DiagLLM: Role-Playing with LLMs for Adaptive and Personalized Interview in Personality Detection

Anonymous ACL submission

Abstract

Personality detection aims to identify an individual's personality traits based on posts generated by users on the internet. Existing methods typically train a data-driven model based on the labeled data, which often requires substantial resources and time to collect. With the emergence of large language models (LLMs), zero-shot personality detection methods based on LLMs present an alternative solution. Recently, researchers have focused on incorporating external psychological knowledge through prompt engineering to enhance the detection effectiveness of a single LLM. However, the results are still far from satisfactory. Inspired by guided clinical interviews in psychology, this paper proposes DiagLLM, a Role-Playing method with LLMs for adaptive and personalized interview to identify a user's personality. Specifically, DiagLLM involves two LLM-driven agents: an AI psychologist and an AI user. The AI psychologist is responsible for adaptively generating a personalized questionnaire from the user's posts as interview questions to accurately capture the user's psychological information. Meanwhile, the AI user is tasked with learning user characteristics from the original posts and providing answers to the interview questions. After several turns of the interview conversation, the AI psychologist summarizes the dialogue history and provides the final personality classification results. Experiments on three benchmark datasets demonstrate that our method significantly improves the LLM's performance in zero-shot personality detection.

1 Introduction

Personality refers to an individual's internal characteristics, reflecting psychological differences and similarities among individuals. Personality detection aims to extract underlying personality traits from user-generated text, categorizing them according to psychological theories such as the Myers-Briggs Type Indicator (MBTI) and the Big Five personality taxonomy (Myers, 1987; Goldberg, 2013). This task is critical as it can provide valuable insights into an individual's personality for various user-centric applications, including character dialogues (Yang et al., 2021), psychotherapy (Bagby et al., 2016), recommendation systems (Jeong et al., 2020; Wu et al., 2024), and other domains (Zhang et al., 2025; Wang et al., 2023).

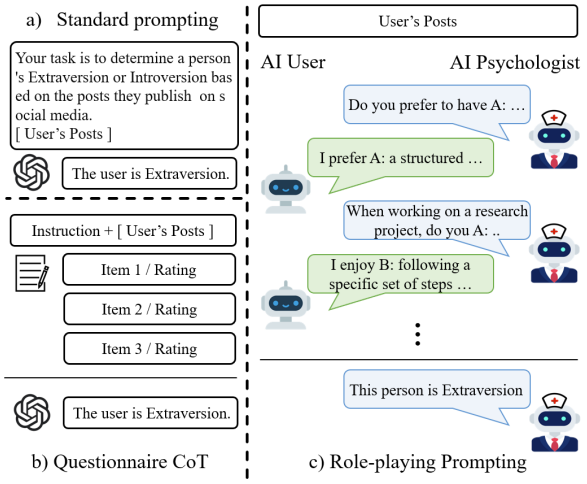

Figure 1: Existing zero-shot personality detection methods include: (a) standard zero-shot prompting, (b) using existing personality questionnaires as a chain-of-thought process, and (c) our proposed RPPD (Role-Playing enhanced Personality Detection) framework that involves an AI psychologist and an AI user conducting an adaptive and personalized interview.

Previous personality detection methods typically involve training or fine-tuning an expert model to identify a person's personality traits from texts using supervised learning. However, these methods require vast amounts of high-quality labeled data, typically derived from self-report questionnaires, which can be both time-consuming and costly to collect. Furthermore, this process raises significant concerns regarding user privacy, posing challenges to the scalability of data-driven approaches. These

issues are even more pronounced in low-resource scenarios, such as non-English languages, where available data is often scarce.

Recently, Large Language Models (LLMs) (Touvron et al., 2023; Hu et al., 2024b) have demonstrated strong performance across a wide range of text reasoning tasks, even in zero-shot settings. However, their direct application to personality detection remains limited. Personality detection requires both psychological knowledge and the ability to infer latent traits from user-generated text. Moreover, personality cues are often implicit and multidimensional, which makes reliable extraction difficult (Li et al., 2024; Shahnazari and Ayyoubzadeh, 2025). To address these challenges, prior studies have incorporated psychological knowledge into prompting strategies. For example, Yang et al. (2023b) used items from human-designed questionnaires as intermediate reasoning steps within a chain-of-thought framework, together with detailed label explanations to guide personality inference. Nevertheless, such approaches remain relatively shallow, failing to adequately capture individual differences or enable personalized analysis of users' posts.

Inspired by clinical interviews in psychology (Motowidlo et al., 1992; Magaldi and Berler, 2020), we propose a role-playing personality detection framework (RPPD) for adaptive, personalized interviewing. RPPD uses two LLM agents: an AI psychologist that generates personalized probing questions, and an AI user that extracts user characteristics from the original texts and answers those questions (Figure 1). RPPD proceeds in two stages: (1) the psychologist constructs a personalized questionnaire by generating probing questions grounded in psychological constructs and the user's posts; (2) the agents conduct a multi-round adaptive interview, after which the psychologist predicts the final personality label based on the dialogue history.

Our approach has two main advantages. First, by setting different roles, we can effectively decompose the complex personality detection task into subtasks, enabling the model to adaptively analyze and address specific aspects of the task more effectively. This targeted focus allows each role to specialize, leading to improved overall performance and more accurate results. Second, conventional self-report scales require users to respond to predetermined questions based on their personal experiences, and frequently fail to capture the nuanced information embedded in users' social media posts. This limitation arises because the static and generic nature of these scales cannot adequately align with the content of online posts. *For example, a scale might ask, "Do you enjoy large parties?" to assess extroversion. However, a user's posts express their social hobbies like participating in team sports like basketball games. In this case, the question does not align with their shared experiences, the user cannot provide a meaningful response, causing the scale to overlook their true personality as revealed online.* Thus, the tailored questions generated by our framework, which are grounded in the user's own posts, can more accurately capture their psychological features.

Overall, our main contributions can be summarized as follows:

- To the best of our knowledge, this is the first effort to enhance LLM's personality detection ability through a role-playing framework with LLMs for adaptive and personalized interview in zero-shot personality detection.
- Our framework involves two roles: the AI psychologist and the AI user. The AI psychologist adaptively generates personalized questions tailored to the user and conducts multiple rounds of questioning throughout the interview process. Meanwhile, the user generates contextually coherent responses, emulating user behavior by grounding its outputs in the historical posts of the target user.
- Experimental results on three datasets have demonstrated that our proposed method significantly outperforms the state-of-the-art zero-shot baselines, even achieving results comparable to those of trained expert models.

2 Related Work

2.1 Personality Detection

Traditional personality detection methods mostly rely on manually designed features. One common approach is to use psycholinguistic dictionaries such as LIWC (Tausczik and Pennebaker, 2010) and MRC (Coltheart, 1981), which categorize vocabulary into psychological labels and count their frequency as features.

With the advancement of deep learning, various neural models including CNN, LSTM, GRU and pre-trained language models like BERT have been widely adopted for end-to-end personality

detection (Xue et al., 2018; Tighe et al., 2016a; Lynn et al., 2020a; Jiang et al., 2020). Recent work further explores graph-based modeling to capture inter-post structural information for enhanced user representation, including dynamic graph construction and graph contrastive learning paradigms (Yang et al., 2023a; Zhu et al., 2022).

With the development of large language models (LLMs), recent approaches leverage their strong text comprehension capabilities to enhance detection performance: TAE (Hu et al., 2024a) uses LLMs to generate multi-perspective analysis texts for contrastive distillation, while other work explores graph structure optimization and parameter-efficient tuning for improved efficiency (Shen et al., 2025b,a).

Despite impressive performance, these data-driven methods remain heavily reliant on the quantity and quality of labeled training data. To alleviate this limitation, emerging work explores LLM-based zero-shot personality detection, which operates without labeled data and is particularly suitable for privacy-sensitive and low-resource settings. V Ganesan et al. (2023) first evaluated GPT-3's zero-shot capability and identified effective knowledge types to boost performance, while Yang et al. (2023b) enhanced GPT-3.5's personality reasoning by leveraging questionnaire items as intermediate chain-of-thought steps.

2.2 LLMs in Role-playing

As LLMs develop, extensive training corpora have endowed them with rich world knowledge. Additionally, instruction-based fine-tuning and in-context learning enable language models to effectively follow role-playing prompts, demonstrating great potential for diverse applications.

Existing work such as Role-LLM (Wang et al., 2024), ChatHaruhi (Li et al., 2023) and Character-LLM (Shao et al., 2023) have validated LLMs' strong capabilities in high-fidelity role-playing. Recently, role-playing LLMs have also been applied to psychological scenarios: Yang et al. (2025) proposed a personalized dialogue generation framework infused with distinct personality traits.

Differently, in this work, we present the first effort to enhance LLMs' personality detection capabilities through a multi-role playing framework for adaptive and personalized interview in personality detection.

3 Method

Personality detection can be defined as a multi-document multi-label classification task. Formally, given a set $\mathcal{X} = \{p_1, p_2, \ldots, p_N\}$ of N posts from a user. The target of this task is to predict T-dimensional personality traits of the user from the trait-specific label space $Y = \{\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_T\}$ based on $\mathcal{X}$. For the MBTI classification, $T = 4$, whereas for the Big Five classification, $T = 5$. In the context of zero-shot scenarios, the generative LLMs can accomplish the classification tasks through the instruction I and user's posts $\mathcal{X}$. The standard zero-shot prompting can be formulated as $\hat{y}^t = \mathcal{LLM}(\mathcal{I}, \mathcal{X})$. In our LLM based role-playing framework for adaptive and personalized interview in personality detection, we design two characters: AI psychologist $\mathcal{P}$ and AI user $\mathcal{U}$ with character prompts $\mathcal{I}_P$ and $\mathcal{I}_U$. We first instruct the AI psychologist $\mathcal{P}$ to adaptively generate relevant questions $\mathcal{Q}$ based on the context information $\mathcal{C}$ of posts $\mathcal{X}$. Then, we have AI user $\mathcal{U}$ imitate the user to answer these questions and obtain responses $\mathcal{A}$. After interview based on generated questions, the AI psychologist $\mathcal{P}$ is instructed to provide the final classification result based on the dialogue history. The overall framework RPPD is illustrated in Figure 2.

3.1 Personalized Questionnaire Generation

Self-reported questionnaires are traditional way to analyze one person's personality, but they often fail to capture subtle psychological features, as these fixed questions may not align with the content of users' posts. While LLMs can be directly used to analyze the content of users' posts and determine personality labels, they fail to leverage the psychological insights embedded in the traditional questionnaires. To address this issue, we introduce a novel role-playing framework with LLMs for adaptive and personalized interview in personality detection. In the interview, an AI psychologist adaptively generates relevant, user-tailored questionnaires. This approach leverages two key advantages of LLMs: 1) LLMs' strong text comprehension abilities to extract keywords reflecting personality and emotions, and 2) LLMs' extensive psychological knowledge, enabling them to adaptively generate context-aware, insightful questions that more accurately reflect personality traits.

Specifically, for each personality dimension d, we first instruct the AI psychologist to extract a

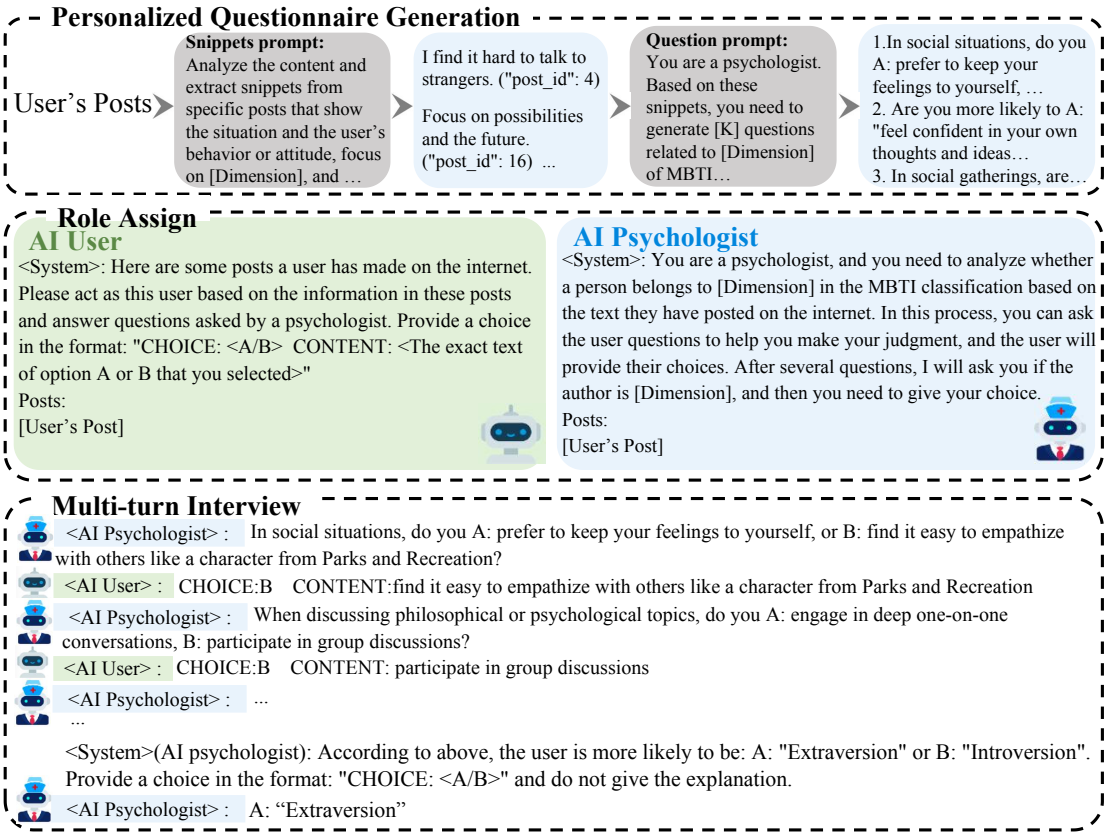

Figure 2: Overview of our LLM based Role-Playing framework for adaptive and personalized interview in personality detection (RPPD). RPPD comprises two LLM-driven agents: an AI Psychologist, which adaptively generates personalized and psychologically insightful questions, and an AI User, which simulates responses grounded in the user's posts.

set of snippets $\mathcal{C}^d$ from the user's posts. These snippets are extracted to capture the user's unique characteristics and interests relevant to the dimension d, and serve as crucial contextual information for subsequent personalized question generation:

$$\mathcal{C}^d = \mathcal{P}(\mathcal{I}_C, \mathcal{X}, d), \quad d \in \mathcal{D}. \tag{1}$$

Based on the guidance of contextual information, we use the AI psychologist to adaptively generate a personalized questionnaire. Specifically, for each psychological dimension, we employ in-context learning, using existing questions corresponding to the specific dimension as examples to generate K questions. This approach enables the generation of K new questions that are not only aligned with the specific dimension but also highly relevant to the user's profile, thereby enhancing the questionnaire's personalization and effectiveness. The formal representation can be expressed as:

$$\mathcal{Q}^d = \mathcal{P}(\mathcal{I}_Q, \mathcal{C}^d) \tag{2}$$

where $\mathcal{Q}^d$ denotes the set of personalized questions for dimension d, and $\mathcal{I}_C$ and $\mathcal{I}_Q$ are the corresponding prompts.

3.2 Multi-round Interview

After adaptively generating the personalized questionnaire, the AI psychologist $\mathcal{P}$ and the AI user $\mathcal{U}$ engage in a multi-round question-answering process. In each round, the AI psychologist first selects the next most informative question $\mathcal{Q}_t$ from the remaining candidate questions based on the current dialogue history, and then the AI user provides a response grounded in the user's posts and the interview history. Formally, for the t-th question, the set of historical question-answer pairs can be represented as:

$$\mathcal{H}_{t-1} = \{(\mathcal{Q}_0, \mathcal{A}_0), \ldots, (\mathcal{Q}_{t-1}, \mathcal{A}_{t-1})\} \tag{3}$$

The next question is selected by the AI psychologist based on the dialogue history and the remaining candidate questions $\mathcal{Q}_{\text{rem}}$:

$$\mathcal{Q}_t = \mathcal{P}\left(\mathcal{I}_{\text{sel}}, \mathcal{H}_{t-1}, \mathcal{Q}_{\text{rem}}\right) \tag{4}$$

The response from the AI user can be denoted as:

$$\mathcal{A}_t = \mathcal{U}\left(\mathcal{I}_U, \mathcal{X}, \mathcal{H}_{t-1}, \mathcal{Q}_t\right) \quad (5)$$

where $\mathcal{A}_t$ is a structured output containing a forced choice, supporting evidence quoted verbatim from the posts, and a confidence score. After all questions have been asked, the AI psychologist is instructed with $\mathcal{I}_J$ to make a final judgment for each personality dimension d based on the user's posts $\mathcal{X}$ and the full dialogue history $\mathcal{H}$:

$$\hat{y}^d = \mathcal{P}\left(\mathcal{I}_P, \mathcal{X}, \mathcal{H}_t, \mathcal{I}_J\right). \quad (6)$$

The overall procedure of RPPD is described in Algorithm 1 in the Appendix A.1.

4 Experiments

4.1 Datasets

Following previous studies (Zhu et al., 2022; Hu et al., 2024a; Yang et al., 2023b), we choose three widely used public datasets, two based on the MBTI taxonomy and one on the Big Five taxonomy. The Kaggle[1] and the Pandora[2] are based on MBTI taxonomy which divides personality into four dimensions: Introversion or Extroversion (I/E), Sensing or iNtuition (S/N), Thinking or Feeling (T/F), and Perception or Judging (P/J). Both datasets are collected from the online social platforms where each user contributes 40–50 posts. Following the experimental setup of the most related PsyCoT work (Yang et al., 2023b), we randomly sampled 200 instances from the test set for evaluation. Note that we also include a large-scale Kaggle test set consisting of 1,655 instances, to further assess the scalability of our method beyond small-sample evaluation. The Essays dataset is based on Big Five taxonomy that defines five personality traits, namely Openness, Conscientiousness, Extraversion, Agreeableness and Neuroticism.

We use Accuracy and Macro-F1 score as evaluation metrics on all datasets.

4.2 Baselines

We employ the following previous methods as baselines. A detailed description can be found in Appendix A.2:

Training-based Methods: SVM (Cui and Qi, 2017; Tighe et al., 2016b), Regression (Park et al., 2015), W2V+CNN (Rahman et al., 2019), SN+Attn (Lynn et al., 2020b), BERT (Keh et al., 2019), D-DGCN (Yang et al., 2023a) and TAE (Hu et al., 2024a).

LLM-based Methods: Standard Prompt, Zero-shot-CoT and PsyCoT (Yang et al., 2023b).

4.3 Implementation Details

In the experiments, considering cost and efficiency, we employed `GPT-4o` and set the temperature to 0 to make the outputs mostly deterministic for identical inputs. Following previous works (Yang et al., 2023a; Hu et al., 2024a; Yang et al., 2023b), in the cases of the Kaggle and Pandora, we removed words that match any personality label (e.g., INTP) to prevent information leaks. For fair comparison, we also limited each user to a maximum of 50 posts and restricted the length of each post to 70 words. For the Essays dataset, we also limited the length for the author's text to 3200 words. In the experiment, for each user, we chose to generate $K = 10$ questions and conducted experiments to assess the impact of the number of questions on the results.

4.4 Overall Results

Table 1 presents the experimental results of our proposed framework RPPD and the baselines on the Kaggle and Pandora datasets, which are based on the MBTI taxonomy, while Table 2 shows the results on the Essays dataset based on the Big Five taxonomy. From the two tables, we make the following observations.

First, our method outperforms all LLM-based baselines on both the MBTI taxonomy and the Big Five taxonomy across three datasets. Specifically, our method achieves higher average accuracy and Macro-F1 scores than PsyCoT on all three datasets. On the Kaggle dataset, improvements are 3.02% in average accuracy and 5.11% in Macro-F1. For the Pandora dataset, the gains are 1.89% and 2.95%, respectively. On the Essays dataset, our method outperforms PsyCoT by 5.67% in accuracy and 10.68% in Macro-F1. We believe the advantage comes from two main factors: (1) Our method benefits from the generated adaptive and personalized questionnaire, which extracts dimension-specific diagnostic evidence from a user's posts. (2) Assigning roles such as AI psychologist and AI user, and conducting guided interviews, allows each role to focus on its own specific subtask, thereby enhancing the overall classification results. Compared to the fine-tuned methods, our method outperforms traditional personality detection methods (SVM,

[1] https://www.kaggle.com/datasnaek/mbti-type
[2] https://psy.takelab.fer.hr/datasets/all

Table 1: Overall results of RPPD and baselines on the Kaggle and Pandora datasets. Methods above the line are training-based, while methods below the line are prompt-based. The best results of the LLM-based methods are in **bold** and the second best results are shown with underline. The note * indicates our method significantly outperforms the best prompt-based baseline PsyCoT ($p < 0.05$)

Methods	I/E		S/N		T/F		J/P		Average	
	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
Kaggle										
SVM	71.00	44.94	79.50	46.38	75.00	74.25	61.50	58.59	71.75	56.04
SN+Attn	–	62.34	–	57.08	–	69.26	–	63.09	–	62.94
BERT	77.30	62.50	84.90	54.04	78.30	77.93	69.50	68.80	77.50	65.82
DDGCN	78.10	70.26	84.40	60.66	79.30	78.91	73.30	71.73	78.78	70.39
TAE	79.98	70.90	83.04	66.21	81.14	81.17	71.74	70.20	78.97	72.07
Standard Prompt	77.00	69.33	85.00	58.33	79.00	78.69	**66.00**	55.82	76.75	65.55
Zero-shot-CoT	76.38	65.99	80.40	63.23	74.87	74.87	65.33	**59.76**	74.25	65.96
PsyCoT	78.50	**69.88**	84.50	55.96	73.00	72.96	**66.00**	50.47	75.50	62.32
RPPD	**80.40***	65.81	**86.93***	**69.20***	**82.41***	**81.82***	64.32	**52.90***	**78.52***	**67.43***
Pandora										
SVM	65.05	44.01	73.06	44.80	64.13	64.77	52.27	56.43	63.63	52.50
SN+Attn	–	56.98	–	54.78	–	60.95	–	54.81	–	56.88
BERT	71.52	54.54	84.06	52.73	70.99	66.47	58.47	56.26	71.25	57.50
DDGCN	72.54	59.78	84.78	56.36	70.72	68.61	59.59	56.26	71.90	60.25
TAE	76.46	62.57	84.39	61.01	73.81	69.28	60.47	59.34	73.78	63.05
Standard Prompt	71.50	56.67	57.00	48.31	68.50	62.68	61.00	56.55	64.50	56.05
Zero-shot-CoT	69.04	55.73	41.12	37.56	68.02	59.10	59.39	**57.91**	59.39	52.57
PsyCoT	62.81	56.27	**84.42**	53.81	**74.87**	66.16	57.79	47.81	69.97	56.01
RPPD	**80.90***	**62.71***	71.86	**55.35***	72.36	**68.20***	**62.31***	49.58*	**71.86***	**58.96***

Table 2: Overall results of RPPD and baselines on the Essays dataset. Methods above the line are training-based, while methods below the line are prompt-based.

Methods	AGR		CON		EXT		NEU		OPN		Average	
	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
SVM	51.78	47.50	51.99	52.00	51.22	49.20	51.09	50.90	54.05	52.40	52.03	50.40
Regression	50.96	51.01	54.65	54.66	55.06	55.06	57.08	57.09	59.51	59.51	55.45	55.47
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
Standard	59.11	57.98	57.49	49.55	59.92	54.39	**61.13**	**59.95**	55.87	49.11	58.70	54.20
Zero-shot-CoT	58.94	58.09	55.14	42.49	57.55	55.63	57.49	54.63	58.78	54.40	57.58	53.05
PsyCoT	61.53	61.11	**59.51**	48.80	53.44	48.85	56.28	52.80	42.91	32.76	54.73	48.86
RPPD	**63.97***	**63.35***	56.28	**52.80***	**60.73***	**60.66***	59.92*	59.88*	**61.13***	**60.99***	**60.40***	**59.54***

SN+Attn, W2V+CNN), even surpassing the pre-trained model Bert on three datasets. Compared to the specially designed expert models like TAE and DDGCN, our method still exhibits a discernible gap. Nevertheless, our method employs a zero-shot personality detection approach, thereby effectively circumventing the challenges associated with gathering labeled psychological data.

4.5 Ablation Study

We conduct ablation experiments on the Kaggle and Pandora datasets to verify the effectiveness of three key components in RPPD.

Snippet Extraction. As shown in Table 3, removing snippet extraction leads to consistent performance degradation across both datasets, with a substantially larger drop on Pandora than on Kaggle, demonstrating that extracting structured, dimension-focused evidence snippets from posts provides essential grounding for generating psychologically relevant questions.

Personalized Questionnaire. Instead of generating personalized questions from the extracted snippets, this variant replaces the question generation step with a fixed pre-defined question bank, which is shared across all users. The results show consistent accuracy degradation across both datasets, most notably on Pandora, confirming that personalized question generation, which tailors questions to each user's specific behavioral evi-

Table 3: Ablation study results on the Kaggle and Pandora datasets. We report the averaged Accuracy (Acc.) and Macro-F1 across four MBTI dimensions.

Methods	Kaggle		Pandora	
	Acc.	F1	Acc.	F1
RPPD	78.52	67.43	71.86	58.96
w/o snippet extraction	76.63	67.16	64.75	54.07
w/o personalized generation	76.02	67.84	64.21	56.52
w/ free-dialog	76.91	64.80	70.94	57.25

dence, captures more diagnostically informative psychological signals than generic fixed questionnaires.

Interview Format. In the standard RPPD, the AI user responds to each question in a structured forced-choice format, providing a choice (A or B), supporting evidence quoted from the posts, and a confidence score. In this variant, the AI user instead provides brief, spontaneous free-form responses without any structural constraint. The results show that the forced-choice format consistently outperforms free-dialog across both datasets, though the margin is more pronounced on Kaggle than on Pandora. This is likely because free-form responses, while more natural, introduce noise that obscures the psychological signal, whereas the structured format ensures that each response directly maps to a diagnostic dimension.

4.6 Evaluation on Large-Scale Kaggle Set.

Table 4 reports the results on the large-scale Kaggle test set containing 1,655 instances. RPPD achieves the highest average Accuracy of 78.30%, outperforming all baselines on this metric. Its average Macro-F1 of 65.60% is slightly below Standard Prompt (67.28%), a gap primarily driven by the I/E dimension as discussed below. These results suggest that RPPD generalizes effectively to large-scale evaluation beyond the 200-instance sampled subset. The largest gain is observed on the T/F dimension. PsyCoT exhibits a noticeable bias toward the Thinking (T) label, likely because its fixed questionnaire items are not grounded in users' posts and therefore are less effective at eliciting Feeling-related cues. In contrast, RPPD uses a personalized questionnaire together with a structured forced-choice interview to elicit user-specific behavioral evidence, yielding more balanced and more accurate predictions on this dimension. On the I/E dimension, RPPD underperforms Standard Prompt and PsyCoT in Macro-F1, which accounts for the gap in average F1. One possible reason is the severe class imbalance in the Kaggle dataset, where the Extroversion class is heavily underrepresented. This imbalance makes it difficult for the adaptive interview to accumulate sufficient evidence for Extroversion, causing the model to default toward Introversion predictions.

4.7 Quality of Generated Personalized Questionnaire

LLMs, through pre-training on extensive text corpora, appear to acquire substantial knowledge of personality-related language patterns. To examine this, we first evaluate GPT-4o's ability to classify items from two conventional personality assessments—the Myers–Briggs Type Indicator and the 16 Personalities Questionnaire—into their corresponding MBTI dimensions. As reported in Table 5, GPT-4o achieves near-perfect classification accuracy (26/26 and 45/48), indicating that LLMs can effectively distinguish the linguistic and semantic features associated with different personality dimensions.

Inspired by the evaluation of traditional questionnaires based on their validity and reliability (Taherdoost, 2016), we further evaluate the quality of our generated questionnaires along the same two dimensions. Since GPT-4o demonstrates high accuracy in classifying established questionnaire items into their target dimensions, we adopt it as an automated evaluator for our generated questions.

For validity, we sample 100 generated questions per dimension and use GPT-4o to classify them into their target MBTI dimensions. Table 5 shows that the generated questions achieve high classification accuracy (384/400), suggesting that they are well-aligned with the intended personality dimensions. We note that using the same family of model for both generation and evaluation may inflate this metric; however, the near-ceiling accuracy on established questionnaires (Table 5) suggests that GPT-4o's classification behavior is grounded in meaningful personality constructs rather than superficial generation artifacts.

To assess reliability, we conduct two standard tests. For test-retest reliability, we shuffle the order of each user's posts and regenerate the questionnaire, then compare the resulting personality predictions with those obtained from the original post order; this measures the stability of the full pipeline (questionnaire generation and personality prediction) against variation in input order. For in-

Table 4: Performance comparison on the large-scale Kaggle test set.

Methods	I/E		S/N		T/F		J/P		Average	
	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
Standard Prompt	81.11	71.42	84.17	61.61	77.87	77.85	67.19	58.25	77.59	67.28
Zero-shot-CoT	80.69	70.31	77.41	61.99	73.20	73.05	65.71	59.35	74.25	66.17
PsyCoT	**82.31**	**72.56**	**86.22**	57.04	71.18	70.47	65.94	51.95	76.41	63.01
RPPD	80.94	65.13	85.43	**63.01**	**80.22**	**80.15**	**66.61**	**54.05**	**78.30**	**65.60**

Table 5: The classification results of GPT-4o on the traditional questionnaires and our generated personalized questionnaire. We report the number of correctly classified questions out of the total questions.

Questionnaire	I/E	S/N	T/F	P/J	Total
Myers-Briggs Personality Test	8/8	7/7	3/3	8/8	26/26
16 personalities Quest.	12/12	10/12	12/12	11/12	45/48
Our Generated Personalized Quest.	95/100	92/100	97/100	100/100	384/400

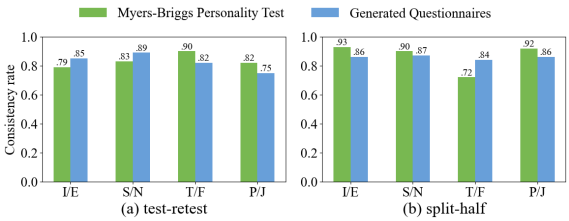

Figure 3: Reliability test results of the Myers-Briggs Personality Test and the generated questionnaire. Consistency rate refers to the proportion of identical predicted labels within each reliability test.

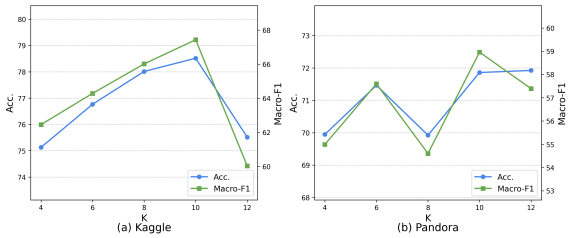

Figure 4: Impact of question number on the Kaggle dataset and the Pandora dataset.

ternal consistency, we perform a split-half test: we partition questions into odd- and even-numbered subsets, derive a personality prediction from each subset independently, and compute the proportion of users who receive identical labels from both halves. As shown in Figure 3, both consistency rates are comparable to those obtained with the standard Myers-Briggs Personality Test under the same protocol, suggesting that our generated questionnaires achieve a level of reliability on par with this widely used instrument.

4.8 The Impact of Number of Questions

We conduct experiments to examine the impact of the number of questions in a personalized questionnaire on the model's ability to capture users' psychological information. Varying the number of questions from 4 to 12, we evaluate performance on the Kaggle and Pandora datasets, with results shown in Figure 4. Performance improves as the number of questions increases, peaking at K=10 for both datasets. Beyond this point, however, performance declines, likely because the essential information from the limited user posts has already been captured, and additional questions offer diminishing returns.

5 Conclusion

In this work, we propose a novel LLM-based role-playing framework that simulates an adaptive and personalized clinical interview in personality detection. Inspired by the clinical interview, our approach involves two LLM-driven agents: an AI psychologist and an AI user. The AI psychologist generates personalized probing questions based on dimension-specific diagnostic evidence snippets extracted from the user's posts, while the AI user adaptively answers these questions based on the interview history and the online posts. After multiple rounds, the AI psychologist then determines the user's personality based on the whole interview question-answer pairs. Experiments on three datasets show that our framework outperforms existing zero-shot methods and performs comparably to fine-tuned expert models. This work is the first effort to use role-playing with LLMs for adaptive and personalized interviews in personality detection, offering a new perspective for zero-shot personality detection.

Limitations

The prospective limitations of our method are discussed as follows. Our method involves a two-stage design, including the generation of personalized questionnaires and interaction between roles, which introduces more reasoning costs compared to standard prompting methods. Our research has been conducted exclusively in English contexts, focusing on Western personality trait theories. Studies in other languages and cultural contexts require further exploration.

Ethics Statement

Personality detection has been a longstanding research task in the field of psycholinguistics. This study primarily aims to investigate and enhance the capabilities of the LLM in this specific task, not to develop a tool for privacy intrusion. Both the Kaggle and the Pandora datasets are publicly accessible and all anonymized, aiming to minimize privacy concerns. We strictly follow data usage policies and state that research or applications stemming from this work are intended solely for research purposes. Any use of this method to infer an individual's personality is prohibited.

References

R. Michael Bagby, Tara M. Gralnick, Nadia Al-Dajani, and Amanda A. Uliaszek. 2016. The role of the five-factor model in personality assessment and treatment planning. *Clinical Psychology: Science and Practice*, 23(4):365–381.

Max Coltheart. 1981. The mrc psycholinguistic database. *The Quarterly Journal of Experimental Psychology Section A*, 33(4):497–505.

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*.

Lewis R Goldberg. 2013. An alternative "description of personality": The big-five factor structure. In *Personality and Personality Disorders*, pages 34–47.

Linmei Hu, Hongyu He, Duokang Wang, Ziwang Zhao, Yingxia Shao, and Liqiang Nie. 2024a. 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.

Linmei Hu, Zeyi Liu, Ziwang Zhao, Lei Hou, Liqiang Nie, and Juanzi Li. 2024b. A survey of knowledge enhanced pre-trained language models. *IEEE Transactions on Knowledge and Data Engineering*, 36(4):1413–1430.

Chi-Seo Jeong, Jong-Yong Lee, and Kye-Dong Jung. 2020. Adaptive recommendation system for tourism by personality type using deep learning. *International journal of internet, broadcasting and communication*, 12(1):55–60.

Hang Jiang, Xianzhe Zhang, and Jinho D Choi. 2020. Automatic text-based personality recognition on monologues and multiparty dialogues using attentive networks and contextual embeddings (student abstract). In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 13821–13822.

Sedrick Scott Keh, I Cheng, and 1 others. 2019. Myers-briggs personality classification and personality-specific language generation using pre-trained language models. *arXiv preprint arXiv:1907.06333*.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Bohan Li, Jiannan Guan, Longxu Dou, Yunlong Feng, Dingzirui Wang, Yang Xu, Enbo Wang, Qiguang Chen, Bichen Wang, Xiao Xu, and 1 others. 2024. Can large language models understand you better? an mbti personality detection dataset aligned with population traits. *arXiv preprint arXiv:2412.12510*.

Cheng Li, Ziang Leng, Chenxi Yan, Junyi Shen, Hao Wang, Weishi Mi, Yaying Fei, Xiaoyang Feng, Song Yan, HaoSheng Wang, and 1 others. 2023. Chatharuhi: Reviving anime character in reality via large language model. *arXiv preprint arXiv:2308.09597*.

Veronica Lynn, Niranjan Balasubramanian, and H. Andrew Schwartz. 2020a. Hierarchical modeling for user personality prediction: The role of message-level attention. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5306–5316.

Veronica Lynn, Niranjan Balasubramanian, and H Andrew Schwartz. 2020b. Hierarchical modeling for user personality prediction: The role of message-level attention. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 5306–5316.

Danielle Magaldi and Matthew Berler. 2020. Semi-structured interviews. *Encyclopedia of personality and individual differences*, pages 4825–4830.

Stephan J Motowidlo, Gary W Carter, Marvin D Dunnette, Nancy Tippins, Steve Werner, Jennifer R Burnett, and Mary Jo Vaughan. 1992. Studies of the structured behavioral interview. *Journal of applied psychology*, 77(5):571.

Isabel Briggs Myers. 1987. *Introduction to type: A description of the theory and applications of the Myers-Briggs Type Indicator*. Consulting Psychologists Press.

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

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*, pages 1–6.

Kourosh Shahnazari and Seyed Moein Ayyoubzadeh. 2025. Who are you behind the screen? implicit mbti and gender detection using artificial intelligence. *arXiv preprint arXiv:2503.09853*.

Yunfan Shao, Linyang Li, Junqi Dai, and Xipeng Qiu. 2023. Character-LLM: A trainable agent for role-playing. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13153–13187.

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

Lingzhi Shen, Yunfei Long, Xiaohao Cai, Guanming Chen, Yuhan Wang, Imran Razzak, and Shoaib Jameel. 2025b. Ll4g: Self-supervised dynamic optimization for graph-based personality detection. *arXiv preprint arXiv:2504.02146*.

Hamed Taherdoost. 2016. Validity and reliability of the research instrument; how to test the validation of a questionnaire/survey in a research. *International Journal of Academic Research in Management*, 5:28–36.

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

Edward P Tighe, Jennifer C Ureta, Bernard Andrei L Pollo, Charibeth K Cheng, and Remedios de Dios Bulos. 2016a. Personality trait classification of essays with the application of feature reduction. In *SAAIP@ IJCAI*, pages 22–28.

Edward P Tighe, Jennifer C Ureta, Bernard Andrei L Pollo, Charibeth K Cheng, and Remedios de Dios Bulos. 2016b. Personality trait classification of essays with the application of feature reduction. In *SAAIP@ IJCAI*, pages 22–28.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, and 1 others. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Adithya V Ganesan, Yash Kumar Lal, August Nilsson, and H. Schwartz. 2023. Systematic evaluation of GPT-3 for zero-shot personality estimation. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 390–400, Toronto, Canada. Association for Computational Linguistics.

Quanxiu Wang, Xinlei Cao, Jianyong Wang, and Wei Zhang. 2023. Knowledge-aware collaborative filtering with pre-trained language model for personalized review-based rating prediction. *IEEE Transactions on Knowledge and Data Engineering*, 36(3):1170–1182.

Zekun Moore Wang, Zhongyuan Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhan Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Man Zhang, and 1 others. 2024. Rolellm: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. In *Findings of the Association for Computational Linguistics*, pages 14743–14777.

Yiqing Wu, Ruobing Xie, Yongchun Zhu, Fuzhen Zhuang, Xu Zhang, Leyu Lin, and Qing He. 2024. Personalized prompt for sequential recommendation. *IEEE Transactions on Knowledge and Data Engineering*, 36(7):3376–3389.

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.

Runzhe Yang, Jingxiao Chen, and Karthik Narasimhan. 2021. Improving dialog systems for negotiation with personality modeling. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 681–693.

Tao Yang, Jinghao Deng, Xiaojun Quan, and Qifan Wang. 2023a. Orders are unwanted: Dynamic deep graph convolutional network for personality detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 13896–13904.

Tao Yang, Tianyuan Shi, Fanqi Wan, Xiaojun Quan, Qifan Wang, Bingzhe Wu, and Jiaxiang Wu. 2023b. PsyCoT: Psychological questionnaire as powerful chain-of-thought for personality detection. In *Findings of the Association for Computational Linguistics: EMNLP*, pages 3305–3320.

Tao Yang, Yuhua Zhu, Xiaojun Quan, Cong Liu, and Qifan Wang. 2025. Psyplay: Personality-infused role-playing conversational agents. *arXiv preprint arXiv:2502.03821*.

Puning Zhang, Miao Fu, Rongjian Zhao, Hongbin Zhang, and Changchun Luo. 2025. Pure: Personality-coupled multi-task learning framework for aspect-based multimodal sentiment analysis. *IEEE Transactions on Knowledge and Data Engineering*, 37(1):462–477.

Yangfu Zhu, Linmei Hu, Xinkai Ge, Wanrong Peng, and Bin Wu. 2022. Contrastive graph transformer network for personality detection. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*, pages 4559–4565.

A Appendix

A.1 The RPPD Algorithm

Algorithm 1: Adaptive and Personalized Interview

Input : Characters: $\mathcal{P} \leftarrow \mathcal{LLM}^{\mathcal{I}_P}$, $\mathcal{U} \leftarrow \mathcal{LLM}^{\mathcal{I}_U}$; Instructions: $\mathcal{I}_C$, $\mathcal{I}_Q$, $\mathcal{I}_{\text{sel}}$, $\mathcal{I}_J$; Personality dimensions: $\mathcal{D}$; User posts: $\mathcal{X}$.

Output : Personality traits: $\{\hat{y}^d\}_{d\in\mathcal{D}}$

for $d \in \mathcal{D}$ **do**
  $\mathcal{C}^d = \mathcal{P}(\mathcal{I}_C, \mathcal{X}, d)$
  $\mathcal{Q}^d = \mathcal{P}(\mathcal{I}_Q, \mathcal{C}^d)$
  Initialize dialogue history $\mathcal{H}_0 \leftarrow \emptyset$; $\mathcal{Q}_{\text{rem}} \leftarrow \mathcal{Q}^d$
  while $\mathcal{Q}_{\text{rem}} \neq \emptyset$ **do**
    $\mathcal{Q}_t = \mathcal{P}(\mathcal{I}_{\text{sel}}, \mathcal{H}_{t-1}, \mathcal{Q}_{\text{rem}})$
    $\mathcal{Q}_{\text{rem}} \leftarrow \mathcal{Q}_{\text{rem}} \setminus \{\mathcal{Q}_t\}$
    $\mathcal{A}_t = \mathcal{U}(\mathcal{I}_U, \mathcal{X}, \mathcal{H}_{t-1}, \mathcal{Q}_t)$
    $\mathcal{H}_t \leftarrow \mathcal{H}_{t-1} \cup \{(\mathcal{Q}_t, \mathcal{A}_t)\}$
  end while
  $\hat{y}^d = \mathcal{P}(\mathcal{I}_P, \mathcal{X}, \mathcal{H}_t, \mathcal{I}_J)$
end for
return $\{\hat{y}^d\}_{d\in\mathcal{D}}$

A.2 Details of Baselines

The detailed descriptions of the baselines are as follows: **SVM** (Cui and Qi, 2017; Tighe et al., 2016b): These approaches utilize TF-IDF or LIWC for feature extraction and employ the SVM classifier for categorization.

Regression (Park et al., 2015): Uses a regression model with labels ranging from 0 to 1 and the quantile discretization method for classification.

W2V+CNN (Rahman et al., 2019): Encodes the context using the Word2Vec algorithm and then applies a CNN to obtain the context representation.

SN+Attn (Lynn et al., 2020b): A hierarchical network uses GRU and attention mechanisms to encode sequences for user representation.

BERT (Keh et al., 2019): It uses BERT to encode each post individually, then employs mean pooling to derive the user representation.

TrigNet: TrigNet constructs a tri-graph for each user consisting of posts, words, and word categories, based on LIWC dictionary. It then employs a modified GAT to encode the graph and uses average pooling to obtain a user representation.

D-DGCN (Yang et al., 2023a): Uses a dynamic graph to autonomously learn post structures and employs GCN to obtain user representation.

TAE (Hu et al., 2024a): TAE utilizes LLMs for psycho-linguistic augmentation of user posts and uses contrastive learning to improve the post encoder's ability to capture psychological features. A soft-label training approach based on label semantics is also designed to mitigate model overconfidence.

Standard Prompt: Standard prompting refers to directly inputting classification instructions and user's posts in zero-shot setting personality detection.

Zero-shot-CoT: It refers to zero-shot CoT method (Kojima et al., 2022), which inserts a reasoning prompt "Let's think step by step." before inferring user's personality traits.

PsyCoT (Yang et al., 2023b): PsyCoT utilizes items from an existing questionnaire[3] as intermediate steps in the CoT. The LLM first assigns a score to each item, and then delivers the final classification result.

A.3 Prompt Specifications

The prompt employed for generating snippets in our method is as follows:

These are posts published by a user on the internet. Analyze the content and summarize snippets or phrases that reflect the user's characteristics, interests, and behaviors. Please provide the summary in JSON format, with categories for different aspects such as Interests, Personality Traits, Emotional Tone and Interaction with Others. Make sure the snippets are as comprehensive as possible. Here are the posts:

[User's Posts]

The prompt used for generating personalized questions in our method is:

You are a psychologist. There are some snippets here. Based on these snippets, you need to gen-

[3] https://wedgworthleadership.com/wpcontent/uploads/2016/08/Myers-Briggs-Personality-Test.pdf

erate 10 questions related to the Introversion and Extraversion dimension of the MBTI classification. Here, definitions of Introversion and Extraversion are provided for you, as well as some questions collected from a survey questionnaire as examples. Please return in JSON format and do not repeat the examples.

Definitions:

[Personality trait's Definitions]

Snippets:

[User's Snippets]

where "*Definition*" refers to the definition of the corresponding MBTI personality traits[4] and Big Five personality traits[5]. The specific definitions are shown in Figure A1 and Figure A2.

A.4 Case Study

We present samples from the Kaggle dataset that are correctly classified by the RPPD framework but not by zero-shot-CoT, in order to qualitatively analyze the effectiveness of RPPD. The similarities and differences in the analysis process are shown in Figure A3. From the figure, we can see that CoT focuses on only a small amount of information in the text and incorrectly analyzes the user's tendency on the Introversion and Extroversion dimensions. This is because the LLM does not select the appropriate key information or provide the correct judgment of the user's tendency. In contrast, within the RPPD framework, the AI psychologist first identifies all key information in the user's text, then selects suitable content to generate personalized questions. The AI user is responsible for accurately answering each question. Finally, the AI psychologist summarizes all the information and provides the correct classification.

[4] https://www.myersbriggs.org/

[5] https://www.elaborer.org/cours/psy7124/lectures/John2008.pdf

Introversion: People who prefer Introversion feel energized when reflecting on concepts and ideas in their inner world. Their attention is naturally drawn in this inward direction.
Extraversion: People who prefer Extraversion feel energized by interaction in the outer world of people and things. Their attention is naturally drawn in this outward direction.

Sensing: People who prefer Sensing pay more attention to information that is concrete and tangible. They focus on what is, by paying attention to specific details and facts. They rely on perception through their five senses (trust experience).
Intuition: People who prefer Intuition pay more attention to the patterns and possibilities in the information they receive. They focus on what could be, by looking at the big picture and making connections between the facts. They use their five senses, too, but rely on perception through insights and hunches (trust inspiration).

Thinking: People who prefer Thinking put more weight on objective principles and impersonal facts when decision-making. They focus on logic and analysis.
Feeling: People who prefer Feeling put more weight on personal concerns and the people involved when decision making. They focus on values and relationships.

Judging: People who prefer Judging like a more structured and organized lifestyle. They like to control their environment by making plans or at least knowing what the plans are when others make them. Interaction with the outside world is through their decision-making (judging) mental process of Thinking or Feeling. Making decisions, coming to closure, and then moving on is important to people with a Judging preference.
Perceiving: People who prefer Perceiving like a more flexible and open-ended lifestyle. Rather than control their environment, they want to experience it through exploring options. Interaction with the outside world is through their information gathering (perceiving) mental process of Sensing or Intuition. Staying open to new information, last minute options, and being adaptable is important to people with a Perceiving preference.

Figure A1: Definition of MBTI personality traits.

Extraversion: Implies an energetic approach toward the social and material world and includes traits such as sociability, activity, assertiveness, and positive emotionality.

Agreeableness: Contrasts a prosocial and communal orientation toward others with antagonism and includes traits such as altruism, tender-mindedness, trust, and modesty.

Conscientiousness: Describes socially prescribed impulse control that facilitates task- and goal-directed behavior, such as thinking before acting, delaying gratification, following norms and rules, and planning, organizing, and prioritizing tasks.

Neuroticism: Contrasts emotional stability and even-temperedness with negative emotionality, such as feeling anxious, nervous, sad, and tense.

Openness: Describes the breadth, depth, originality, and complexity of an individual's mental and experiential life.

Figure A2: Definition of Big Five personality traits.

User's Posts

"Pardon the thread necromancy, but I've been thinking about this a lot lately and Google found this thread for me. This is me, only probably more so. I'm really struggling with whether I'm...",
"I really don't know; it's something I hadn't even considered until I saw the way is sort of a mirror image of along the I/E axis. Ne and Ni are both really high for me; I can't remember..." ,
"[x] You have a tablet device What if I choose not to drink alcohol sometimes? I had to Google lifestyle block, so just to save other non-Aotearoans the trouble it apparently means a small..." ,
"I think it's quite likely that s are much the same way, and the only I know is definitely like that. For some reason, though, I have encountered claims in the MBTI literature that s are..." ,
"You know, I never noticed until now that there was no t in your name!" ,
"Yeah, it's definitely more the falling for me, too. If I'm in an enclosed space like an airplane or otherwise secured in a way that falling wouldn't be possible without something going extremely,..." ,
"I think it would be really nice to have a way to interact with other people on this forum that's different from making posts in topical threads. Something more spontaneous and freeform would be..." ,
"From French salon to Monty Python clip in the first posts. This is an thread." ,
"My daughter is and there are times when I feel very much that she and I relate very well, other times when I think she relates better to her mom (), and still other times when my wife and I..." ,
"I would like to find a way to do this but I'm not sure I want to use Facebook for it. It might be fun, though. I dunno. I'm glad you brought it up and I don't think you should feel dumb." ,
"Here I thought you were going to tell us that Trent Reznor is an ! :) "
…

Zero-shot CoT

Based on the author's posts, they seem to engage in discussions, respond to other users' comments, express their thoughts openly, and participate actively in various topics. They also mention interacting with others on forums and expressing opinions freely. These behaviors align more closely with traits of Extraversion rather than Introversion.

RPPD

< AI Psychologist >When discussing religion, does the author prefer to A: "reflect on their own beliefs internally" or B: "engage in discussions with others"?. Provide a choice ID in the format: "CHOICE: <A/B>", and do not give the explanation.
< AI User >CHOICE: B
< AI Psychologist >When learning about scientific concepts, does the author prefer to A: "study them on their own" or B: "discuss them with others"?
< AI User >CHOICE: A
< AI Psychologist > When considering topics related to homosexuality, does the author prefer to A: "reflect on them privately" or B: "engage in open discussions with others"?
< AI User >CHOICE: B
< AI Psychologist > When facing medical procedures, does the author prefer to A: "process the information internally" or B: "seek support and advice from others"?
< AI User >CHOICE: A
< AI Psychologist >When exploring cognitive functions, does the author prefer to A: "analyze them independently" or B: "engage in group discussions"?
< AI User >CHOICE: A
…

Figure A3: Case study for Standard, Zero-shot-CoT, and RPPD frameworks.