Ask, Answer, and Detect: Role-Playing LLMs for Personality Detection with Question-Conditioned Mixture-of-Experts

Yifan Lyu
The Hong Kong University of Science and Technology (Guangzhou)
ylyu634@connect.hkust-gz.edu.cn

Liang Zhang*
The Hong Kong University of Science and Technology (Guangzhou)
liangzhang@hkust-gz.edu.cn

Abstract

Understanding human personality is crucial for web applications such as personalized recommendation and mental health assessment. Existing studies on personality detection predominantly adopt a "posts → user vector → labels" modeling paradigm, which encodes social media posts into user representations for predicting personality labels (*e.g.*, MBTI labels). While recent advances in large language models (LLMs) have improved text encoding capacities, these approaches remain constrained by limited supervision signals due to label scarcity, and underspecified semantic mappings between user language and abstract psychological constructs. We address these challenges by proposing ROME, a novel framework that explicitly injects psychological knowledge into personality detection. Inspired by standardized self-assessment tests, ROME leverages LLMs' role-play capability to simulate user responses to validated psychometric questionnaires. These generated question-level answers transform free-form user posts into interpretable, questionnaire-grounded evidence linking linguistic cues to personality labels, thereby providing rich intermediate supervision to mitigate label scarcity while offering a semantic reasoning chain that guides and simplifies the text-to-personality mapping learning. A question-conditioned Mixture-of-Experts module then jointly routes over post and question representations, learning to answer questionnaire items under explicit supervision. The predicted answers are summarized into an interpretable answer vector and fused with the user representation for final prediction within a multi-task learning framework, where question answering serves as a powerful auxiliary task for personality detection. Experiments on two real-world datasets show that ROME outperforms state-of-the-art baselines, achieving relative improvements of **13.60%** and **19.72%** on Kaggle and Pandora, respectively. Our code is available at https://github.com/JasmineLyf/ROME.

1 Introduction

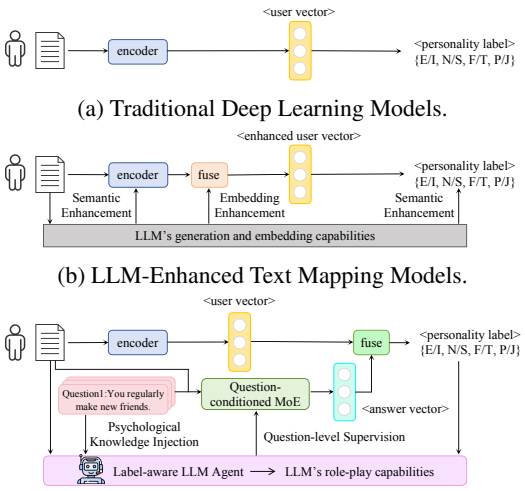
Understanding human personality is essential for explaining individual differences in cognition, emotion, and behavior and serves as the foundation for web applications such as personalized recommendation, mental-health assessment, and career development [17]. Among typological instruments, the Myers–Briggs Type Indicator (MBTI) remains one of the most widely adopted, categorizing individuals into sixteen personality types along four binary dimensions [26]. With the rapid growth of social media, individuals increasingly share their thoughts and experiences online, creating a rich source of behavioral signals for understanding human personality [10]. There is now substantial evidence that digital traces and linguistic expressions on social media can reveal users' personality traits [31, 49]. These developments have spurred the emerging task of *personality detection*, which

*Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

aims to infer personality labels (*i.e.*, MBTI labels) from social media posts automatically and has attracted substantial research attention in recent years [35].

These aforementioned studies typically formulate the task as a text classification problem following the "posts → user vector → labels" paradigm as illustrated in Figure 1(a), where the crucial challenge is to learn a mapping function that encodes user posts into a single user representation. In this line of work, various strategies have been proposed. Traditional approaches primarily rely on handcrafted features, such as those derived from the Linguistic Inquiry and Word Count (LIWC) lexicon [29, 25]. With the advent of deep learning, subsequent approaches have shifted from handcrafted features to neural architectures, which extract user representations from raw text automatically. For example, building on pretrained language models, some studies [16, 13] adopt BERT [6] to encode user content, with strategies ranging from post-level encoding to concatenating multiple posts into one sequence. Some alternative methods [44, 52, 47] adopt a cross-post perspective, using BERT to extract semantic features and graph neural networks to further enhance post representations. More recent extensions further enrich user representations via psychological guidance [45] or multi-view modeling [51]. For example, MvP [51] leverages multiple encoders with consistency regularization to help capture diverse perspectives of user posts. Although these methods have achieved notable progress, their limited representational capacity often produces suboptimal embeddings that fail to fully exploit the rich personality signals embedded in user posts, especially when the posts are heterogeneous and noisy.



(a) Traditional Deep Learning Models.

(b) LLM-Enhanced Text Mapping Models.

(c) Our LLM Role-Play Enhanced, Questionnaire-Grounded Model.

Figure 1: Overview of modeling paradigms for personality detection.

Recently, large language models (LLMs) have shown remarkable text understanding abilities and strong performance across a wide range of natural language processing tasks [15, 3, 40, 41] compared with traditional smaller models, which in turn has motivated their application to the personality detection task. The typical modeling paradigm is illustrated in Figure 1(b). Specifically, TAE [11] leverages the generative capabilities of LLMs to produce multi-perspective textual analyses of posts and labels within a contrastive learning framework, thereby enriching post semantics and strengthening user representations. Building on TAE, ETM [2] further employs LLMs to exploit their strong embedding capacity, augmenting the user vector with long-text embeddings directly and strengthening the mapping between user vectors and personality labels. Extending this line, LL4G [34] utilizes LLM embeddings to model inter-post relationships for dynamic graph construction, enhancing graph-based user representation learning. More recently, EmoPerso [33] incorporates LLM-based generative mechanisms including style-conditioned paraphrasing and contextual feature completion to construct and enrich emotion-aware user representations and achieves state-of-the-art performance. However, despite these advances, existing LLM-based approaches largely adhere to the traditional "posts → user vector → labels" paradigm, incorporating only limited augmentations. The semantic mapping from user language to abstract psychological constructs is achieved through simple text encoding and relies primarily on supervision from coarse-grained personality labels which are often scarce, costly to collect and prone to privacy concerns.

To address this limitation, we observe that humans often infer their own personality by completing standardized self-assessment tests. Inspired by this process, we posit that structured, well-validated psychometric questionnaires provide a natural bridge between user language and abstract psychological constructs. Beyond coarse-grained personality labels, this questionnaires decompose each label into multiple semantically meaningful questions (*i.e.*, psychometric items), and their fine-grained question–answer design provides rich intermediate supervision to mitigate label scarcity while offering a semantic reasoning chain that simplifies the text-to-personality mapping learning.

Building on this insight, we propose **ROME** (ROle-play enhanced question-conditioned Mixture-of-Experts), an LLM role-play enhanced, questionnaire-grounded framework for personality detection, with its modeling paradigm illustrated in Figure 1(c). Our key idea is to treat the standardized personality questionnaire as a structured retrieval source and design a label-aware LLM agent to role-play each user, generating question-level answers from their free-form posts. These answers constitute *questionnaire-grounded evidence* that explicitly links linguistic cues to specific personality labels, bringing psychological knowledge directly into the modeling pipeline. We then design a question-conditioned Mixture-of-Experts (MoE) module that jointly routes over posts and question representations and learns to answer the psychometric questionnaire items under explicit supervision. These predicted answers are then summarized into an answer vector, a compact representation that captures the model's question-by-question reasoning chain for personality detection. This vector is finally fused with the encoder's user representation for prediction within a multi-task learning framework. In summary, our main contributions are as follows:

- Conceptually, we are the first to leverage the strong role-play capabilities of LLMs to explicitly inject psychological knowledge into personality detection and formulate an interpretable, questionnaire-grounded modeling paradigm.
- Methodologically, we design a multi-task learning framework that treats question answering as a powerful auxiliary task for personality detection and develop a question-conditioned Mixture-of-Experts (MoE) architecture that learns to answer questionnaire items through a post–question joint routing mechanism.
- Experimentally, extensive results on two real-world datasets demonstrate that ROME significantly and consistently outperforms state-of-the-art baselines, achieving substantial improvements of 13.60% on the KAGGLE dataset and 19.72% on the PANDORA dataset.

2 Method

2.1 Problem Definition

Personality detection from social media can be formulated as a *user-level, multi-label* classification problem. Given a collection of posts authored by a user u, denoted as $\mathcal{P}_u = \{p_{u,1}, \ldots, p_{u,N_u}\}$, where N_u represents the total number of posts, the goal of this task is to infer the user's *m*-dimensional personality profile $\hat{\mathbf{y}}_u = \left(\hat{y}_u^{(1)}, \hat{y}_u^{(2)}, \cdots, \hat{y}_u^{(m)}\right)$ based on posts $\mathcal{P}_u$.

Under the MBTI taxonomy, the personality dimension set is $\mathcal{M} = \{\mathrm{I/E}, \mathrm{S/N}, \mathrm{T/F}, \mathrm{P/J}\}$. Each dimension $m \in \mathcal{M}$ consists of two opposing trait poles, and the ground-truth label $y_u^{(m)} \in \{0, 1\}$ indicates which pole the user u aligns with. The objective is therefore to learn the mapping function from $\mathcal{P}_u$ to $\hat{\mathbf{y}}_u$:

$$f: \mathcal{P}_u \longmapsto \hat{\mathbf{y}}_u = \left(\hat{y}_u^{(m)}\right)_{m \in \mathcal{M}}.$$

2.2 Model Overview

To facilitate and ease the learning process from the free-form user language space $\mathcal{P}_u$ to the abstract psychological construct space $\hat{\mathbf{y}}_u$, we incorporate well-validated psychometric questionnaires as a natural intermediate bridging layer, drawing inspiration from the standardized self-assessment procedure widely adopted in human personality measurement in psychology studies. Building on this insight, we propose **ROME** (ROle-play enhanced question-conditioned Mixture-of-Experts), an LLM role-play enhanced, questionnaire-grounded framework for personality detection, with its overall architecture illustrated in Figure 2.

Conceptually, ROME comprises three key stages: *Ask*, *Answer*, and *Detect*. Specifically, in the Ask stage, a role-play LLM agent is instructed to "take on the perspective of the user" and produce answers to each questionnaire item [2] based on the training user's posts and the corresponding training label. These questionnaire-grounded evidence serves as a structured and interpretable bridge between raw linguistic expressions and psychological constructs. Importantly, this evidence generation process is performed offline during training only and is no longer required at inference time. Then, in the

[2] Henceforth, we use the terms "questionnaire item(s)" and "question(s)" interchangeably for convenience and readability throughout this paper.

Answer stage, we enable the model to reproduce these evidence directly from the user posts and the questionnaire item texts, thereby establishing a psychology-informed semantic reasoning chain that simplifies the text-to-personality mapping learning. To account for the heterogeneity across psychometric questionnaire items, we introduce a question-conditioned Mixture-of-Experts (MoE) module, where each expert specializes in a particular type of psycholinguistic cue and the gating network selects the most relevant experts for answering each question. Finally, in the Detect stage, the questionnaire-grounded evidence is aggregated with the user representation learned from posts to perform the final prediction, preserving the expressive capacity of the post encoder while injecting a concise and interpretable decision pathway grounded in standardized psychological knowledge.

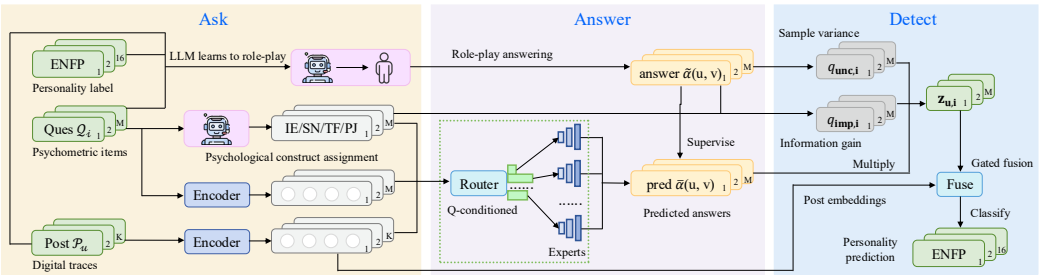


Figure 2: The overall pipeline of ROME.

2.3 Ask: Role-Play-Based Questionnaire Response Generation

To ease the learning of the mapping function $f(\cdot)$, we introduce standardized psychometric questionnaires to decompose each personality label into multiple semantically meaningful questions, thereby providing rich, fine-grained intermediate supervision signals for the model. Ideally, one could present these self-assessment tests directly to each user u and collect their responses to obtain such supervision; however, this procedure is costly, intrusive, and often impractical. Motivated by the recent success of LLMs in simulating human behaviors across diverse contexts [1, 42], we propose an LLM agent to role-play each user with given characteristics (*i.e.*, user posts $\mathcal{P}_u$ and personality label $\mathbf{y}_u$) and produce question-level answers. The posts provide evidence of the user's linguistic patterns and behavioral tendencies, while the label supplies a stable trait orientation; conditioning on both factors enables the LLM agent to more accurately approximate how the user would respond in a self-assessment setting. Here, it is important to note that this generation process is performed offline during training only to augment the training dataset and is not required at inference time.

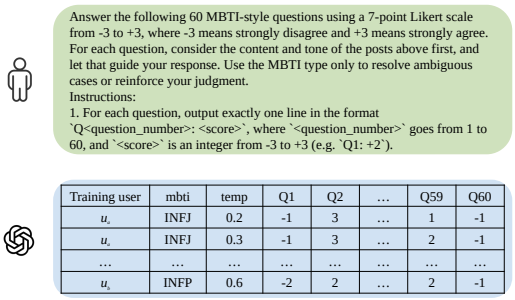


Figure 3: Prompt design for role-play-based questionnaire response generation.

Formally, for each user u with posts $\mathcal{P}_u = \{p_{u,1}, \ldots, p_{u,N_u}\}$ and each questionnaire item $Q_i \in \mathcal{Q}$, we apply the prompt shown in Figure 3 to simulate the user's response. The LLM agent outputs a score on an ordered scale that approximates how user u would answer Q_i. Since LLM generation is inherently stochastic, we draw T samples and average them to reduce sampling variance and obtain a more robust estimate, as follows.

$$\tilde{a}_{u,i} = \frac{1}{T}\sum_{t=1}^{T} a_{u,i}^{(t)}, \quad u \in \mathcal{U},\ Q_i \in \mathcal{Q}, \tag{1}$$

where $a_{u,i}^{(t)}$ denotes the t-th sampled answer for the pair (u, Q_i). These answers bring psychological knowledge explicitly into the modeling framework.

2.4 Answer: Question-Conditioned Mixture-of-Experts Framework

The goal of the Answer stage is to enable the model to predict the question-level answers produced in the Ask stage directly from the user's posts which are accessible during inference. By learning to

reproduce these answers, the model internalizes the psychology-informed semantic reasoning chain, thereby facilitating the text-to-personality mapping process.

A straightforward approach is to train a neural encoder to map a user's posts to each question-level answer. However, different questionnaire items probe distinct psychological constructs (*e.g.*, emotional stability vs. social expressiveness), and the linguistic cues associated with these constructs may also differ substantially across user posts. Therefore, a single shared encoder tends to conflate heterogeneous semantic signals, which leads to entangled representations and suboptimal generalization across questions. To address these limitations, we propose a question-conditioned Mixture-of-Experts (MoE) architecture. Instead of relying on a single shared encoder, the MoE structure allows different experts to specialize in capturing distinct psychological constructs and their associated linguistic cues, while still enabling shared experts to model knowledge that is broadly useful across questions.

Formally, we assume a set of K experts $\mathcal{E} = \{E_1, \ldots, E_K\}$, each parameterized as a neural encoder. Given a user u and a specific question Q_i, the model should first identify the experts that are most informative to generate the corresponding answer. This selection is governed by a routing mechanism. Conceptually, effective routing requires understanding *who* is being assessed (the user's linguistic profile), *what* is being asked (the content of the question), and *which* psychological construct the question is designed to measure. Guided by this intuition, we introduce a lightweight question-conditioned router to dynamically select and aggregate experts as follows.

$$\mathbf{x}_{u,i} = [\mathbf{v}_u;\ \mathbf{q}_i;\ \mathbf{e}_{m_i}],\ \mathbf{g}_{u,i} = \text{softmax}\left(\text{MLP}(\mathbf{x}_{u,i})\right). \tag{2}$$

where $\mathbf{v}_u$ and $\mathbf{q}_i$ denote the text embeddings of the user posts $\mathcal{P}_u$ and the question Q_i respectively, computed by the shared textual encoder $t(\cdot)$. The vector $\mathbf{e}_{m_i}$ is a one-hot indicator specifying the psychological construct that question Q_i is designed to measure (*e.g.*, I/E or S/N personality dimension). In our implementation, we employ an LLM to obtain the construct assignment once, and keep this assignment fixed throughout training and evaluation. The prompt design used by the LLM agent is provided in Appendix B.

Each expert E_k is a small MLP-based expert function denoted as $f_k(\cdot)$. The final answer is then predicted as:

$$\hat{a}_{u,i} = \sum_{k=1}^{K} g_{u,i}^{(k)} \cdot f_k(\mathbf{x}_{u,i}). \tag{3}$$

This component is trained to predict the answer for a given user and psychometric question pair (u, Q_i) using the role-played target $\tilde{a}_{u,i}$ as supervision, and optimized via a regression objective function as follows.

$$\mathcal{L}_q = \frac{1}{|\mathcal{U}| \cdot |\mathcal{Q}|} \sum_{u \in \mathcal{U}} \sum_{Q_i \in \mathcal{Q}} \ell\left(\hat{a}_{u,i},\ \tilde{a}_{u,i}\right), \tag{4}$$

where $\ell(\cdot, \cdot)$ denotes a robust L1 regression loss. During inference, the predicted $\hat{a}_{u,i}$ serves as a strong psychologically grounded evidence for final personality detection.

2.5 Detect: Psychological Evidence-Guided Personality Prediction

2.5.1 Psychological Evidence Modulation

Given the predicted $\hat{a}_{u,i}$ as psychologically grounded evidence, one may simply concatenate them to form an evidence vector for personality prediction. However, two critical challenges remain. First, the "ground truth" answers used to supervise $\hat{a}_{u,i}$ are obtained through LLM role-play generation. Due to the stochastic nature of LLM outputs, the augmented answers may contain noise, leading to varying levels of uncertainty across questions. Second, not all questions in a questionnaire are equally informative for personality inference. Some questions provide strong diagnostic signals compared to others. Therefore, we propose to adaptively weight the contribution of each predicted answer based on both its reliability and its importance when forming the final answer-based representation.

Reliability Score. Specifically, when estimating the reliability, our intuition is that if the responses of a question exhibit high variance across the T sampled role-play generations, the corresponding evidence is more likely to be unreliable. Formally, we can define it as follows.

$$q_{\text{unc},i} = \frac{1}{|\mathcal{U}|} \sum_{u \in \mathcal{U}} \text{Var}\left(\{a_{u,i}^{(t)}\}_{t=1}^{T}\right),\ q_{\text{rel},i} = 1 - \text{Norm}\left(q_{\text{unc},i}\right). \tag{5}$$

where $q_{\mathrm{rel},i}$ denotes the reliability score associated with question Q_i, and $\mathrm{Norm}(\cdot)$ represents the min–max normalization operator.

Importance Score. While for the importance, our intuition is that each question is primarily designed to probe a specific psychological construct (*e.g.*, I/E or S/N personality dimension). Thus, if the LLM-generated answers for a question exhibit stronger separability and induce larger between-class shifts with respect to its intended construct, then that question should be considered more important for personality detection. Formally, for each question Q_i, we first retrieve its one-hot indicator vector $\mathbf{e}_{m_i}$ which is a fixed mapping introduced in the previous section, and denote the corresponding psychological construct by $m \in \mathcal{M}$ (*e.g.*, m = I/E). The importance is then defined as:

$$\mu_{i,m}^{+} = \frac{1}{|\mathcal{U}_m^{+}|} \sum_{u \in \mathcal{U}_m^{+}} \tilde{a}_{u,i},\ \mu_{i,m}^{-} = \frac{1}{|\mathcal{U}_m^{-}|} \sum_{u \in \mathcal{U}_m^{-}} \tilde{a}_{u,i},\ q_{\mathrm{imp},i} = \left|\mu_{i,m}^{+} - \mu_{i,m}^{-}\right|, \tag{6}$$

where $\mathcal{U}_m^{+} = \{u \mid y_u^{(m)} = 1\}$ and $\mathcal{U}_m^{-} = \{u \mid y_u^{(m)} = 0\}$, $q_{\mathrm{imp},i}$ is the importance score associated with question Q_i, which is min–max normalized to the range $[0, 1]$.

Adaptive Weight. For each question Q_i, the final weight w_i is computed under the intuition that more important and more reliable questions should contribute more to the final prediction, as follows.

$$w_i = q_{\mathrm{imp},i} \cdot q_{\mathrm{rel},i}. \tag{7}$$

Both reliability and importance scores are computed using training users only, and the resulting question weights are kept fixed during validation and testing. And the final answer-based representation is obtained by:

$$\mathbf{s}_u = \mathbf{w} \odot \hat{\mathbf{a}}_u. \tag{8}$$

where $\hat{\mathbf{a}}_u$ denotes the vector of predicted question-level answers and $\odot$ denotes element-wise multiplication.

2.5.2 Personality Prediction

Each questionnaire item Q_i is primarily designed to probe a specific psychological construct and contributes only minimally to others. Directly applying question-level signals without regard to their associated constructs may introduce irrelevant evidence and increase noise. To prevent such cross-construct interference, we apply construct-specific binary masks to ensure that each construct only integrates evidence from its corresponding questions. Specifically, for each $m \in \mathcal{M}$ (*e.g.*, m = I/E), we define the mask vector $\boldsymbol{\mu}^{(m)} \in \{0, 1\}^{|\mathcal{Q}|}$ by:

$$\boldsymbol{\mu}_i^{(m)} = \begin{cases} 1, & \arg\max(\mathbf{e}_{m_i}) = m, \\ 0, & \text{otherwise.} \end{cases} \tag{9}$$

And the construct-specific answer representation is then computed by applying the obtained binary mask that preserves only the evidence associated with m as follows.

$$\mathbf{s}_u^{(m)} = \boldsymbol{\mu}^{(m)} \odot \mathbf{s}_u. \tag{10}$$

The post-derived representation $\mathbf{v}_u$ captures broad linguistic patterns across a user's posts and is highly expressive, but it is not explicitly aligned with psychological constructs. In contrast, the construct-specific evidence $\mathbf{s}_u^{(m)}$ is questionnaire-grounded, providing a semantic and interpretable reasoning pathway to psychological constructs. However, its quality can vary due to differences in diagnostic strength across questionnaire items and the uncertainty inherent in role-play supervision. Therefore, the two sources offer complementary advantages and with varying reliability. Their contributions should be adaptively balanced during personality prediction. Specifically, we introduce a lightweight gating network that learns to dynamically weight the contributions of $\mathbf{v}_u$ and $\mathbf{s}_u^{(m)}$ as:

$$\boldsymbol{\gamma}_u^{(m)} = \sigma\left(\mathrm{MLP}_{\mathrm{gate}}^{(m)}\left(\mathbf{v}_u \,\|\, \mathbf{s}_u^{(m)}\right)\right),\ \mathbf{z}_u^{(m)} = \boldsymbol{\gamma}_u^{(m)} \odot \mathbf{v}_u + \left(\mathbf{1} - \boldsymbol{\gamma}_u^{(m)}\right) \odot \mathbf{s}_u^{(m)}. \tag{11}$$

Subsequently, the personality prediction for a given $m \in \mathcal{M}$ is produced by a lightweight construct-specific classifier as follows.

$$\hat{y}_u^{(m)} = \sigma\left(\mathrm{MLP}_{\mathrm{cls}}^{(m)}\left(\mathbf{z}_u^{(m)}\right)\right), \tag{12}$$

Table 1: Performance comparison on the KAGGLE and PANDORA datasets.

Methods	KAGGLE					PANDORA				
	I/E	S/N	T/F	P/J	Avg	I/E	S/N	T/F	P/J	Avg
XGBoost	56.67	52.85	75.42	65.94	62.72	45.99	48.93	63.51	55.55	53.50
BiLSTM	57.82	57.87	69.97	57.01	60.67	48.01	52.01	63.48	56.21	54.93
BERTconcat	58.33	53.88	69.36	60.88	60.61	54.22	49.15	58.31	53.14	53.71
BERTmean	64.05	57.82	77.06	65.25	66.04	56.60	48.71	64.70	56.07	56.52
AttRCNN	59.74	64.08	78.77	66.44	67.25	48.55	56.19	64.39	57.26	56.60
AttnSeq	65.43	62.15	78.05	63.92	67.39	56.98	54.78	60.95	54.81	56.88
Transformer-MD	66.08	69.10	79.19	67.50	70.47	55.26	58.77	69.26	60.90	61.05
TrigNet	69.54	67.17	79.06	67.69	70.86	56.69	55.57	66.38	57.27	58.98
D-DGCN	68.41	65.66	79.56	67.22	70.21	61.55	55.46	1.07	59.96	62.01
D-DGCN+ℓ_0	69.52	67.19	80.53	68.16	71.35	59.98	55.52	70.53	59.56	61.40
PQ-Net	68.94	67.65	79.12	69.57	71.32	57.07	55.26	65.64	58.74	59.18
MvP	67.68	69.89	80.99	68.32	71.72	60.08	56.99	69.12	61.19	61.85
TAE	70.90	66.21	81.17	70.20	72.12	62.57	61.01	69.28	59.34	63.05
ETM	68.97	71.21	86.19	**84.78**	77.79	68.57	64.91	66.07	63.53	65.77
LL4G	79.05	78.27	86.03	76.91	80.06	66.24	67.55	71.30	66.91	68.00
EmoPerso	81.12	78.23	84.45	79.21	80.75	68.45	67.98	70.10	69.77	69.08
ROME (BERT)	**92.24±0.47**	**95.90±0.17**	**97.45±0.12**	81.32±0.28	**91.73±0.11**	**74.93±0.24**	**95.45±0.04**	**82.21±0.09**	**78.22±0.07**	**82.70±0.11**
Improvement	13.71%	22.52%	13.06%	-4.08%	13.60%	9.28%	40.41%	15.30%	12.11%	19.72%

which is optimized using a binary classification loss function:

$$\mathcal{L}_{\text{cls}} = \frac{1}{|\mathcal{U}| \cdot |\mathcal{M}|} \sum_{u \in \mathcal{U}} \sum_{m \in \mathcal{M}} \text{BCE}\left(\hat{y}_u^{(m)},\, y_u^{(m)}\right). \tag{13}$$

Together with the answer prediction loss defined in Equation (4), the model is finally trained in a multi-task learning framework with the following joint objective function:

$$\mathcal{L} = \lambda_q \, \mathcal{L}_q + \lambda_{cls} \, \mathcal{L}_{cls}, \tag{14}$$

where λ_q and λ_{cls} are hyperparameters that balance the contributions of the answer prediction and personality classification tasks, respectively.

Remark. During training, we first pre-train the MoE module with the answer prediction objective (*i.e.*, Equation (4)) to stabilize the question-aligned reasoning space, and then fine-tune the entire framework end-to-end using the joint objective in Equation (14). At inference time, the user posts are encoded once, and the question-conditioned MoE module produces psychologically grounded evidence $\{\hat{a}_{u,i}\}$, which is then fused with the encoded post representations to generate the final personality prediction.

3 Experiment

Experiment Overview. We first report the overall comparison results with existing methods in Section 3.1, followed by component ablations in Section 3.2. We then compare ROME with representative inference-time LLM methods in Section 3.3 and examine its data efficiency in Section 3.4. Additional experiments on backbone LLMs and text encoders are provided in Appendix C. Parameter analyses, including questionnaire length and the number of experts, are presented in Appendix D. A three-part qualitative case study covering question–post evidence chains, learned question weights, and expert activation patterns is included in Appendix E.

Dataset. In line with prior work [44, 47, 11], we conduct our evaluation on two widely used personality datasets: KAGGLE and PANDORA [9]. Given pronounced class imbalance, we report Macro-F1 per dimension and the overall score as the average of the four Macro-F1s. Following [47], we shuffle users and partition data into train/validation/test with a 60/20/20 split. Dataset statistics are summarized in Table 5 in Appendix B.

Baselines. We compare our approach against **16** strong baseline methods, including traditional and basic neural models (XGBoost [5], BiLSTM [36], BERTconcat [13], and BERTmean [16]), deep and structure-enhanced models (AttRCNN [43], AttnSeq [23], Transformer-MD [44], TrigNet [46], PQ-Net [45], D-DGCN and D-DGCN+ℓ_0 [47], and MvP [51]), and recent LLM-assisted methods (TAE [11], ETM [2], LL4G [34], and EmoPerso [33]). For recent LLM-assisted baselines that support different backbone LLMs (*e.g.*, LL4G and EmoPerso), we evaluate their `GPT-4o`-based variants in our experiments to ensure a fair comparison. The implementation details are provided in Appendix B.

3.1 Overall Results

The main results on the two datasets are provided in Table 1. We observe that traditional feature-based models and standard PLM baselines perform poorly on both datasets. For example, BERTmean achieves only 66.04 and 56.52 average Macro-F1 on Kaggle and Pandora, respectively. In contrast, recent LLM-enhanced baselines deliver substantially stronger performance. In particular, EmoPerso achieves the strongest baseline results, reaching 80.75 and 69.08 average Macro-F1 on the two datasets. ROME further improves over these strong LLM-augmented baselines, achieving 91.73 and 82.70 Macro-F1 on Kaggle and Pandora, corresponding to relative improvements of 13.60% and 19.72% over EmoPerso.

We further observe that ROME consistently and significantly outperforms ETM even though it relies only on a small-scale text encoder (`BERT-base-uncased`), whereas ETM exploits long-text embeddings from a much larger `Meta-Llama-3-8B-Instruct` model (about $70\times$ larger than `BERT-base-uncased`). This advantage mainly stems from ROME's questionnaire-grounded architecture, thereby providing substantially richer intermediate supervision than coarse-grained personality labels alone. Additional experiments in Appendix C demonstrate that substituting BERT with LLaMA will further increase the performance from 91.73 to 93.61 on the KAGGLE dataset.

3.2 Ablation Studies

To assess the contribution of each component in our ROME framework, we conduct an ablation study on the KAGGLE dataset, as shown in Table 2 and Table 3.

Benefits of adaptive weighting. We first evaluate the impact of the question-level weighting mechanism. Removing the reliability- and importance-based weights ($\text{ROME}_{\text{w/o q-weighting}}$) reduces the average Macro-F1 from 91.73 to 88.47. This degradation reflects the heterogeneity across questionnaire items: different items vary in their uncertainty in LLM-simulated answers and in their discriminative power for distinguishing personality classes.

Benefits of gated fusion. We next examine the fusion mechanism between post-derived and questionnaire-grounded representations. Replacing the learnable gated fusion with a simple unweighted average ($\text{ROME}_{\text{w/o gated fusion}}$) yields a similar decline in performance. This result indicates that the relative contributions of posts and questionnaire evidence should be adaptively determined.

Benefits of psychological knowledge. We further ablate each information source in isolation to examine their individual contributions and complementarity. Specifically, using only questionnaire-grounded evidence while discarding post representations ($\text{ROME}_{\text{w/o posts}}$) reduces the average Macro-F1 to 75.12. Conversely, using only post representations ($\text{ROME}_{\text{w/o evidence}}$) leads to an even larger drop to 65.77. This pattern aligns with our intuition: post-derived features capture broad and expressive linguistic cues, whereas questionnaire-grounded evidence provides semantically aligned and psychologically meaningful signals. ROME performs best when both sources are jointly leveraged.

Table 2: Results of the ablation study on the KAGGLE dataset.

Methods	I/E	S/N	T/F	P/J	Avg
$\text{ROME}_{\text{w/o q-weighting}}$	90.13	94.82	96.36	72.55	88.47
$\text{ROME}_{\text{w/o gated fusion}}$	89.03	94.45	96.36	74.16	88.50
$\text{ROME}_{\text{w/o posts}}$	86.56	71.21	94.18	48.51	75.12
$\text{ROME}_{\text{w/o evidence}}$	42.33	92.35	75.11	53.27	65.77
$\text{ROME}_{\text{label-only Ask}}$	47.02	91.53	79.02	56.09	68.42
$\text{ROME}_{\text{posts-only Ask}}$	83.69	94.02	92.42	71.51	85.41
$\text{ROME}_{\text{w/o pretrain}}$	47.58	92.19	80.14	56.93	69.21
ROME	**92.24**	**95.90**	**97.45**	**81.32**	**91.73**

Benefits of combining posts and labels in role-play. We also examine the contributions of posts and labels to Ask-stage response generation while keeping all other components fixed. Removing personality labels from the role-playing LLM's inputs ($\text{ROME}_{\text{posts-only Ask}}$) reduces the average Macro-F1 from 91.73 to 85.41. Conversely, retaining only personality labels ($\text{ROME}_{\text{label-only Ask}}$) leads to a larger drop to 68.42, only 2.65 points above removing the evidence pathway entirely. These results indicate that label expansion alone provides limited auxiliary supervision, while grounding role-play responses in user posts is important for effective question-level supervision. Further evidence is provided in Appendix E.4.

Benefits of question-conditioned expert routing. We introduce two orthogonal design choices: (i) whether question information enters the predictor, and (ii) what the router conditions on.

(1) Question information improves predictor. Predicting all 60 items from user representation alone ($\text{ROME}_{\text{user-only MLP}}$) forces a single mapping to serve heterogeneous items and recovers only 3.05 points over removing the evidence pathway entirely. Conditioning the predictor on the item text and its construct indicator ($\text{ROME}_{\text{q-cond. MLP}}$) improves the average Macro-F1 from 68.82 to 78.14, confirming that different items require different mappings from user language to responses.

Table 3: Ablation study of the Answer module on the KAGGLE dataset.

Methods	I/E	S/N	T/F	P/J	Avg
$\text{ROME}_{\text{user-only MLP}}$	45.15	91.96	80.27	57.90	68.82
$\text{ROME}_{\text{q-cond. MLP}}$	64.14	92.83	88.82	66.76	78.14
$\text{ROME}_{\text{uniform MoE}}$	82.58	93.16	93.30	71.16	85.05
$\text{ROME}_{\text{user-cond. MoE}}$	<u>91.46</u>	93.46	95.29	<u>76.98</u>	<u>89.30</u>
ROME	**92.24**	**95.90**	**97.45**	**81.32**	**91.73**

(2) Question information improves router. We note that the uniform MoE, user-conditioned MoE, and full ROME variants have identical MoE structure, differing only in the router design. Replacing uniform mixing ($\text{ROME}_{\text{uniform MoE}}$) with user-conditioned routing ($\text{ROME}_{\text{user-cond. MoE}}$) improves the average Macro-F1 from 85.05 to 89.30, and additionally conditioning the router on question information (ROME) further improves it from 89.30 to 91.73. This isolates the contribution claimed by our method: user-dependent routing alone is insufficient, and the router needs to know which construct is being probed in order to select experts specialized for the corresponding psycholinguistic cues and associated linguistic cues.

Benefits of model pretraining. Finally, skipping the pretraining stage of the answer module ($\text{ROME}_{\text{w/o pretrain}}$) leads to a substantial performance reduction to 69.21. This indicates that directly optimizing the full model without first stabilizing the question-level prediction task makes it difficult for ROME to fully leverage item-level supervision signals.

3.3 Inference-Time LLM Methods

To better understand different ways of using LLMs for personality detection, we compare three representative settings on the KAGGLE test set, as summarized in Table 4: direct prompting with ChatGPT, recent inference-time LLM reasoning methods, and our ROME framework with different Ask-stage LLM backbones. Following [11], we first consider two prompting configurations for `gpt-3.5-turbo` and `gpt-4o-2024-08-06` under zero-shot and 3-shot settings. We further include recent online inference methods, including `PsyCoT-3.5`, `HIPPD-4o`, and `EERPD-3.5`, all of which rely on inference-time LLM interactions and substantially heavier generation pipelines, making them fundamentally different from our setting.

Compared with direct prompting, online reasoning-based methods improve performance by introducing more structured inference procedures, such as multi-turn questionnaire prompting or explicit reasoning over psychologically relevant cues. Nevertheless, both `ROME-3.5` and `ROME-4o` outperform all these alternatives in overall performance. In particular, `ROME-3.5` already surpasses the strongest online reasoning method `EERPD-3.5`, while `ROME-4o` achieves the best average Macro-F1 of 91.73. This result suggests that the advantage of ROME does not merely come from using a stronger backbone LLM, but from its questionnaire-grounded Ask–Answer–Detect framework. Instead of directly relying on raw inference-time LLM outputs for each test instance, ROME transforms LLM-based reasoning into offline supervision by generating question-level answers during training. Importantly, because such generated answers can be noisy and vary in reliability across questions, ROME further calibrates this supervision by considering both the stability of generated responses and the diagnostic importance of questionnaire items.

Table 4: Inference-Time LLM performance on the KAGGLE dataset.

Methods	I/E	S/N	T/F	P/J	Avg
ChatGPT-3.5	65.86	51.69	78.60	63.93	65.02
$\text{ChatGPT-3.5}_{\text{3-shot}}$	70.61	58.35	76.58	65.43	67.74
ChatGPT-4o	81.70	78.15	85.13	74.67	79.91
$\text{ChatGPT-4o}_{\text{3-shot}}$	80.55	76.53	84.87	78.13	80.02
PsyCoT-3.5	66.56	61.70	74.80	57.83	65.22
HIPPD-4o	73.61	77.25	84.98	79.33	78.79
EERPD-3.5	86.63	90.59	89.15	**82.12**	87.12
ROME-3.5	<u>88.32</u>	<u>94.91</u>	<u>95.86</u>	73.76	<u>88.21</u>
ROME-4o	**92.24**	**95.90**	**97.45**	<u>81.32</u>	**91.73**

More importantly, unlike online inference-time methods such as `PsyCoT`, `HIPPD`, and `EERPD`, ROME uses the LLM only in the offline Ask stage during model training, while inference is performed

entirely by a compact supervised model. This design yields not only better overall accuracy, but also substantially lower latency and monetary cost, making ROME more practical for deployment.

3.4 Data Efficiency under Limited Supervision

To examine how well ROME performs when labeled data are scarce, a situation frequently encountered in practice because personality annotation is costly and often subject to privacy concerns, we simulate limited-supervision settings by down-sampling the KAGGLE training set while keeping the validation and test sets fixed. Concretely, we train the same ROME configuration using only 40%, 60%, 80%, or 100% of the training data and report the resulting average Macro-F1 in Figure 4. As the available training data increases, performance improves smoothly from 82.87 (40%) to 87.61 (60%), 89.16 (80%), and 91.73 with the full dataset, indicating that the model behaves stably across data scales and can effectively exploit additional supervision when it is available.

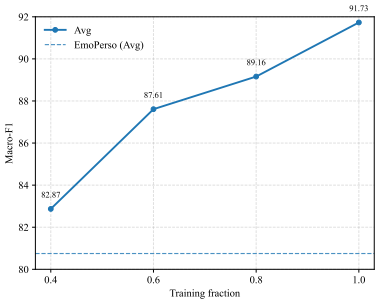


Figure 4: Effect of training data size on the KAGGLE dataset.

More importantly, ROME remains highly competitive even in the most data-poor regime. With only 40% of the training data, our model (82.87) already surpasses the best baseline method EmoPerso, whose average Macro-F1 under full-data training is 80.75 (Table 1). These results highlight the data efficiency of the Ask–Answer–Detect framework: by leveraging standardized questionnaires and question-level supervision, ROME benefits from a much denser and more structured training than traditional "posts $\rightarrow$ user vector $\rightarrow$ labels" models, and thus maintains strong performance even when labeled users are very limited.

3.5 Additional Experimental Analyses

To complement the main experiments, we conduct further analyses in the Appendix. We first examine the effects of backbone LLMs and text encoders (Appendix C), as well as questionnaire length and the number of MoE experts (Appendix D). We further investigate the interpretability of ROME through question–post evidence chains, item-removal sensitivity, and expert activation patterns (Appendix E). In addition, we analyze the diversity of role-played questionnaire responses to examine whether they preserve user-specific variation beyond personality labels (Appendix E.4). Finally, we provide an additional comparison with EERPD on PANDORA and evaluate the applicability of ROME to Big Five personality traits on the Essays dataset (Appendix F). These experiments provide supplementary evidence on the robustness, interpretability, and broader applicability of ROME.

4 Conclusion

In this paper, we propose a role-play–enhanced, question-conditioned Mixture-of-Experts framework for personality detection that replaces the prevailing "posts $\rightarrow$ user vector $\rightarrow$ labels" paradigm with an Ask–Answer–Detect pipeline. By leveraging an LLM-based role-play agent to generate question-level answers and treating these answers as questionnaire-grounded evidence, our framework injects structured psychological knowledge into the modeling process and provides an explicit question-level reasoning chain from raw text to personality labels. Extensive experiments on two real-world datasets demonstrate substantial and consistent gains, with relative Macro-F1 improvements of up to 13.60% and 19.72% over state-of-the-art baselines.

Acknowledgments and Disclosure of Funding

This work was supported by the Guangzhou-HKUST(GZ) Joint Funding Program (No. 2024A03J0630). The authors declare no competing interests.

References

[1] Zahra Abbasiantaeb, Yifei Yuan, Evangelos Kanoulas, and Mohammad Aliannejadi. Let the LLMs talk: Simulating human-to-human conversational QA via zero-shot LLM-to-LLM interactions. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining (WSDM '24)*, pages 8–17, New York, NY, USA, 2024. Association for Computing Machinery. doi: 10.1145/3616855.3635856. URL https://doi.org/10.1145/3616855.3635856.

[2] Weihong Bi, Feifei Kou, Lei Shi, Yawen Li, Haisheng Li, Jinpeng Chen, and Mingying Xu. Leveraging the dual capabilities of LLM: LLM-enhanced text mapping model for personality detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 23487–23495, 2025. doi: 10.1609/aaai.v39i22.34517. URL https://ojs.aaai.org/index.php/AAAI/article/view/34517.

[3] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html.

[4] Guanming Chen, Lingzhi Shen, Xiaohao Cai, Imran Razzak, and Shoaib Jameel. HIPPD: Brain-inspired hierarchical information processing for personality detection. In *Proceedings of the 17th Asian Conference on Machine Learning*, volume 304 of *Proceedings of Machine Learning Research*, pages 910–925. PMLR, 2025. URL https://proceedings.mlr.press/v304/chen26c.html.

[5] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794, New York, NY, USA, 2016. Association for Computing Machinery. doi: 10.1145/2939672.2939785. URL https://doi.org/10.1145/2939672.2939785.

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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*, pages 4171–4186, 2019. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423/.

[7] Nan Du, Yanping Huang, Andrew M. Dai, Simon Tong, Dmitry Lepikhin, Yuanzhong Xu, Maxim Krikun, Yanqi Zhou, Adams Wei Yu, Orhan Firat, Barret Zoph, Liam Fedus, Maarten P. Bosma, Zongwei Zhou, Tao Wang, Emma Wang, Kellie Webster, Marie Pellat, Kevin Robinson, Kathleen Meier-Hellstern, Toju Duke, Lucas Dixon, Kun Zhang, Quoc Le, Yonghui Wu, Zhifeng Chen, and Claire Cui. GLaM: Efficient scaling of language models with mixture-of-experts. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 5547–5569. PMLR, 2022. URL https://proceedings.mlr.press/v162/du22c.html.

[8] William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23 (120):1–39, 2022. URL https://www.jmlr.org/papers/v23/21-0998.html.

[9] Matej Gjurković, Vanja Mladen Karan, Iva Vukojević, Mihaela Bošnjak, and Jan Snajder. PANDORA talks: Personality and demographics on Reddit. In *Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media*, pages 138–152, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.socialnlp-1.12. URL https://aclanthology.org/2021.socialnlp-1.12/.

[10] Jeffrey Gottfried. Americans' social media use. Technical report, Pew Research Center, January 2024. URL https://www.pewresearch.org/internet/2024/01/31/americans-social-media-use/. Accessed 2025-10-02.

[11] Linmei Hu, Hongyu He, Duokang Wang, Ziwang Zhao, Yingxia Shao, and Liqiang Nie. 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, 2024. doi: 10.1609/aaai.v38i16.29782. URL https://ojs.aaai.org/index.php/AAAI/article/view/29782.

[12] Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Guillaume Lample, et al. Mixtral of experts. *arXiv preprint arXiv:2401.04088*, 2024. doi: 10.48550/arXiv.2401.04088. URL https://arxiv.org/abs/2401.04088.

[13] Hang Jiang, Xianzhe Zhang, and Jinho D. Choi. 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, 2020. doi: 10.1609/aaai.v34i10.7182. URL https://ojs.aaai.org/index.php/AAAI/article/view/7182.

[14] Oliver P. John, Eileen M. Donahue, and Robert L. Kentle. The big five inventory—versions 4a and 54. Technical report, Institute of Personality and Social Research, University of California, Berkeley, Berkeley, CA, 1991.

[15] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, et al. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020. URL https://arxiv.org/abs/2001.08361.

[16] Sedrick Scott Keh and I-Tsun Cheng. Myers-briggs personality classification and personality-specific language generation using pre-trained language models. *arXiv preprint arXiv:1907.06333*, 2019. URL https://arxiv.org/abs/1907.06333.

[17] Otto F. Kernberg. What is personality? *Journal of Personality Disorders*, 30(2):145–156, 2016. doi: 10.1521/pedi.2106.30.2.145. URL https://doi.org/10.1521/pedi.2106.30.2.145.

[18] Dmitry Lepikhin, HyoukJoong Lee, Yuanzhong Xu, Dehao Chen, Orhan Firat, Yanping Huang, Maxim Krikun, Noam Shazeer, and Zhifeng Chen. GShard: Scaling giant models with conditional computation and automatic sharding. In *International Conference on Learning Representations*, 2021. URL https://mlanthology.org/iclr/2021/lepikhin2021iclr-gshard/.

[19] Guohao Li, Hasan Abed Al Kader Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. CAMEL: Communicative agents for "mind" exploration of large language model society. *arXiv*, 2023. doi: 10.48550/arXiv.2303.17760. URL https://arxiv.org/abs/2303.17760.

[20] Zheng Li, Sujian Li, Dawei Zhu, Qilong Ma, and Weimin Xiong. EERPD: Leveraging emotion and emotion regulation for improving personality detection. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 7721–7734, Abu Dhabi, UAE, January 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.coling-main.516/.

[21] Aixin Liu, Bei Feng, Bin Wang, et al. Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model. *arXiv preprint arXiv:2405.04434*, 2024. URL https://arxiv.org/abs/2405.04434.

[22] Ryan Louie, Ananjan Nandi, William Fang, Cheng Chang, Emma Brunskill, and Diyi Yang. Roleplay-doh: Enabling domain-experts to create LLM-simulated patients via eliciting and adhering to principles. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 10570–10603, Miami, Florida, USA, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.591. URL https://aclanthology.org/2024.emnlp-main.591/.

[23] Veronica Lynn, Niranjan Balasubramanian, and H. Andrew Schwartz. 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, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.472. URL https://aclanthology.org/2020.acl-main.472/.

[24] Jiaqi Ma, Zhe Zhao, Xinyang Yi, Jilin Chen, Lichan Hong, and Ed H. Chi. Modeling task relationships in multi-task learning with multi-gate mixture-of-experts. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1930–1939, 2018. doi: 10.1145/3219819.3220007. URL https://dl.acm.org/doi/10.1145/3219819.3220007.

[25] François Mairesse, Marilyn A. Walker, Matthias R. Mehl, and Roger K. Moore. Using linguistic cues for the automatic recognition of personality in conversation and text. *Journal of Artificial Intelligence Research*, 30:457–500, 2007. doi: 10.1613/jair.2349. URL https://www.jair.org/index.php/jair/article/view/10520.

[26] Isabel Briggs Myers, Mary H. McCaulley, Naomi L. Quenk, and Allen L. Hammer. *MBTI Manual: A Guide to the Development and Use of the Myers–Briggs Type Indicator*. Consulting Psychologists Press, Palo Alto, CA, 3rd edition, 1998. ISBN 9780891061304. URL https://isbnsearch.org/isbn/9780891061304.

[27] Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST '23)*, 2023. doi: 10.1145/3586183.3606763. URL https://dl.acm.org/doi/10.1145/3586183.3606763.

[28] James W. Pennebaker and Laura A. King. Linguistic styles: Language use as an individual difference. *Journal of Personality and Social Psychology*, 77(6):1296–1312, 1999. doi: 10.1037/0022-3514.77.6.1296. URL https://doi.org/10.1037/0022-3514.77.6.1296.

[29] James W. Pennebaker, Roger J. Booth, Ryan L. Boyd, and Martha E. Francis. *Linguistic Inquiry and Word Count: LIWC2015*. Austin, TX, 2015. URL https://liwc.app/static/documents/LIWC2015%20Manual%20-%20Operation.pdf.

[30] Zihan Qiu, Zeyu Huang, Shuang Cheng, Yizhi Zhou, Zili Wang, Ivan Titov, and Jie Fu. Layerwise recurrent router for mixture-of-experts. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=eWNEqdH0vk.

[31] H. Andrew Schwartz, Johannes C. Eichstaedt, Margaret L. Kern, Lukasz Dziurzynski, Stephanie M. Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stillwell, Martin E. P. Seligman, and Lyle H. Ungar. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PLOS ONE*, 8(9):e73791, 2013. doi: 10.1371/journal.pone.0073791. URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0073791.

[32] Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *International Conference on Learning Representations*, 2017. URL https://arxiv.org/abs/1701.06538.

[33] Lingzhi Shen, Xiaohao Cai, Yunfei Long, Imran Razzak, Guanming Chen, and Shoaib Jameel. EmoPerso: Enhancing personality detection with self-supervised emotion-aware modelling. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, pages 2577–2587. Association for Computing Machinery, 2025. doi: 10.1145/3746252.3761247. URL https://doi.org/10.1145/3746252.3761247.

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

[35] Sanja Štajner and Seren Yenikent. A survey of automatic personality detection from texts. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6284–6295, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.553. URL https://aclanthology.org/2020.coling-main.553/.

[36] Tommy Tandera, Hendro, Derwin Suhartono, Rini Wongso, and Yen Lina Prasetio. Personality prediction system from Facebook users. *Procedia Computer Science*, 116:604–611, 2017. doi: 10.1016/j.procs.2017.10.016. URL https://doi.org/10.1016/j.procs.2017.10.016.

[37] Yu-Min Tseng, Yu-Chao Huang, Teng-Yun Hsiao, Wei-Lin Chen, Chao-Wei Huang, Yu Meng, and Yun-Nung Chen. Two tales of persona in LLMs: A survey of role-playing and personalization. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 16612–16631, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.969. URL https://aclanthology.org/2024.findings-emnlp.969/.

[38] Lei Wang, Jianxun Lian, Yi Huang, Yanqi Dai, Haoxuan Li, Xu Chen, Xing Xie, and Ji-Rong Wen. Characterbox: Evaluating the role-playing capabilities of LLMs in text-based virtual worlds. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6372–6391, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.naacl-long.323. URL https://aclanthology.org/2025.naacl-long.323/.

[39] Noah Wang, Z.Y. Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhan Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Jian Yang, Man Zhang, Zhaoxiang Zhang, Wanli Ouyang, Ke Xu, Wenhao Huang, Jie Fu, and Junran Peng. RoleLLM: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14743–14777, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.878. URL https://aclanthology.org/2024.findings-acl.878/.

[40] Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=gEZrGCozdqR.

[41] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, 2022. doi: 10.52202/068431-1800. URL https://proceedings.neurips.cc/paper/2022/hash/9d5609613524ecf4f15af0f7b31abca4-Abstract.html.

[42] Chengxing Xie, Canyu Chen, Feiran Jia, Ziyu Ye, Shiyang Lai, Kai Shu, Jindong Gu, Adel Bibi, Ziniu Hu, David Jurgens, et al. Can Large Language Model Agents simulate Human Trust Behavior? In *Advances in Neural Information Processing Systems 37*, pages 15674–15729. Neural Information Processing Systems Foundation, 2024. doi: 10.52202/079017-0501.

[43] Di Xue, Lifa Wu, Zheng Hong, Shize Guo, Liang Gao, Zhiyong Wu, Xiaofeng Zhong, and Jianshan Sun. Deep learning-based personality recognition from text posts of online social networks. *Applied Intelligence*, 48(12):4232–4246, 2018. doi: 10.1007/s10489-018-1212-4. URL https://doi.org/10.1007/s10489-018-1212-4.

[44] Feifan Yang, Xiaojun Quan, Yunyi Yang, and Jianxing Yu. Multi-document transformer for personality detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14221–14229, 2021. doi: 10.1609/aaai.v35i16.17673. URL https://ojs.aaai.org/index.php/AAAI/article/view/17673.

[45] Feifan Yang, Tao Yang, Xiaojun Quan, and Qinliang Su. Learning to answer psychological questionnaire for personality detection. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1131–1142, Punta Cana, Dominican Republic, November

2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-emnlp.98. URL https://aclanthology.org/2021.findings-emnlp.98/.

[46] Tao Yang, Feifan Yang, Haolan Ouyang, and Xiaojun Quan. Psycholinguistic tripartite graph network for personality detection. 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 4229–4239, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.326. URL https://aclanthology.org/2021.acl-long.326/.

[47] Tao Yang, Jinghao Deng, Xiaojun Quan, and Qifan Wang. 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, 2023. doi: 10.1609/aaai.v37i11.26627. URL https://ojs.aaai.org/index.php/AAAI/article/view/26627.

[48] Tao Yang, Tianyuan Shi, Fanqi Wan, Xiaojun Quan, Qifan Wang, Bingzhe Wu, and Jiaxiang Wu. PsyCoT: Psychological questionnaire as powerful chain-of-thought for personality detection. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3305–3320, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.216. URL https://aclanthology.org/2023.findings-emnlp.216/.

[49] Wu Youyou, Michal Kosinski, and David Stillwell. Computer-based personality judgments are more accurate than those made by humans. *Proceedings of the National Academy of Sciences of the United States of America*, 112(4):1036–1040, 2015. doi: 10.1073/pnas.1418680112. URL https://www.pnas.org/doi/10.1073/pnas.1418680112.

[50] Hao Zhao, Zihan Qiu, Huijia Wu, Zili Wang, Zhaofeng He, and Jie Fu. HyperMoE: Towards better mixture of experts via transferring among experts. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10605–10618, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.571. URL https://aclanthology.org/2024.acl-long.571/.

[51] Haohao Zhu, Xiaokun Zhang, Junyu Lu, Liang Yang, and Hongfei Lin. Integrating multi-view analysis: Multi-view mixture-of-expert for textual personality detection. In *Natural Language Processing and Chinese Computing: NLPCC 2024, Proceedings, Part IV*, volume 15362 of *Lecture Notes in Computer Science*, pages 359–371. Springer, 2024. doi: 10.1007/978-981-97-9440-9_28. URL https://doi.org/10.1007/978-981-97-9440-9_28.

[52] Yangfu Zhu, Linmei Hu, Xinkai Ge, Wanrong Peng, and Bin Wu. Contrastive graph transformer network for personality detection. In *Proceedings of the 31st International Joint Conference on Artificial Intelligence (IJCAI 2022)*, pages 4559–4565, 2022. doi: 10.24963/ijcai.2022/633. URL https://www.ijcai.org/proceedings/2022/633.

A Related Work

A.1 Personality Detection

Early evidence shows that digital traces and language use carry robust signals of human personality [31, 49]. Classical approaches therefore relied on handcrafted or lexicon-based features (*e.g.*, LIWC) to engineer user representations for personality classification [25, 29]. With the advent of deep learning, the community has shifted toward neural encoders that aggregate multiple posts into a unified user vector automatically, typically built upon pretrained language models such as BERT [6], with strategies ranging from simple concatenation to hierarchical pooling [13, 16]. To better capture dependencies across posts, subsequent studies further introduce cross-document modeling and graph-based architectures (*e.g.*, multi-document transformers and dynamic graph convolution) to enhance user-level semantic representations [44, 47, 52]. Some recent extensions further enhance user representations by incorporating psychological guidance or multi-view modeling. For example, [45] leverages psychological domain knowledge to extract personality-relevant cues from user posts, while [51] models users from multiple perspectives using consistency regularization.

More recently, LLM-augmented pipelines have advanced the frontier by leveraging LLMs' strong generation and embedding capacities. For instance, TAE [11] leverages LLMs to generate multi-perspective analyses of posts and labels within contrastive learning frameworks; ETM [2] strengthens user vectors with LLM-derived long-text embeddings and label-aware alignment; LL4G [34] incorporates LLM-derived semantic embeddings to facilitate graph construction for better graph-based user representation learning; and EmoPerso [33] introduces LLM-based generative mechanisms for conditioned paraphrasing and contextual feature completion to enrich emotion-aware user representations. Despite these advances, the prevailing modeling paradigm remains "posts $\rightarrow$ user vector $\rightarrow$ labels", which learns text-to-personality mappings only from coarse-grained labels, offering limited interpretability and intermediate supervision.

In parallel, some recent studies further explore inference-time LLM reasoning for personality detection, including multi-turn questionnaire prompting [48], retrieval-augmented few-shot inference [20], and online hierarchical LLM-based reasoning [4]. However, these approaches require online interaction with LLMs and incur substantial higher inference costs, making them fundamentally different from our setting.

A.2 Role-play Prompting

Role-play has emerged as a popular prompt-based paradigm for conditioning LLMs on personas, goals, and contextual constraints across both simulation and evaluation settings, enabling coherent agent behaviors and goal-directed reasoning in open-ended environments [19, 27]. Recent studies have systematized the use of role-play in LLMs by introducing benchmarks to quantify role adherence [39] and domain-specific frameworks that co-design standardized, faithful interactions such as clinician-guided virtual patients and character-grounded agents in text-based environments [22, 38]. These developments, further synthesized in a recent survey [37], collectively underscore the growing interest in aligning LLM behavior with contextual roles. In this paper, we do not employ role-play for open-ended dialogue generation. Instead, we elicit question-level answers to validated psychometric questionnaires from users' posts, providing psychologically grounded intermediate supervision for personality detection. To the best of our knowledge, this is the first work to harness LLM's role-play capacity to construct an explicit questionnaire-grounded evidence space for this task, tightly integrated with our question-conditioned Mixture-of-Experts architecture that learns to answer questionnaire items before final personality classification.

A.3 Mixture-of-Experts (MoE)

Mixture-of-Experts (MoE) realizes *conditional computation* by routing each input to a small subset of experts, thereby achieving scalable model capacity through sparse activation [32, 18, 8, 7]. Beyond sheer scale, task- and signal-aware variants encourage functional specialization. For instance, multi-gate MoE architectures have been widely applied in multi-task learning scenarios [24]. More recent systems further develop different routing mechanisms that address various aspects, including improving routing stability, load balance, and cross-task transferability [12, 21, 50, 30]. These advances collectively show that what conditions the router matters: conditioning on the appropriate

Table 5: Statistics of the two datasets.

Dataset	Types	Train	Validation	Test
Kaggle	I/E	4032/1173	1330/405	1314/421
	S/N	724/4481	230/1505	243/1492
	T/F	2388/2817	802/933	791/944
	P/J	3160/2045	1007/728	1074/661
Pandora	I/E	4314/1126	1425/388	1403/411
	S/N	621/4819	202/1611	205/1609
	T/F	3527/1913	1160/653	1164/650
	P/J	3211/2229	1064/749	1035/779

structure enables experts to capture distinct facets of the problem. Unlike prior work that conditions primarily on token- or task-level signals to improve efficiency or general accuracy, we repurpose the MoE framework for question-conditioned specialization: standardized questionnaire signals drive routing and supervision at the item level, producing intermediate, psychologically grounded evidence that is subsequently aggregated for personality inference. This reframes the role of the MoE from scaling the encoder to structuring the supervision and providing an interpretable bridge for the under-specified text-to-personality mapping emphasized in our introduction.

B Implementation Details

B.1 Question-to-Construct Assignment

To support question-conditioned routing in the Answer stage, each questionnaire item is associated with a fixed MBTI construct indicator, *i.e.*, one of the four MBTI dimensions: I/E, S/N, T/F, or P/J. We obtain this assignment once before training by prompting an LLM to identify the primary MBTI dimension measured by each questionnaire item, and then keep the resulting mapping fixed throughout training and evaluation. The prompt used for this assignment is shown in Figure 5. The resulting fixed mapping from questionnaire items to MBTI dimensions is reported in Table 6.

Table 6: Fixed mapping from questionnaire items to MBTI dimensions.

Question IDs	Assigned dimensions
Q1–Q10	I/E, S/N, T/F, P/J, T/F, I/E, P/J, T/F, P/J, T/F
Q11–Q20	I/E, S/N, T/F, P/J, T/F, I/E, S/N, T/F, S/N, T/F
Q21–Q30	I/E, S/N, T/F, P/J, T/F, I/E, T/F, T/F, P/J, T/F
Q31–Q40	I/E, S/N, T/F, P/J, P/J, I/E, S/N, T/F, P/J, T/F
Q41–Q50	I/E, S/N, I/E, P/J, T/F, S/N, T/F, T/F, P/J, T/F
Q51–Q60	I/E, S/N, I/E, T/F, T/F, P/J, S/N, T/F, P/J, T/F

B.2 Training Procedure and Computational Setup

Before role-play generation, we first perform the user-level 60/20/20 train/validation/test split. All role-play-based questionnaire responses in the Ask stage are generated exclusively for users in the training partition; no such responses are generated for validation or test users. The reliability score $q_{\mathrm{unc},i}$, importance score $q_{\mathrm{imp},i}$, and resulting question weight w_i are likewise estimated using only the training partition.

In the Ask stage, we employ ChatGPT (`gpt-4o-2024-08-06`) as a role-playing LLM to generate question-level answers to the MBTI-style questionnaire from user posts. The question-conditioned Mixture-of-Experts (MoE) module in the Answer stage consists of 32 experts, each instantiated as a lightweight MLP with a hidden dimension of 1024. In the first stage, we pretrain the Answer module with the regression loss $\mathcal{L}_q$ using a learning rate of 5×10^{-4}, a mini-batch size of 64, and 120 training epochs. In the second stage, we jointly optimize the entire model with the multi-task objective $\mathcal{L} = \lambda_q \mathcal{L}_q + \lambda_{\mathrm{cls}} \mathcal{L}_{\mathrm{cls}}$, where $\lambda_q = 1$ and $\lambda_{\mathrm{cls}} = 0.05$, using a learning rate of 1×10^{-4}

Prompt for Question-to-Construct Assignment

Role and Task:
You are a research assistant in computational psychology and psychometrics. Your task is to assign each questionnaire item to exactly one primary MBTI dimension.

MBTI Dimensions:

- I/E = Introversion vs. Extraversion
- S/N = Sensing vs. Intuition
- T/F = Thinking vs. Feeling
- P/J = Perceiving vs. Judging

Instructions:

1. For each item, choose exactly one primary dimension from: I/E, S/N, T/F, P/J.
2. Assign the dimension that best matches the main psychological construct measured by the item.
3. If an item seems ambiguous, still choose the single closest dimension.
4. Do not provide explanations, rationales, confidence scores, or any extra text.

Output Format:
Return exactly one line per item in the following format:
`Q<item_number>: <dimension>`

Example:

- `Q1: I/E`
- `Q2: S/N`
- `Q3: T/F`

Items to Classify:
`[Paste the 60 questionnaire items here]`

Figure 5: Prompt used for question-to-construct assignment.

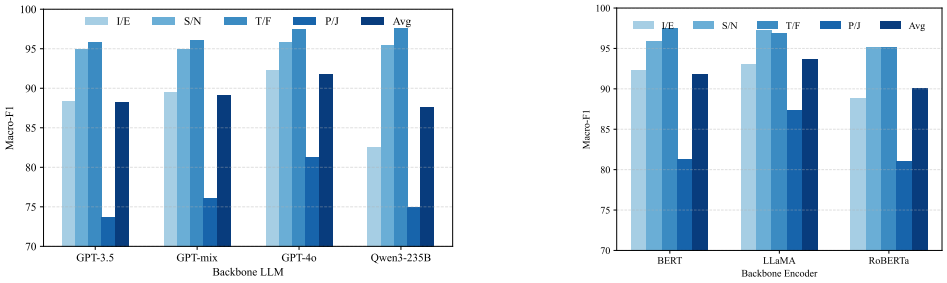
and a mini-batch size of 32. For the main results of ROME, we repeat training and evaluation over five times with different random seeds and report the mean and standard deviation of Macro-F1.

All experiments are conducted on a single NVIDIA RTX 5880 Ada GPU. The Ask stage relies on the ChatGPT API (`gpt-4o-2024-08-06`) only during data preprocessing, while the Answer stage is trained locally using PyTorch. The full training pipeline completes within approximately 1 hour, indicating the efficiency and practicality of the proposed framework.

C Effect of Backbone LLMs and Encoders

To analyze the effect of backbone LLM choice, we vary the LLM deployed as the agent in the *Ask* stage while holding all other components of ROME fixed. Concretely, we consider Qwen3-235B, GPT-3.5, GPT-4o, and a simple ensemble variant (GPT-mix) that averages the question-level answers generated by GPT-3.5 and GPT-4o before training the downstream modules. As shown in Figure 6a, all four settings yield strong results, but their average Macro-F1s follow a clear ordering: Qwen3-235B attains 87.65, GPT-3.5 attains 88.21, GPT-mix improves to 89.16, and GPT-4o achieves the best performance at 91.73. This pattern is consistent with the view that the *quality of the role-play supervision* matters: GPT-4o produces more faithful, less noisy questionnaire answers during Ask stage, which in turn provide a reliable training signal for the Answer and Detect stages. Even so, the Qwen3-235B configuration already significantly exceeds the best baseline method EmoPerso (Table 1, 80.75), indicating that ROME remains effective under more economical LLM choices.

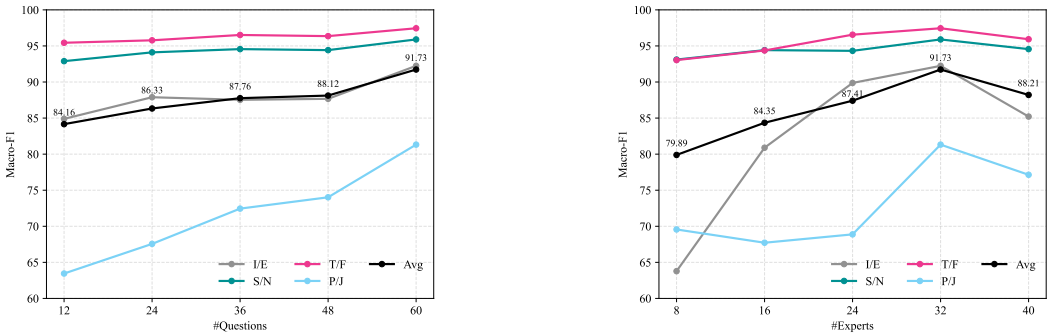
We then examine the impact of the backbone text encoder used to represent posts and questionnaire items. Figure 6b reports results when using BERT-base-uncased (the default configuration in Table 1), Meta-Llama-3-8B-Instruct as a frozen encoder, and RoBERTa-base, with the Ask-stage LLM and all other components unchanged. BERT already supports a strong average Macro-F1 of 91.73, while replacing it with LLaMA further raises the score to 93.61. Overall, these findings indicate that ROME is relatively robust across encoder choices—both BERT and RoBERTa configurations already outperform the best baseline methods—while the LLaMA-based variant further benefits from its stronger capacity to encode heterogeneous, long-range semantics in user posts and questionnaire items.



(a) Backbone LLM comparison.

(b) Backbone encoder comparison.

Figure 6: Effect of backbone LLMs and text encoders on KAGGLE dataset.



(a) Effect of questionnaire length

(b) Effect of the number of experts

Figure 7: Effect of questionnaire length and the number of experts on the KAGGLE dataset.

D Parameters Analysis

D.1 Effect of questionnaire length

We first investigate how many questionnaire items are required for ROME to perform well. As shown in Figure 7a, we randomly subsample a fixed number of items from the 60-item inventory (12, 24, 36, 48, or 60) and retrain the model on KAGGLE under each setting. The average Macro-F1 increases steadily from 84.16 with 12 items to 91.73 with all 60 items, indicating that additional evidence systematically strengthens the psychology-grounded reasoning process and yields more accurate personality predictions. Despite this clear upward trend, ROME remains highly competitive even with substantially shortened questionnaires. With 12 and 24 randomly sampled items, the model attains average Macro-F1s of 84.16 and 86.33, respectively, already outperforming the best baseline method EmoPerso (80.75, Table 1) by relative margins of 4.22% and 6.91%. This demonstrates that the Ask–Answer–Detect framework can extract robust, psychology-aligned evidence even from incomplete questionnaires.

D.2 Effect of the number of experts

We next study how the capacity of the question-conditioned MoE affects performance by varying the number of experts while keeping all other settings fixed. As shown in Figure 7b, increasing the number of experts from 8 to 32 steadily improves the average Macro-F1 on KAGGLE dataset from 79.89 to 91.73, whereas further enlarging the expert pool to 40 yields a slight drop to 88.21. This pattern suggests that a small expert set provides insufficient capacity to capture the heterogeneous psycholinguistic cues associated with different questionnaire items and personality dimensions, while a moderately sized expert pool enables more effective specialization. Beyond this regime, however, the additional experts introduce extra complexity and routing difficulty without translating into further gains.

Table 7: Illustrative question–post evidence for a representative ISTJ user.

Dim.	Q#	Question	Direction	Δlogit	Text evidence (snippet)
I/E	Q21	You enjoy solitary hobbies or activities more than group ones.	I	-3.19	"...Feeling cozy in over sized sweaters, reading a good book in a warm bed..."
I/E	Q41	You avoid making phone calls.	I	-0.62	"The very thought of having my personal space violated in that manner by a complete stranger is terrifying to me."
S/N	Q2	Complex and novel ideas excite you more than simple and straightforward ones.	S	-1.71	"Usually, I don't really like discussing topics like these but the posts here are so relatable."
S/N	Q42	You enjoy exploring unfamiliar ideas and viewpoints.	N	$+0.83$	"That's an interesting perspective – that we shouldn't hold others' imperfections against them."
T/F	Q3	You usually feel more persuaded by what resonates emotionally with you than by factual arguments.	T	-1.54	"That map isn't very reliable. I know INFPs who live outside of Europe/Americas and there isn't a single dot..."
T/F	Q15	You rarely worry about whether you make a good impression on people you meet.	F	$+1.05$	"As a child and adolescent, I would lay awake for nights in a row with my stomach knotting and turning out of guilt, even if the person said it didn't bother them."
P/J	Q4	Your living and working spaces are clean and organized.	J	$+1.02$	"If an assignment, essay, or other work was not perfect in my eyes, I really struggled just handing it in."
P/J	Q34	You find it challenging to maintain a consistent work or study schedule.	J	$+1.00$	"Lately, I have been evaluating my behaviors and life and have concluded that I need to step..."

E Case Study

To understand how ROME integrates questionnaire items and textual evidence to form personality predictions, we conduct a three-part interpretability analysis that examines the model's behavior.

E.1 Question–Post Evidence Chains

To make the Ask–Answer–Detect pipeline more concrete, we present a qualitative case study for a randomly selected test user whose true type is ISTJ and for whom ROME correctly predicts all four MBTI dimensions. As shown in Table 7, for each dimension we select a handful of questionnaire items with the largest impact on the final decision, measured by the change in the predicted logit when the item is removed from the evidence aggregation. We then retrieve representative snippets from the user's posts that GPT-4o, in a post-hoc analysis, identifies as semantically aligned with these high-impact items, with key phrases highlighted for readability.

Across dimensions, the question–post pairs reveal coherent, psychologically plausible evidence patterns. For example, on the I/E axis, the most influential items are supported by snippets describing repeatedly turning down social invitations, a strong need for personal space, and discomfort with close contact or crowded environments, all pointing toward low social needs and a clear inclination for solitude. And for P/J, the evidence clusters around perfectionistic attitudes toward assignments, ongoing self-evaluation, and a strong drive to accomplish personally meaningful goals, reflecting a structured, planful, and self-disciplined J-oriented pattern.

Overall, this case study demonstrates that ROME's personality predictions are supported by a layered chain of question-level diagnostics and post-level linguistic cues.

E.2 Sensitivity of Item Removal

To evaluate the functional role of individual questionnaire items in ROME's inference process and verify whether the learned question weights (w_i in Equation (7)) capture truly diagnostic evidence, we conduct a sensitivity analysis based on item removal where the trained model is kept fixed and only the questionnaire evidence is altered. For each user and each MBTI dimension, we remove exactly one item from the weighted question summary in three ways: deleting the item with the highest learned weight ($\text{ROME}_{\text{drop-max-question}}$), deleting the item with the lowest weight ($\text{ROME}_{\text{drop-min-question}}$), or randomly deleting one item ($\text{ROME}_{\text{drop-rand-question}}$). As shown in Table 8, removing the highest-weight item leads to the largest degradation, reducing the average Macro-F1 from 91.73 to 89.39, while removing the lowest-weight item leaves the overall performance almost unchanged. This pattern

indicates that the learned question weights meaningfully reflect the diagnostic value of individual items: high-weight questions carry indispensable evidence for the final decision, while low-weight questions play a marginal role.

Table 8: Sensitivity of item removal on KAGGLE dataset.

Methods	**I/E**	**S/N**	**T/F**	**P/J**	**Avg**
$\text{ROME}_{\text{drop-max-question}}$	92.06	88.16	96.33	81.02	89.39
$\text{ROME}_{\text{drop-rand-question}}$	92.01	95.08	97.21	81.09	91.35
$\text{ROME}_{\text{drop-min-question}}$	92.16	95.81	**97.45**	**81.43**	91.71
ROME	**92.24**	**95.90**	**97.45**	81.32	**91.73**

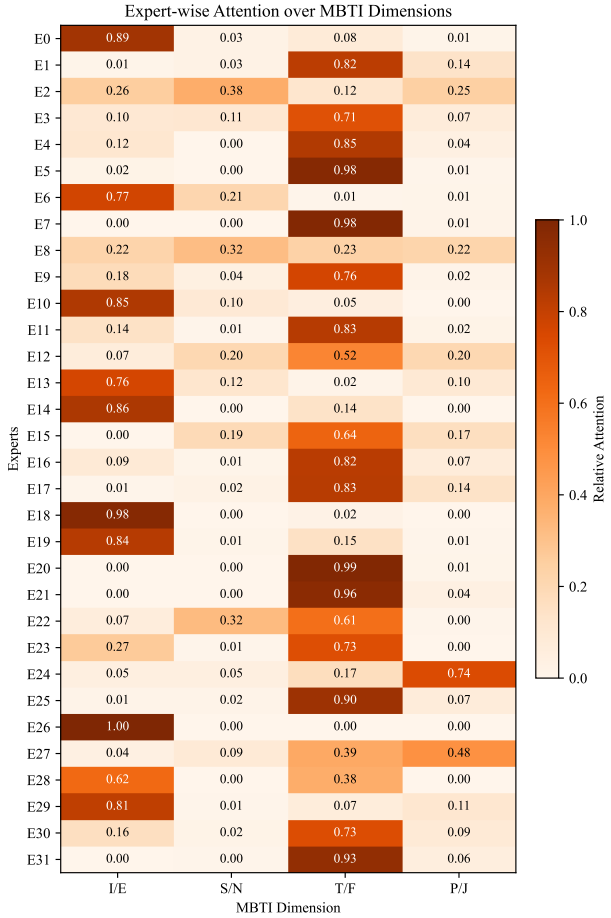


Figure 8: Expert activation analysis on KAGGLE dataset.

E.3 Expert Activation Analysis: Internal Specialization in the MoE

To further interpret our model, we analyze how the question-conditioned MoE allocates expert capacity across personality dimensions. For each user and each questionnaire item, we record the soft routing probabilities assigned to the experts and aggregate them over all users and all items associated with each MBTI dimension. Normalizing these aggregated values yields a 32×4 matrix reflecting the *relative attention* each expert devotes to the four MBTI dimensions (*i.e.*, I/E, S/N, T/F, and P/J).

As visualized in Figure 8, many experts exhibit a clear dominant dimension, concentrating the majority of their attention on one of the four MBTI axes, while others display more mixed routing patterns across dimensions. This specialization suggests that the MoE develops a meaningful division of labor,

with different experts focusing on distinct psychological constructs and personality dimensions, while a small group of shared experts capturing cross-dimensional knowledge.

E.4 Diversity of Role-Played Questionnaire Responses

Complementing the Ask-stage input ablations in Section 3.2, we further examine whether the role-played questionnaire responses preserve user-specific variation or collapse into similar response patterns for users sharing the same personality label. Specifically, we compute pairwise mean absolute error (MAE) between the 60-dimensional response vectors under three comparison settings: repeated generations for the same user, different users sharing the same MBTI type, and different users with different MBTI types.

Table 9: Pairwise MAE between role-played questionnaire-response vectors.

Comparison Setting	Mean MAE	Median MAE
Same user, repeated generations	0.282	0.267
Same MBTI type, different users	0.529	0.477
Different MBTI types, different users	1.912	1.843

As shown in Table 9, repeated generations for the same user yield a mean pairwise MAE of 0.282, reflecting variation introduced by stochastic generation. The distance increases to 0.529 for different users sharing the same MBTI type, indicating that users with the same label still exhibit distinct response profiles beyond within-user generation variability. The distance further increases to 1.912 for users with different MBTI types, showing that the generated responses also preserve broader trait-level distinctions. Overall, the generated responses do not collapse into a single response template determined solely by the personality label.

Table 10 provides an illustrative item-level example. Although both users are labeled ISTP, their responses to Q37, "You are drawn to various forms of creative expression," are consistently opposite across five generations. The corresponding post excerpts also express contrasting attitudes toward creative activities.

Table 10: Illustrative role-played responses to Q37 for two users sharing the same ISTP label.

User ID	Role-Played Responses to Q37	Representative Post Evidence
3782	$-3, -3, -2, -3, -3$	"I have zero desire to write anything."
6170	$+3, +3, +3, +3, +3$	"I'm enveloped in making and watching films. I'm always writing..."

This example complements the aggregate distance analysis by showing that users sharing the same MBTI type can receive substantially different role-played responses when their post evidence differs.

F Additional Experiments

F.1 Comparison with EERPD on PANDORA

To complement the inference-time LLM comparisons in Section 3.3, we further evaluate EERPD on the PANDORA dataset using the same user-level data split as ROME. Table 11 reports the Macro-F1 scores for all four MBTI dimensions.

Table 11: Comparison with EERPD on the PANDORA dataset.

Methods	I/E	S/N	T/F	P/J	Avg
EERPD	62.14	56.98	70.68	60.57	62.59
ROME	**74.93**	**95.45**	**82.21**	**78.22**	**82.70**

ROME achieves an average Macro-F1 of 82.70, outperforming EERPD (62.59) by 20.11 points, with higher scores across all four dimensions. One possible explanation lies in the different modeling strategies: EERPD relies on emotion and emotion-regulation cues, which may be less informative for topic-driven or context-dependent Reddit comments. In contrast, ROME learns questionnaire-grounded evidence covering all four personality constructs without relying exclusively on affective information. These results further demonstrate the effectiveness of ROME on the heterogeneous user-generated content in PANDORA.

F.2 Generalization to Big Five Personality Traits

To examine the applicability of ROME beyond MBTI, we further evaluate it on the Essays dataset [28] under the Big Five personality framework and compare with the reported results of EERPD [20]. Specifically, we replace the MBTI-style questionnaire with the 44-item Big Five Inventory (BFI-44) [14], assigning each item to one of five traits: Agreeableness (AGR), Conscientiousness (CON), Extraversion (EXT), Neuroticism (NEU), and Openness (OPN).

Table 12: Performance comparison on the Essays dataset under the Big Five personality framework.

Methods	AGR	CON	EXT	NEU	OPN	Avg
EERPD	65.01	**68.64**	63.02	56.00	60.93	62.72
ROME	**65.19**	63.31	**65.06**	**61.69**	**65.47**	**64.14**

As shown in Table 12, ROME achieves an average Macro-F1 of 64.14, outperforming EERPD (62.72) by 1.42 points. ROME performs better on four out of five traits, with particularly notable improvements on NEU (+5.69) and OPN (+4.54), although EERPD performs better on CON. These results suggest that the questionnaire-grounded intermediate supervision and Ask–Answer–Detect framework can be extended beyond MBTI to a different personality construct system without changing the main architecture.

G Limitations and Broader Impacts

Limitations. The current evaluation focuses on personality detection using English-language datasets, and its generalization to other languages, downstream tasks, or clinical scenarios remains an important direction for future work. Nevertheless, this limitation is orthogonal to our core methodological contributions.

Broader impacts. ROME has the potential to benefit personality-aware web applications by substantially improving the interpretability and transparency of personality detection. By producing questionnaire-level evidence, it enables researchers and practitioners to trace predictions back to specific psychometric items, offering a more structured and theoretically grounded view of user traits. Beyond application performance, ROME contributes to more accessible and efficient AI systems. Despite its strong performance, ROME relies on a relatively lightweight backbone model, making it computationally efficient and easier to deploy in real-world settings with limited resources. This efficiency may facilitate broader adoption in academic and industry applications, especially in scenarios where large-scale LLM deployment is impractical or cost-prohibitive.

At the same time, personality detection from user data raises concerns regarding privacy, profiling, and potential misuse. We encourage responsible deployment with appropriate safeguards, including transparency to users, informed consent where applicable, and restrictions on sensitive or high-stakes use cases.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims in the abstract and introduction are supported by the proposed Ask–Answer–Detect framework, the question-conditioned MoE architecture, and the experimental results reported in Section 3.1. The claimed improvements are explicitly reported on the KAGGLE and PANDORA datasets.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The limitations of ROME are discussed in Appendix G.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A]

Justification: The paper does not present formal theoretical results, theorems, or proofs. The proposed framework is evaluated empirically through experiments.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The paper describes the model architecture in Section 2, the experimental protocol in Section 3.1, and implementation details in Appendix B. The code repository is also provided in the abstract to support reproducibility.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The paper provides a public GitHub repository (https://github.com/JasmineLyf/ROME), linked in the abstract, to support reproducibility. The experiments use existing personality detection datasets, with implementation details and data splits described in Appendix B.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper specifies the evaluated datasets, train/validation/test split, evaluation metric, baselines, and ablation settings in Section 3.1 and Section 3.2. The training procedure, hyperparameters, model size, and computational setup are further provided in Appendix B.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Our results are reported as the mean and standard deviation over five runs with different random seeds.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.

- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Appendix B reports the computational setup, including the use of a single NVIDIA RTX 5880 Ada GPU, the use of the ChatGPT API during preprocessing, and the approximate runtime of the full training pipeline.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research uses existing publicly available personality datasets and does not involve newly recruited human subjects, interventions, or attempts to re-identify individuals. We also discuss potential privacy and profiling risks and responsible deployment considerations in the Broader Impacts section.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: The broader impacts are discussed in Appendix G.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: The paper does not release a high-risk pretrained model, generative system, or newly scraped dataset. The released materials are intended to reproduce the reported experiments rather than to enable open-ended generation or deployment.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: The creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The paper introduces code and generated questionnaire-level evidence for the proposed Ask–Answer–Detect framework. The generation protocol, question-to-construct assignment, training procedure, and implementation details are documented in Section 2 and Appendix B.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing experiments or newly recruited human subjects. All experiments are conducted on existing personality detection datasets.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: The paper does not involve newly recruited human subjects, interventions, surveys, or crowdsourcing experiments conducted by the authors. The study uses existing publicly available datasets.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: LLM usage is a core component of the proposed method. Section 2.3 describes how a role-playing LLM generates question-level answers in the offline Ask stage, and Appendix B specifies the use of `gpt-4o-2024-08-06` and related implementation details.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.