2-FixedQ's performance is far lower than PsyPath's by an average of 16.08 Acc./11.18 F1 on Kaggle and 10.05 Acc./17.01 F1 on Essays, which further affirms the significant value of LLM's capability to tailor effective diagnostic inquiries for personality detection.

Impact of MCTS Search Optimization We demonstrate the value of MCTS search optimization by comparing PsyPath with **Qwen2-RandPath**. Qwen2-RandPath utilizes randomly selected Q&A paths that lead to correct predic-

Methods	AGR		CON		EXT		NEU		OPN		Average	
	Acc.	**F1**	**Acc.**	**F1**	**Acc.**	**F1**	**Acc.**	**F1**	**Acc.**	**F1**	**Acc.**	**F1**
W2V+CNN	-	46.16	-	52.11	-	39.40	-	58.14	-	59.80	-	51.12
BERT	56.84	54.72	57.57	56.41	58.54	58.42	56.60	56.36	60.00	59.76	57.91	57.13
RoBERTa	59.03	57.62	57.81	56.72	57.98	57.20	56.93	56.80	60.16	59.88	58.38	57.64
Kimi-K2	52.00	52.48	45.50	15.50	54.00	61.02	47.50	45.60	59.00	61.86	51.60	47.29
Gemini-2.5-Pro	53.00	61.16	47.00	34.57	58.00	68.18	55.00	58.87	58.00	62.00	54.20	56.96
Qwen3-8B	53.00	51.86	45.50	15.50	54.00	51.02	52.00	52.48	47.50	45.60	50.4	43.29
LLaMA3-8B	54.00	58.93	51.50	50.26	43.50	48.40	48.00	46.39	48.50	44.32	49.1	49.66
Claude-Sonnet-4	53.00	51.05	38.00	13.89	58.00	65.00	56.00	58.71	42.00	36.9	61.35	45.52
PsyCoT	61.13	61.13	59.92	**57.41**	59.76	59.74	56.68	56.58	**60.73**	57.30	59.64	58.43
Direct	51.00	50.00	47.50	21.05	45.00	26.67	43.00	21.92	53.00	47.19	47.9	33.37
FixedQ	55.10	56.27	51.50	45.81	54.5	38.93	54.04	52.40	51.09	49.90	53.25	48.66
RandPath	58.84	58.09	51.50	55.30	53.5	55.07	53.80	53.06	58.74	54.40	55.28	55.18
BoN	59.09	60.18	57.73	55.50	57.27	55.24	55.45	56.64	57.27	56.07	57.36	56.73
PsyPath (our)	**65.00**	**75.86**	**65.00**	56.34	**65.50**	**69.33**	**60.50**	**59.07**	60.50	**67.76**	**63.30**	**65.67**

Table 2: Overall results of PsyPath and baselines on the Essays dataset. We use Accuracy(%) and Macro-F1(%) as metrics. Best results are listed in bold and the second best results are shown with underline.

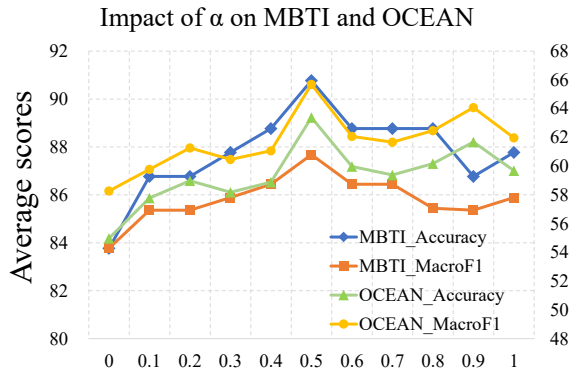

Figure 3: The Performance of Different α's methods.

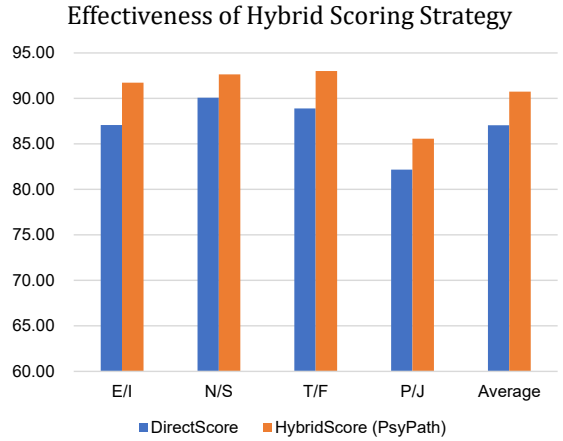

Figure 4: Comparative Performance of Direct Scoring and Hybrid Scoring Strategies.

tions for training, instead of those generated via MCTS optimization. As shown in Table 1 and Table 2, despite the correctness of its paths, Qwen2-RandPath's performance still falls short of PsyPath's by an average of 7.13 Acc./15.43 F1 on Kaggle and 8.02 Acc./10.49 F1 on Essays, underscoring the vital role of MCTS in selecting optimal reasoning sequences.

Impact of α Parameter We examine the impact of the weighting parameter α. As shown in Figure 3, the model performs best when $\alpha = 0.5$. Specifically, the hybrid strategy (90.77 Acc./87.68 F1) outperforms both pure black-box ($\alpha = 0$, 83.77 Acc./F1) and white-box guidance ($\alpha = 1$, 87.77 Acc./85.89 F1). This confirms that balancing both information sources provides optimal guidance for MCTS search.

Effectiveness of Hybrid Scoring Strategy We compared the full PsyPath (using hybrid scoring) with an MCTS guidance method based on model direct scoring (PsyPath DirectScore version). In DirectScore, we prompted the LLM directly to score "whether the question could aid in personality judgment for that dimension" as the MCTS reward signal. As Figure 4 shows, while the DirectScore version achieved average Accuracy 87.1%, it remained significantly lower than PsyPath's average accuracy of 90.75%. This result strongly proves the superiority of our proposed hybrid scoring mechanism by demonstrating how the integration of external psychological alignment (black-box) and internal model confidence (white-box) provides a more effective reward signal for MCTS search, enhancing overall model performance.

5.6 Human Expert Evaluation

To rigorously validate our Black-box Scoring mechanism, we conducted a blind evaluation with five psychology experts. We sampled 50 instances and presented experts with two reasoning paths generated by PsyPath: one with a **High Black-box Score** (top 25%) and one with a **Low Black-box Score** (bottom 25%). Experts performed two tasks: 1-5 Likert ratings on diagnostic dimensions and a pairwise comparison of overall value. See Appendix G for detailed evaluation protocol.

Inter-annotator agreement was assessed using binarized ratings (*Acceptable* ≥ 3; see Appendix G). As shown in Table 3, experts achieved near-perfect consensus (**>99%**) on **Cognitive Style** and strong agreement on other dimensions. In pairwise comparisons, experts preferred High-Score paths (57.5%) significantly more than Low-Score ones (24.0%; $p < 0.001$), validating the scorer's effectiveness. Qualitative analysis attributes lower ratings primarily to redundancy (32%) rather than hallucinations. To further scrutinize the model's limitations, we provide a comprehensive error analysis of representative bad cases in Appendix I.

Dimension	SD	Agree.%
Behavioral Consistency	1.04	79.2%
Cognitive Style	**0.53**	**>99%**
Motivation & Self-Concept	0.82	91.7%

Table 3: Expert agreement statistics. SD denotes the Standard Deviation of raw scores.

6 Related work

The field of personality detection from text has traditionally framed the task as a direct classification problem, mapping text to trait labels. Early approaches used psycholinguistic features (Cui and Qi, 2017), evolving into complex neural networks like W2V+CNN (Rahman et al., 2019), and sophisticated pre-trained language models such as BERT (Devlin et al., 2019), DDGCN (Yang et al., 2022), and RoBERTa (Liu et al., 2019). While these methods advanced mapping capabilities, they often lacked transparency and failed to capture the intricate, iterative psychological reasoning processes that human experts employ.

More recent explorations with Large Language Models (LLMs) have introduced some forms of reasoning to improve personality detection performance (Yang et al., 2023; Hu et al., 2024; Li et al., 2024). However, these still struggle to fully emulate the nuanced, diagnostic inquiry characteristic of human psychological assessment.

To further enhance reasoning capabilities, a growing body of literature explores augmenting LLMs with search-based planning strategies. Concurrently, advanced planning algorithms like Tree of Thoughts (ToT) (Yao et al., 2023) and MCTS-based RAP (Hao et al., 2023; Feng et al., 2023) have revolutionized reasoning in deterministic domains (e.g., math/coding). However, these methods rely on objective verifiers. PsyPath adapts MCTS for subjective psychological reasoning via a novel hybrid scoring mechanism.

Beyond detecting personality in human-authored texts, significant research now focuses on understanding and manipulating LLM personality. This includes evaluating their psychometric properties (Maharjan et al., 2025; Heston and Gillette, 2025; Serapio-García et al., 2025), assessing consistent persona maintenance (Bhandari et al., 2025), and inducing specific traits via latent feature steering or targeted fine-tuning (Jiang et al., 2023; Yang et al., 2025; Chen et al., 2024). Other studies investigate LLM-generated text's personality detection and explanation (Ji et al., 2025), and parameter-efficient fine-tuning for personality detection tasks (Shen et al., 2025). The broader landscape of personality computing also considers critical threats, challenges, and future directions (Celli et al., 2025), alongside deriving new insights from natural language processing (Saeteros et al., 2025).

7 Conclusion

Our paper introduces PsyPath, a new framework that improves model performance in personality trait detection by using a dynamic, psychologically-guided reasoning process. Instead of relying on a static and shallow intermediate reasoning process, PsyPath enables LLMs to iteratively generate and answer diagnostic questions, forming psychologically aligned and confidence-sensitive reasoning paths that reflect how human psychologists structure personality assessment. This dynamic inference process is guided by a hybrid scoring mechanism that combines psychological alignment captured via black-box evaluation, and model confidence shifts captured via white-box signals within an MCTS framework. Experiments show that PsyPath significantly outperforms existing methods on two benchmark datasets.

Limitations

This study primarily focuses on improving the LLM's performance by leveraging psychological knowledge. How to exploit more trainable and tunable models to further optimize PsyPath is left for future investigation.

Additionally, we acknowledge the inherent noise and psychometric debates surrounding the Kaggle MBTI dataset. While we utilize it as a standard benchmark for comparison with prior work, we mitigate these validity concerns by also evaluating our framework on the scientifically more rigorous Big Five Essays dataset.

This method carries certain potential risks. Even with well-intentioned use, personality detection may lead to misjudgments, negatively impacting individuals' careers or social relationships.

Reproducibility, Artifacts, and Ethics

Datasets and Artifacts

Public datasets used: Kaggle MBTI, Essays (OCEAN). LLM baselines via APIs (Gemini 2.5 Pro, Claude Sonnet 4 , Kimi-K2, Qwen3-8B, LLaMA3-8B-Instruct). We release PsyPath code and used datasets (details in Licensing).

Dataset Statistics and Splits

Report counts, train/dev/test splits, and preprocessing steps.

Intended Use and Compliance

All uses restricted to research/education; no high-stakes deployment. Derivatives from research-only sources remain research-only; users must comply with original terms.

Data Privacy and Safety

No intentional collection of PII; we do not redistribute raw third-party text. Offensive content may exist in-source; we follow dataset terms and apply basic filtering where applicable.

Compute, Model Sizes, and Budget

List parameter counts (e.g., 8B), hardware (GPU type/num), API token usage/cost ranges.

Experimental Setup and Hyperparameters

Training/inference settings, search spaces, best hyperparameters, prompts/configs.

Reporting and Variance

State whether results are single-run vs. mean$\pm$std over k seeds; note any deterministic settings.

Software and Package Versions

Experiments were conducted on Linux (Ubuntu 20.04) using Python 3.10. We implemented the models using `Torch` 2.6.0 and `Transformers` 4.51.3. Evaluation metrics (Accuracy, F1-score) were calculated using `scikit-learn` 1.6.1. Data processing was performed using `Pandas` and `NumPy`. Full dependency versions are provided in the `environment.yml` included in the supplementary materials.

Use of AI Assistants and Synthetic Data Generation

We distinguish between the use of AI tools for manuscript preparation and their role in our experimental methodology:

- **Manuscript Preparation:** AI assistants (e.g., ChatGPT) were used limitedly for grammatical polishing, LaTeX formatting, and code refactoring. All scientific claims and analyses remain the authors' original work.
- **Methodological Generation:** As explicitly described in our proposed framework (PsyPath), we utilized Large Language Models (specifically Llama-3-8B-Instruct) to generate synthetic reasoning paths and pseudo-labels. This AI-generated data is central to our distillation process and is documented as a core component of the research contribution rather than an undisclosed artifact.

References

Anthropic. 2025. Introducing claude 4. https://www.anthropic.com/news/claude-4. Accessed 2025-10-05.

Pranav Bhandari, Nicolas Fay, Michael Wise, Amitava Datta, Stephanie Meek, Usman Naseem, and Mehwish Nasim. 2025. Can llm agents maintain a persona in discourse? *arXiv preprint arXiv:2502.11843*.

Fabio Celli, Aleksandar Kartelj, Miljan Đorđević, Derwin Suhartono, Vladimir Filipović, Veljko Milutinović, Georgios Spathoulas, Alessandro Vinciarelli, Michal Kosinski, and Bruno Lepri. 2025. Twenty years of personality computing: Threats, challenges and future directions. *Preprint*, arXiv:2503.02082.

Yanquan Chen, Zhen Wu, Junjie Guo, Shujian Huang, and Xinyu Dai. 2024. Extroversion or introversion? controlling the personality of your large language models. *Preprint*, arXiv:2406.04583.

Brandon Cui and Calvin Qi. 2017. Survey analysis of machine learning methods for natural language processing for mbti personality type prediction. *Final Report Stanford University*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Seymour Epstein. 1994. Integration of the cognitive and the psychodynamic unconscious. *American psychologist*, 49(8):709.

Xidong Feng, Ziyu Wan, Muning Wen, Ying Wen, Weinan Zhang, and Jun Wang. 2023. Alphazero-like tree-search can guide large language model decoding and training. *arXiv preprint arXiv:2309.17179*.

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

ME Francis and Roger J Booth. 1993. Linguistic inquiry and word count. *Southern Methodist University: Dallas, TX, USA*.

Google DeepMind. 2025. Gemini 2.5 pro. https://deepmind.google/models/gemini/pro/. Accessed 2025-10-05.

Shibo Hao, Yi Gu, Haodi Ma, Joshua Jiahua Hong, Zhen Wang, Daisy Zhe Wang, and Zhiting Hu. 2023. Reasoning with language model is planning with world model. In *EMNLP*.

Thomas F Heston and Justin Gillette. 2025. Do large language models have a personality? a psychometric evaluation with implications for clinical medicine and mental health ai. *medRxiv*, pages 2025–03.

Linmei Hu, Hongyu He, Duokang Wang, Ziwang Zhao, Yingxia Shao, and Liqiang Nie. 2024. Llm vs small model? large language model based text augmentation enhanced personality detection model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18234–18242.

Jiazhou Ji, Jie Guo, Weidong Qiu, Zheng Huang, Yang Xu, Xinru Lu, Xiaoyu Jiang, Ruizhe Li, and Shujun Li. 2025. " i know myself better, but not really greatly": How well can llms detect and explain llm-generated texts? *arXiv preprint arXiv:2502.12743*.

Guangyuan Jiang, Manjie Xu, Song-Chun Zhu, Wenjuan Han, Chi Zhang, and Yixin Zhu. 2023. Evaluating and inducing personality in pre-trained language models. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Oliver P. John, Richard W. Robins, and Lawrence A. Pervin, editors. 2008. *Handbook of Personality: Theory and Research*, 3rd edition. Guilford Press, New York.

Koray Kavukcuoglu. 2025. Gemini 2.5: Our most intelligent ai model. https://blog.google/technology/google-deepmind/gemini-model-thinking-updates-march-2025/. Accessed 2025-10-05.

Zheng Li, Dawei Zhu, Qilong Ma, Weimin Xiong, and Sujian Li. 2024. Eerpd: Leveraging emotion and emotion regulation for improving personality detection. *arXiv preprint arXiv:2406.16079*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Xinyuan Lu and Min-Yen Kan. 2023. Improving recommendation systems with user personality inferred from product reviews. *Information Processing & Management*, 60(2):103183.

Julina Maharjan, Ruoming Jin, Jianfeng Zhu, and Deric Kenne. 2025. Psychometric evaluation of large language model embeddings for personality trait prediction. *Journal of Medical Internet Research*, 27:e75347.

Paul E. Meehl. 1954. *Clinical versus statistical prediction: A theoretical analysis and a review of the evidence*. University of Minnesota Press.

S. Meskelis and J. L. Whittington. 2020. Driving employee engagement: how personality trait and leadership style impact the process. *Journal of Business & Industrial Marketing*, 35(10):1457–1473.

Meta AI. 2024. Introducing meta llama 3. https://ai.meta.com/blog/meta-llama-3/. Accessed 2025-10-05.

Moonshot AI. 2025. Kimi k2: Open agentic intelligence. https://moonshotai.github.io/Kimi-K2/. Accessed 2025-10-05.

Isabel Myers-Briggs. 1991. Introduction to type: A description of the theory and applications of the myers-briggs indicator. *Consulting Psychologists: Palo Alto*.

James W Pennebaker and Laura A King. 1999. Linguistic styles: language use as an individual difference. *Journal of personality and social psychology*, 77(6):1296.

Qwen Team. 2025a. Qwen3 models. https://github.com/QwenLM/Qwen3. Accessed 2025-10-05.

Qwen Team. 2025b. Qwen3: Think deeper, act faster. https://qwenlm.github.io/blog/qwen3/. Accessed 2025-10-05.

Md Abdur Rahman, Asif Al Faisal, Tayeba Khanam, Mahfida Amjad, and Md Saeed Siddik. 2019. Personality detection from text using convolutional neural network. In *2019 1st international conference on advances in science, engineering and robotics technology (ICASERT)*, pages 1–6. IEEE.

V. Ramachandran, A. Loya, K. P. Shah, S. Goyal, E. A. Hansoti, and A. C. Caruso. 2020. Myers-briggs type indicator in medical education: a narrative review and analysis. *Health Professions Education*, 6(1):31–46.

Carl Ransom Rogers and 1 others. 1959. *A theory of therapy, personality, and interpersonal relationships: As developed in the client-centered framework*, volume 3. McGraw-Hill New York.

David Saeteros, David Gallardo-Pujol, and Daniel Ortiz-Martínez. 2025. Text speaks louder: Insights into personality from natural language processing. *PLoS One*, 20(6):e0323096.

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

Greg Serapio-García, Mustafa Safdari, Clément Crepy, Luning Sun, Stephen Fitz, Peter Romero, Marwa Abdulhai, Aleksandra Faust, and Maja Matarić. 2025. Personality traits in large language models. *Preprint*, arXiv:2307.00184.

N. N. Sharan and D. M. Romano. 2020. The effects of personality and locus of control on trust in humans versus artificial intelligence. *Heliyon*, 6(8):e04572.

Lingzhi Shen, Yunfei Long, Xiaohao Cai, Guanming Chen, Imran Razzak, and Shoaib Jameel. 2025. Less but better: Parameter-efficient fine-tuning of large language models for personality detection. *Preprint*, arXiv:2504.05411.

F. M. Vertue and B. D. Haig. 2008. An abductive perspective on clinical reasoning and case formulation. *Journal of Clinical Psychology*, 64(9):1046–1060.

Di Xue, Lifa Wu, Zheng Hong, Shize Guo, Liang Gao, Zhiyong Wu, Xiaofeng Zhong, and Jianshan Sun. 2018. Deep learning-based personality recognition from text posts of online social networks. *Applied Intelligence*, 48.

Shu Yang, Shenzhe Zhu, Liang Liu, Lijie Hu, Mengdi Li, and Di Wang. 2025. Exploring the personality traits of llms through latent features steering. *Preprint*, arXiv:2410.10863.

Tao Yang, Jinghao Deng, Xiaojun Quan, and Qifan Wang. 2022. Orders are unwanted: Dynamic deep graph convolutional network for personality detection. In *Proceedings of AAAI 2023*.

Tao Yang, Tianyuan Shi, Fanqi Wan, Xiaojun Quan, Qifan Wang, Bingzhe Wu, and Jiaxiang Wu. 2023. Psycot: Psychological questionnaire as powerful chain-of-thought for personality detection. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. In *NeurIPS*.

B. Zarouali, T. Dobber, and J. Schreuder. 2024. Personality and susceptibility to political microtargeting: a comparison between a machine-learning and self-report approach. *Computers in Human Behavior*, 151:108024.

A Question Generation Prompt

To elicit psychologically meaningful diagnostic questions during MCTS search, we use the prompt shown in Figure 5.

```
You are an expert in personality psychology and
    psychometrics.
Your task is to generate 3 new yes/no questions that
    can help identify whether the author exhibits the
    personality trait: **{TraitName}**.

Please use the author's posts and avoid repeating
    previous questions.

Author's Posts:
{UserText}

Previously asked questions:
{UsedQuestions}

What makes a good question:
- It explores consistent behavioral patterns across
    time or situations
- It reflects the author's cognitive style
- It reveals internal motivations, values, or self-
    concept
- It must be answerable based on the text

Rules:
- Generate exactly 3 questions.
- Format:
  Q1: ...
  Q2: ...
  Q3: ...
- Use 'the author' instead of 'you'
- Avoid cliches and vague phrasing

New diagnostic questions:
Q1: ...
Q2: ...
Q3: ...
```

Figure 5: Prompt used for dynamic question generation.

B Answering Prompt

To simulate step-by-step reasoning over personality traits, the model answers each generated question using the prompt shown in Figure 6.

```
You are an AI assistant who specializes in MBTI
    personality trait detection.

Given a post by an author, I will ask you a question
    about one MBTI dimension. You need to choose the
    most likely answer.

Author's Posts:
{UserText}

Question:
{Question}

Please answer in the following format:
Answer: (a single-word response: 'yes' or 'no')
Explanation: (1 sentence with evidence from the text.)
```

Figure 6: Prompt used for answering diagnostic questions.

C Black-box Scoring Prompt

The black-box scorer evaluates the Q&A path using psychological constructs via the prompt shown in Figure 7.

```
You are an expert in personality psychology. Given the
    author's posts and a Q&A sequence, your task is
    to evaluate how helpful these Q&As are for
    inferring the trait: '{TraitName}'.

Rate the Q&A path on three dimensions (1-10):
1. Behavioral Consistency
2. Cognitive Style
3. Motivation and Self-Concept

Author's Posts:
{UserText}

Q&A Path:
Q1: ...
A1: ...
Q2: ...
A2: ...

Your output format:
Behavioral Consistency: <score>
Cognitive Style: <score>
Motivation and Self-Concept: <score>
```

Figure 7: Prompt used for black-box psychological scoring.

D White-box Scoring Prompts

White-box scoring measures belief shift using pre- and post-Q&A prompts, as detailed in Figure 8.

```
# Baseline Confidence (p_before)
You are an AI assistant trained to determine if the
    author is {TraitName}.
Read the text and decide whether their {TraitName} is
    high or low.

Author's Posts:
{UserText}

Question: Is the author's {TraitName} high or low?
Answer: (only 'high' or 'low')


# Updated Confidence (p_after)
You are also given diagnostic Q&A pairs to refine your
    prediction.

Author's Posts:
{UserText}

Q&A Pairs:
Q1: ...
A1: ...
Q2: ...
A2: ...

Question: Is the author's {TraitName} high or low?
Answer: (only 'high' or 'low')
```

Figure 8: Prompts used for white-box scoring before and after reasoning.

E Universal Prompt Template for Personality Prediction

To ensure a fair and consistent evaluation, we employ a **unified prompt template** for both the final inference stage of PsyPath and the zero-shot evaluation of general-purpose LLM baselines (e.g., Kimi-K2, Gemini 2.5 Pro, Qwen3). This template was applied to all LLM baselines to ensure fair comparison.

As shown in Figure 9, this template is designed to elicit a binary personality trait prediction.

```
You are an expert in personality detection.
Given the following posts from one author, generate 3
    new and different yes/no questions and answer
    them to help determine whether the author is {
    type[dimension]}.
Use these questions to explore the author's behavior
    or preferences.
After answering, make a final prediction about
    whether the author is {type[dimension]} or not.

What makes a good question:
- It explores consistent behavioral patterns across
    time or situations
- It reflects the author's cognitive style (how they
    perceive or process the world)
- It reveals internal motivations, value systems, or
    self-concept
- It must be answerable based on the text (avoid
    generic or abstract items)

Note: Your reasoning should be guided by the author's
    original posts, not only by the questions
    themselves.

Author's Post:
{UserText}
Carefully examine the posts and generate 3 insightful
    yes/no questions.
Use not only the answers but also the author's
    original posts to help form your final judgment
    about the author's personality.

Based on the above, is the author {TraitName} high or
    low?
Answer: ('high' or 'low')
```

Figure 9: Prompt used for final personality trait prediction.

F Sample Reasoning Paths

Figure 10 shows two full reasoning paths generated by PsyPath, one leading to high and the other to low extraversion predictions.

G Human Evaluation Details

G.1 Annotation Guidelines

To ensure consistency, experts were provided with detailed criteria for each score level. Table 4 summarizes the rubric used for the "Cognitive Style" dimension as an example.

G.2 Expert Demographics

We recruited five domain experts to ensure high-quality evaluation. To strictly control for domain knowledge, all annotators met the following criteria:

```
# Sample 1: Predicting High Extraversion

Author's Posts:
"I'm pretty happy with my first week and a half of
    classes ..."

Q1: Does the author enjoy participating in group
    activities ?
A1: Yes

Q2: Does the author adapt well to social challenges?
A2: Yes

Q3: Does the author proactively seek out social
    engagement?
A3: Yes

Final Prediction: High Extraversion
White-Box Score: 9.5
Black-Box Score: 8.25

---

# Sample 2: Predicting Low Extraversion

Author's Posts:
"I'm pretty happy with my first week and a half of
    classes ..."

Q1: Does the author prefer solitary relaxation after
    class?
A1: Yes

Q2: Does the author avoid initiating conversations
    with peers?
A2: Yes

Q3: Does the author prioritize introspection over
    socializing ?
A3: Yes

Final Prediction: Low Extraversion
White-Box Score: -4.96
Black-Box Score: 2.35
```

Figure 10: Reasoning paths predicting high and low extraversion.

- **Educational Background:** Current graduate students (Master's or PhD candidates) in Psychology at a major research university.
- **Longitudinal Training:** All experts possess an undergraduate degree in Psychology, ensuring a solid foundation in psychometric theories and cognitive science.
- **Task Familiarity:** Prior to the formal evaluation, experts underwent a tutorial phase where they annotated 5 trial instances (excluded from the final results) to calibrate their understanding of the rubric.

The experts were compensated at a rate exceeding the local minimum hourly wage. The evaluation was conducted in a double-blind manner.

Score	Criteria Description
5 (Excellent)	The generated reasoning perfectly captures the subject's unique cognitive patterns, vocabulary, and sentence structures. No generic phrasing.
4 (Good)	High consistency with the subject's style. Minor generic phrases present but do not disrupt the persona.
3 (Fair)	**Valid Baseline.** The reasoning is logical and does not contradict the subject's profile, but the tone may be somewhat generic or flat.
2 (Poor)	Noticeable inconsistencies in tone or logic. The persona feels "off" or robotic.
1 (Very Poor)	Severe hallucinations or complete mismatch with the subject's known cognitive style.

Table 4: Annotation Rubric for Cognitive Style.

G.3 Questionnaire Design and Sample Instance

The evaluation was conducted using a standard web-based survey platform. Each task presented the experts with a subject profile, two candidate reasoning paths (A and B), and a set of evaluation questions. To facilitate reproducibility, we provide a translated sample of one evaluation instance in Figure 11.

Sample Instance (Target Dimension: Intuition/ Abstract Thinking). In this example, Path A (High-Score) correctly addresses the target dimension, while Path B (Low-Score) focuses on irrelevant traits, illustrating the "Weak Relevance" issue.

G.4 Recruitment and Compensation

We recruited five domain experts to ensure high-quality evaluation. To strictly maintain anonymity while ensuring fair labor practices, we report compensation in USD equivalents. Annotators were compensated at an hourly rate of approximately **$14 USD/hr** (based on an average session duration of 1.5 hours). We confirm that this rate is highly competitive, exceeding the statutory minimum hourly wage in the annotators' region of residence by a factor of approximately **4**×.

[Part 1: Survey Instructions & Guidelines]

1. Overview: This questionnaire evaluates the diagnostic value of two "Reasoning Paths" (Path A & Path B). Please judge which path better helps understand the author's personality traits based on the text.

2. Key Concepts:

- **Behavioral Consistency:** Stability of behavioral tendencies across situations.
- **Cognitive Style:** Preferred thinking mode (abstract vs. concrete, rational vs. emotional).
- **Motivation:** Intrinsic motivations and self-view.

3. Scoring Standards (1–5 Scale):

- **5 (Excellent):** Perfectly captures cognitive patterns/style. No generic phrasing.
- **4 (Good):** High consistency. Minor generic phrases.
- **3 (Fair): Valid Baseline.** Logical, no contradiction, but tone is generic.
- **2 (Poor):** Inconsistent tone or logic. Robot-like.
- **1 (Very Poor):** Severe hallucinations or mismatch.

[Part 2: Task Instance Example]

Context Snippet: *"The author's posts contain significant personal feelings... sharing of interpersonal relationships... enjoying solitude... reflecting on life changes..."*

Path A (High-Score Candidate):

Q1: Does the author tend to base evaluations on logic and objective standards rather than emotional reactions or personal preferences?
Answer: No.
Explanation: The author's posts emphasize emotional experiences, personal dilemmas, and interpersonal interactions, indicating a subjective and emotion-oriented evaluation style.

Q2: Is the writing and expression style focused on clarity, rigor, and logical structure rather than creative expression, emotional resonance, or narrative?
Answer: No.
Explanation: The writing is casual, colloquial, and emotionally charged, focusing on personal sharing rather than structured or logical argumentation.

Q3: When forming opinions or making decisions, does the author tend to analyze information objectively rather than relying on intuition or emotional connection?
Answer: No.
Explanation: The author frequently mentions their "inner world" and "solitude," suggesting a reliance on personal intuition and internal feeling states to understand and judge things.

Path B (Low-Score Candidate):

Q1: Does the author tend to plan and organize time in advance, with clear tasks and activity schedules, rather than acting spontaneously?
Answer: No.
Explanation: The author's posts show a preference for "going with the flow," enjoying leisure and spontaneous activities rather than emphasizing planning and time management.

Q2: Even without explicit planning, does the author maintain a sense of structure and regularity in daily life?
Answer: No.
Explanation: The author repeatedly mentions enjoying "solitude" and "me-time" in a way that suggests a preference for flexibility and mood-based actions over a stable daily structure.

Q3: When engaging in creative activities (like writing or gaming), does the author usually set a clear plan beforehand and execute it strictly?
Answer: No.
Explanation: The author tends to let the creative process unfold naturally based on their current state, rather than pushing forward according to a pre-set plan.

Evaluation Questions:

Q1. Diagnosis Value (Pairwise Comparison)
Which path has greater diagnostic value for the target dimension (Intuition)?
[ ] Path A [ ] Path B [ ] Tie

Q2. Quality Rating (Matrix Scale)
Rate the selected path on (1–5):
• Behavioral Consistency (1–5) • Cognitive Style (1–5) • Motivation (1–5)

Q3. Error Analysis (Multi-select)
What issues exist in the reasoning paths?
[ ] Generic phrasing [ ] Weak relevance [ ] Leading questions [ ] Logic issues [ ] Redundancy

Figure 11: Full layout of the human evaluation interface, including the instruction set, definition of dimensions, and detailed scoring rubric provided to experts.

H Computational Cost Analysis

In this section, we provide a formal analysis of the time complexity and computational cost of our proposed MCTS-based data generation method compared to the Best-of-N (BoN) baseline. We decompose the pipeline into three stages: Data Generation, Model Training, and Inference.

H.1 Variable Definitions

We define the computational cost based on atomic Large Language Model (LLM) API calls or inference steps. Let:

- T_{gen} **(Generation Latency):** The time required to generate candidate questions for a single node.
- T_{ans} **(Answering Latency):** The time required to generate a "Yes/No" response for a specific question given the context.
- T_{eval} **(Evaluation Latency):** The time required to evaluate a complete path (depth $D = 3$), including the final prediction, white-box scoring, and black-box scoring.

H.2 Data Generation (Offline)

This is the stage where the computational difference lies. We compare our MCTS method (with score caching) against the BoN baseline ($N = 8$).

1. Proposed Method: MCTS Our MCTS implementation uses a ternary tree structure (*Branching Factor* $K = 3$) with a maximum depth of $D = 3$. The maximum iteration is set to $M = 50$, and a caching mechanism is implemented to prevent redundant evaluations of identical paths.

- **Tree Expansion Cost:** The tree has internal nodes at Depth 0, 1, and 2. The total number of internal nodes to be expanded is $\sum_{i=0}^{D-1} K^i = 3^0 + 3^1 + 3^2 = 1 + 3 + 9 = 13$ unique expansion points. Each expansion involves generating questions once and obtaining $K = 3$ answers.
- **Evaluation Cost:** With score caching enabled, the evaluation is performed at most once for each unique leaf node. The maximum number of unique leaves (i.e., complete paths) is $K^D = 3^3 = 27$.

The total time cost for MCTS is:

$$\textbf{Cost}_{MCTS} = 13 \cdot T_{gen} + (13 \times 3) \cdot T_{ans} + 27 \cdot T_{eval}$$

$$\textbf{Cost}_{MCTS} = 13 \cdot T_{gen} + 39 \cdot T_{ans} + 27 \cdot T_{eval}$$

2. Baseline: Best-of-N (BoN) The baseline generates $N = 8$ independent paths of depth $D = 3$.

- **Path Generation Cost:** Generating 8 independent chains requires $N \times D = 8 \times 3 = 24$ sequential steps. Each step involves generating a question and an answer.
- **Evaluation Cost:** All 8 completed paths are evaluated.

The total time cost for BoN is:

$$\textbf{Cost}_{BoN} = (8 \times 3) \cdot T_{gen} + (8 \times 3) \cdot T_{ans} + 8 \cdot T_{eval}$$

$$\textbf{Cost}_{BoN} = 24 \cdot T_{gen} + 24 \cdot T_{ans} + 8 \cdot T_{eval}$$

3. Comparative Analysis Comparing the two methods:

- **Question Generation Cost (T_{gen}):** MCTS requires fewer generation calls (13 calls) compared to BoN (24 calls). This represents a reduction of approximately 45.8% (calculated as $(24-13)/24 \times 100\%$) in the number of generation API calls. This efficiency stems from the hierarchical reuse of parent nodes in the tree structure, avoiding redundant generation of initial questions.
- **Answer Generation Cost (T_{ans}):** MCTS requires more answering calls (39 calls) compared to BoN (24 calls), as it must obtain answers for all $K = 3$ branches at each expansion point to explore the full tree.
- **Path Evaluation Cost (T_{eval}):** MCTS invests more in evaluation (27 calls) compared to BoN (8 calls). This is a deliberate design choice: we expend roughly $\approx 3.3\times$ (calculated as $27/8$) the evaluation compute to explore a search space that is $3.3\times$ larger than the baseline, thereby ensuring higher potential data quality.

H.3 Training and Inference (Online)

A crucial advantage of our framework is that the additional computational cost is confined strictly to the **offline data generation phase**.

- **Training (SFT + DPO):** Both methods select the single best path ("Chosen") per sample to construct the training dataset. The model architecture, dataset size, and average

token length remain identical across both approaches. Thus, the training time is invariant:

$$\mathbf{Cost}_{Train}^{MCTS} \approx \mathbf{Cost}_{Train}^{BoN}$$

- **Inference (Test):** During testing, the trained model generates the chain-of-thought and prediction in a single pass (Direct Generation). The inference latency depends solely on the model architecture (e.g., Llama-3-8B) and output length, which are consistent across both methods.

$$\mathbf{Cost}_{Inference}^{MCTS} \approx \mathbf{Cost}_{Inference}^{BoN}$$

H.4 Conclusion

While the proposed MCTS method incurs a higher evaluation cost during the data construction phase ($\approx 3.3\times$ increase in evaluation calls for 3.3x search coverage), it significantly reduces the number of question generation calls by $\approx 45.8\%$. More importantly, this represents an efficient **"Offline Compute for Online Quality"** trade-off. We effectively distill the "System 2" search capabilities of MCTS into a "System 1" model, achieving superior performance in the ultimate trained model without increasing online inference latency.

I Qualitative Error Analysis and Human Evaluation

To strictly address the request for a comprehensive error analysis, we present a qualitative deep-dive into representative failure cases from our human evaluation campaign. We selected four distinct examples that highlight specific limitations in the current PsyPath framework.

For each case, we provide the English translation of the **Input Text**, the generated **Reasoning Paths** (PsyPath vs. Baseline), and the **Human Annotation Details** from 5 independent annotators. The annotation metrics include Preference (Q_1), Dimension Ratings (Q_2: Behavioral Consistency, Q_3: Cognitive Style, Q_4: Motivation), and Error Tags (Q_5).

I.1 Case 1: The "Safe but Redundant" Trap

Observation: PsyPath (Path A) achieves unanimous preference ($5/5$), yet it is heavily penalized for **Redundancy (E)** and **Potential Bias (C)**. This indicates that while the model captures the correct trait direction, it struggles to generate diverse diagnostic questions.

- **Input Text:** "I just cherish my alone time because I am more immersed in my inner world, whereas most of the time I am working... Enjoy the solitude while you can. || Moving to the Denver area to start a new life for myself."

- **Path A (PsyPath):**
  Q1: Does the author tend to evaluate matters based on logical reasoning and objective standards rather than emotional reactions?
  A: No. *Explanation: The author's statements show...*
  Q2: Does the author prioritize analyzing information over relying on intuition or emotional connection?
  A: No. *Explanation: The author repeatedly mentions "inner world" and "alone time," indicating a reliance on internal feelings...*

- **Path B (Baseline):**
  Q1: Does the author tend to plan and organize time in advance rather than being spontaneous?
  A: No. *Explanation: The author emphasizes enjoying the moment...*

Annotator	Pref (Q_1)	Q_2	Q_3	Q_4	Error Tags (Q_5)
#1	**Path A**	3	4	4	Weak Correlation (B), Redundancy (E)
#2	**Path A**	2	4	2	Leading/Bias (C), Redundancy (E)
#3	**Path A**	2	5	4	Redundancy (E)
#4	**Path A**	5	5	5	No Issues (F)
#5	**Path A**	3	4	3	Redundancy (E)

Table 5: Human evaluation details for Case 1.

Root Cause Diagnosis: Semantic Saturation. The model correctly identified the "Introverted" trait but failed to expand the reasoning graph. Instead of exploring *why* the user likes solitude (e.g., for creativity vs. recovery), the MCTS process got stuck in a local optimum, repeatedly generating variations of the same question.

I.2 Case 2: Interpretation Divergence on Ambiguous Text

Observation: While preference for PsyPath is high ($5/5$), the numerical ratings for specific dimensions diverge significantly (e.g., Motivation scores range from 1 to 5). This suggests the model's reasoning logic is persuasive but the evidence cited is ambiguous.

- **Input Text:** "This is another stupid misunderstanding. || Approaching problems logically is the key to unlocking the results you truly

want. || ...If you aren't using critical thinking, you aren't using your brain at all. || I am loyal to the vision itself."

- **Path A (PsyPath):**
  Q1: Does the author tend to evaluate specific events using abstract principles rather than specific contexts?
  A: Yes.
  Q2: When making decisions, do they rely on rational analysis over emotional intuition?
  A: Yes. *Explanation: The author explicitly distinguishes value judgments from emotional reactions...*

- **Path B (Baseline):**
  Q1: Does the author tend to rely on established norms or social consensus to judge right from wrong?
  A: No.

Annotator	Pref (Q_1)	Q_2	Q_3	Q_4	Error Tags (Q_5)
#1	**Path A**	3	5	3	Weak Correlation (B)
#2	**Path A**	2	4	**1**	Redundancy (E)
#3	**Path A**	5	5	**5**	No Issues (F)
#4	**Path A**	4	5	**5**	No Issues (F)
#5	**Path A**	3	3	**2**	Weak Correlation (B)

Table 6: Human evaluation details for Case 2.

Root Cause Diagnosis: Subjectivity of Aggressive Tone. The input text is aggressive. Annotators split on interpretation: some viewed this as strong "Cognitive Style" (Score 5), while others interpreted the aggressive tone as *emotional volatility* disguised as logic, rating the model's "Rational Analysis" explanation poorly (Score 1–2).

I.3 Case 3: Stereotypical Profiling (Hallucination)

Observation: High scores for Cognitive Style, but inconsistent scores for Behavioral Consistency. Multiple annotators flagged the reasoning as **"Generic/Templated"**, indicating a disconnect between high-level theory and text evidence.

- **Input Text:** "Dear INTP, I enjoyed our conversation... || I collect shoes because I like the sense of status. || Logic is a set of internally self-consistent principles. || I find cognition itself more interesting than motivation."

- **Path A (PsyPath):**
  Q1: Does the author frequently show a value for independence and autonomy?
  A: Yes. *Explanation: The author repeatedly expresses...*
  Q2: Are they willing to revise their views when new information appears?
  A: Yes.

Annotator	Pref (Q_1)	Q_2	Q_3	Q_4	Error Tags (Q_5)
#1	**Path A**	4	4	3	**Generic/Templated (A)**
#2	**Path A**	**2**	5	5	**Generic/Templated (A)**
#3	Path B	4	5	5	Weak Correlation (B)
#4	**Path A**	4	5	5	No Issues (F)
#5	**Path A**	3	5	4	**Generic/Templated (A)**

Table 7: Human evaluation details for Case 3.

Root Cause Diagnosis: Prior Knowledge Overreliance (The "Barnum Effect"). The text explicitly mentions "INTP." The model likely ignored specific details (e.g., "collecting shoes for status") and instead hallucinated a generic, textbook description of an INTP type.

I.4 Case 4: The Failure of Abstract Reasoning

Observation: Preference is split (2 votes for A, 2 for B, 1 Tie). The dominant criticism is **Genericity** and **Redundancy**. PsyPath fails to outperform the baseline when the text itself is abstract.

- **Input Text:** "Science is not perfect. || Rational thinking is very valuable... || First learn the whole system... || Religion should be critically examined."

- **Path A (PsyPath):**
  Q1: Does the author tend to understand problems starting from holistic systems?
  A: Yes.
  Q2: Do they show a sustained interest in abstract concepts?
  A: Yes.

Annotator	Pref (Q_1)	Q_2	Q_3	Q_4	Error Tags (Q_5)
#1	**Path B**	4	4	3	Redundancy (E)
#2	**Tie**	1	4	3	Generic (A), Redundancy (E)
#3	**Path B**	5	5	4	Redundancy (E)
#4	Path A	5	5	3	Weak Correlation (B)
#5	Path A	4	4	4	**Generic (A)**

Table 8: Human evaluation details for Case 4.

I.5 Summary of Failure Modes

Based on the analysis above, we categorize the primary failure modes of PsyPath into three types:

1. **Redundancy Loops:** The MCTS sometimes converges on a single salient trait and generates variations of the same question.

2. **Stereotypical Hallucination:** When explicit personality labels are present in the text, the model may default to "textbook definitions" ignoring contradictory details.

3. **Ambiguity in Tone Interpretation:** The model tends to take text literally, mistaking emotional ranting for cognitive rationality.