Evidence-Guided Bidirectional Iterative Reasoning for Interpretable Personality Recognition

Tinghao Chen[1,2]* **Yuxuan Song**[1,2]* **Mengyu Xiang**[1,2] **Qiudan Li**[1]† **Shu Wu**[1,2]†
Daniel Dajun Zeng[1,2]
[1]State Key Laboratory of Multimodal Artificial Intelligence Systems (MAIS)
Institute of Automation, Chinese Academy of Sciences
[2]School of Artificial Intelligence, University of Chinese Academy of Sciences
chentinghao2026@ia.ac.cn, songyuxuan2023@ia.ac.cn, xiangmengyu2024@ia.ac.cn
qiudan.li@ia.ac.cn, shu.wu@nlpr.ia.ac.cn, zengdaniel@outlook.com

Abstract

Identifying personality accurately can improve interpretability and adaptability for human–computer interaction and psychological assessment. Recent studies have introduced evidence utterance into large language models (LLMs) for personality recognition, providing supporting evidence to explain personality labels. However, the mixed background and social expressions in dialogue scenarios interfere with the identification of evidence utterance. At the same time, the overconfidence of LLMs leads to biased personality inference. Inspired by the lens theory in psychology, how to systematically explore the mutual influence between evidence utterance identification and personality recognition while improving the performance of both tasks remains a key challenge. In this paper, we reframe personality recognition as an Evidence-Personality Iterative Reasoning problem (EPIR). A graph model is employed to select an initial evidence utterance set, and personality recognition is performed based on a Generative-Discriminative Feedback mechanism. Then, the inferred results are fed back to the graph mapping network to dynamically adjust evidence utterance filtering. This interactive iterative process continues until the evidence and personality results across different iterations converge. Experimental results on the two public datasets demonstrate the effectiveness of the proposed method.

1 Introduction

With the rapid development of intelligent interactive systems and online communication platforms, dialogue has become a primary medium for individuals to express attitudes, emotions, and interpersonal interactions. The emotional trajectories, language styles, and interaction patterns in conversations can often reflect individuals' personality traits. Therefore, achieving automatic personality recognition not only helps improve the intelligence and adaptability of human-computer interaction systems, but also provides an important auxiliary tool for mental health assessment(Redelmeier et al., 2021; Attig et al., 2017).

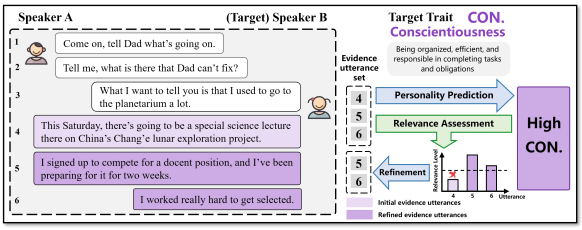


Figure 1: An Illustrative Example of Evidence-Personality Iterative Reasoning

Most existing studies mainly employ graph models (Song et al., 2025; Zhu et al., 2024) or large language models (LLMs) (Yeo et al., 2025; Ma et al., 2025) to model text content and interaction information, treating personality recognition as an end-to-end classification task. The models only output personality labels without evidential support. To improve the interpretability of personality recognition, Sun et al. (2024) first introduced the concept of evidence utterances, using LLMs to explicitly mine utterances containing key personality cues from dialogues and treating them as interpretable outputs along with the personality recognition process. This work points out that accurately identifying evidence utterances and evaluating their impact on personality recognition performance are challenging tasks. In addition, accurately identifying evidence utterances also contributes to enhancing the interpretability of personality recognition.

However, the mixed background and social expressions in dialogue scenarios interfere with the identification of evidence utterances, leading to redundancy in the identification results. Meanwhile, LLMs still exhibit bias in their inferences about

*These authors contributed to the work equally.
†To whom correspondence should be addressed.

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 37877–37896
October 24-29, 2026 ©2026 Association for Computational Linguistics

personality levels due to overconfidence. To address these challenges, we draw inspiration from the Lens Model (Brunswik, 1952) in classical psychological theory, which highlights the mutual influence between personality labels and identifying evidence utterances. As shown in Figure 1, in a dialogue between speaker A and B discussing a planetarium presentation activity, the evidence utterances are initially identified as utterances 4, 5, and 6, inferring speaker B's Conscientiousness level as high. However, utterance 4 is merely background information and cannot reflect Conscientiousness on its own, whereas utterances 5 and 6 directly reflect careful preparation and therefore indicate Conscientiousness. Guided by the Lens Model, the predicted trait, combined with relevance assessment, can help remove irrelevant utterances like 4 and refine key evidence utterances as 5 and 6. Therefore, separately locating evidence or predicting personality is insufficient, an explainable personality prediction system should provide an interaction mechanism that bridges these two components. Specifically, evidence utterance is introduced into the personality recognition process to provide interpretable support for personality inference, while the personality recognition results further inform and guide the selection and refinement of evidence utterances.

Motivated by this insight, we reframe personality recognition as an Evidence-Personality Iterative Reasoning problem (EPIR), which takes a first step towards establishing a bidirectional iterative reasoning paradigm between evidence utterance identification and personality recognition. Specifically, a graph model is employed to select an initial evidence utterance set, and personality recognition is performed based on a Generative-Discriminative Feedback mechanism (GDF). The LLM first generates utterances reflecting possible speaker behaviors under different personality levels. Then, an LLM-as-a-judge discriminator evaluates which level of behavior the original utterance most closely resembles, and subsequently feeds it back to the graph mapping network as personality result to dynamically adjust evidence utterance filtering. This interactive iterative process continues until the evidence and personality results across different iterations converge.

Given the continued scarcity of publicly available evidence-annotated dialogue personality data, we conduct experiments on PersonalityEvd (Sun et al., 2024), the largest publicly available dataset with evidence utterance annotations. We further introduce PANDORA (Gjurković et al., 2021) as a complementary dataset to evaluate the proposed framework in a different data setting. Accordingly, we investigate the following three research questions:

- ***RQ1: What is the potential gain of evidence utterances as explicit intermediate information for personality recognition performance?***

  Injecting ground-truth evidence utterances into the LLM's personality recognition process yields an average 3.47% improvement over direct LLM prediction.

- ***RQ2: Which type of model performs better in the evidence utterance identification task?***

  Comparing the performance of graph models and different large model approaches on evidence utterance identification, the former achieves the highest F1 score of 83.10%, surpassing the second-best LLM result of 80.19%, revealing that graph structures better capture the personality cues and their relational interactions in dialogues.

- ***RQ3: Can the EPIR framework and GDF mechanism bring gains to the personality recognition task?***

  We compare EPIR against existing personality recognition methods and analyze component contributions through ablation studies. EPIR outperforms the strongest baselines on both PersonalityEvd and PANDORA, while achieving a 2.73% improvement over direct evidence injection on PersonalityEvd. Ablation results confirm that both the EPIR framework and the generative-discriminative feedback mechanism contribute to the improvement.

2 Related Work

Traditional personality prediction research primarily relies on explicit linguistic features, such as LIWC (Tausczik and Pennebaker, 2010) and bag-of-words models (Zhang et al., 2010). With the advancement of deep learning, neural network architectures such as CNN (Xue et al., 2018), LSTM (Tandera et al., 2017), and Transformer (Leonardi et al., 2020) have been adopted for personality recognition, outperforming traditional methods in

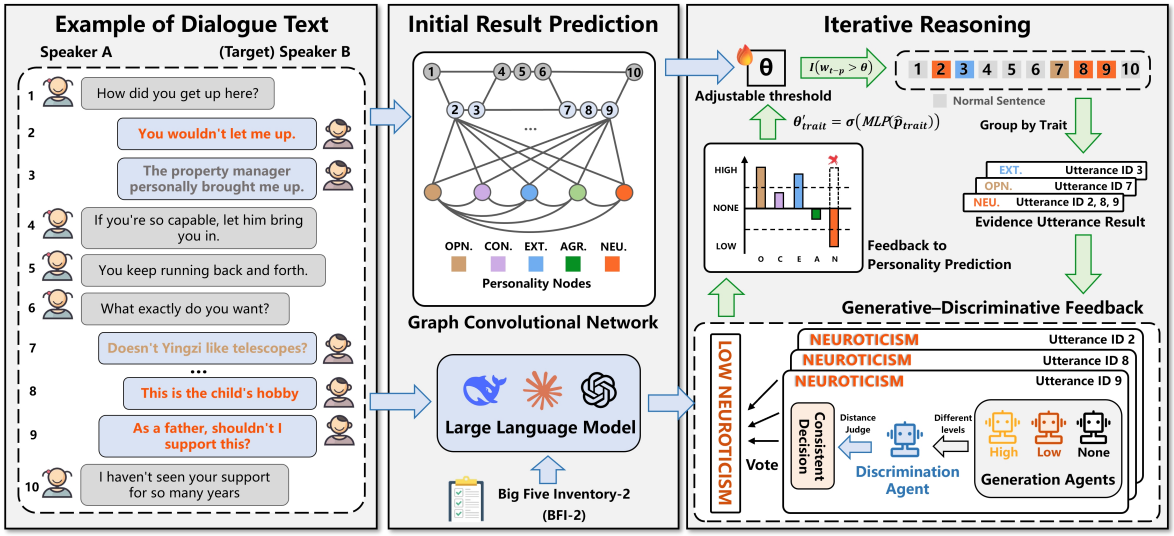


Figure 2: Overall architecture of EPIR.

feature representation and recognition performance. In particular, pre-trained language models such as BERT (Devlin et al., 2019) can model contextual information through effective semantic encoding, thereby extracting richer features than traditional explicit linguistic feature-based methods (Arijanto et al., 2021; Jun et al., 2021) and achieving better performance. As research progresses, graph neural networks have been introduced to model complex structured relationships (Song et al., 2025; Yang et al., 2023a; Yang et al., 2021), such as speaker interactions, contextual changes, and emotional fluctuations (Fu et al., 2024; Han et al., 2023; He et al., 2025). In recent years, LLMs have shown strong potential for personality prediction due to their zero-shot capability and chain-of-thought reasoning (Ganesan et al., 2023; Yang et al., 2023b). Some methods leverage LLMs to extract personality-related features from multiple perspectives, such as emotion, semantics and interaction (Hu et al., 2024; Yeo et al., 2025), while others fine-tune LLMs to improve performance (Ma et al., 2025; Shen et al., 2025b).

To advance interpretable personality prediction, CoPE (Sun et al., 2024) constructs a reasoning process from short-term personality states to long-term stable personality traits and introduces evidence utterances that reflect key personality cues into the prediction pipeline. However, this approach performs personality prediction and evidence utterance identification separately. Meanwhile, recent studies have explored closer integration between structured evidence modeling and task-level reasoning, ranging from label-informed graph structure learning (Wang et al., 2021) and LLM-enhanced graph-based prediction (Xia et al., 2024) to LLM-guided dynamic graph adaptation (Xiang et al., 2025). Nevertheless, there remains a lack of systematic investigation into the interplay between these two tasks that would yield deeper interpretability of the prediction process. This paper aims to address this gap by reframing personality prediction as an Evidence-Personality Iterative Reasoning problem, and answering the following questions: how personality predictions can inform and refine evidence selection, and how accurate evidence identification contributes to personality prediction performance.

3 Preliminary

Given a multi-turn dialogue $D = \{u_1, u_2, \ldots, u_n\}$ and a target speaker s, where each utterance u_i carries a speaker identifier, the goal is to jointly perform personality prediction and evidence utterance identification for s. Following the Big Five personality theory, the personality trait set is defined as $\mathcal{T} = \{OPN, CON, EXT, AGR, NEU\}$, representing Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism. For evidence utterance identification, the model predicts a binary label $\mathcal{E}_{i,t} \in \{0, 1\}$ for each utterance u_i of the target speaker and each personality trait $t \in \mathcal{T}$, where $\mathcal{E}_{i,t} = 1$ indicates that u_i serves as an evidence utterance for trait t. For personality recognition, the model predicts the target speaker's personality level $\hat{p}_t \in \{Low, High, Cannot\ judge\}$ for each trait t. We formulate the task as the follow-

ing mapping: $f(D, s) \rightarrow (\{\hat{p}_t\}_{t\in\mathcal{T}}, \{\mathcal{E}_{i,t}\})$, given a dialogue and the target speaker, jointly predict the five trait levels and the corresponding evidence utterances for each trait.

4 Methodology

The overall framework of the proposed approach *Evidence-Personality Iterative Reasoning (EPIR) Framework* is shown in Figure 2, which consists of two main components: (1) *Graph-based evidence utterance identification*, it employs a Graph Convolutional Network (GCN) (Kipf and Welling, 2017) to capture the associations between utterances and different personality trait representations to identify key utterances reflecting specific traits; (2) *Explicit evidence-aligned LLM personality prediction*, it uses a generative-discriminative feedback mechanism to improve the LLM's current personality recognition results by identified evidence utterances. Building upon this, EPIR feeds the output of the former as explicit evidence input to the latter, and further feeds the latter's prediction results back to the evidence utterance identification process to adjust evidence filtering conditions, thereby forming an iterative reasoning framework for evidence utterance identification and personality prediction.

4.1 Evidence-Personality Iterative Reasoning (EPIR) Framework

In dialogue, evidence utterance identification and personality prediction are inherently interdependent tasks, thus, we introduce a bidirectional feedback mechanism to enable collaborative modeling. It begins with separate initialization on both sides. On the evidence side, a graph structural representation is constructed from the dialogue input, and an initial evidence utterance set $\mathcal{E}^{(0)}$ is mined from the target speaker's utterances. For personality trait t, its initial evidence utterance set is defined as:

$$\mathcal{E}_t^{(0)} = \{u_i \mid s(u_i, t) > \theta^{(0)},\ u_i \in D_s\}, \quad (1)$$

where D_s denotes the set of utterances by the target speaker, $s(u_i, t)$ is the relevance score between utterance u_i and trait t, and $\theta^{(0)}$ is the evidence utterance filtering threshold learned from the similarity between utterances and traits in the graph model. On the personality side, chain-of-thought prompts combining the dialogue content with personality trait descriptions from the Big Five Inventory-2 (BFI-2) (Soto and John, 2017; Zhang et al., 2022) psychological questionnaire guide the LLM to perform step-by-step analysis and complete initial personality inference under a zero-shot setting, obtaining the initial personality prediction result $\hat{\mathbf{p}}^{(0)}$:

$$\hat{\mathbf{p}}^{(0)} = \mathcal{M}(Dialogue_{raw}, Scale_{BFI-2}), \quad (2)$$

The initial evidence utterance set and initial personality prediction results together constitute the start for the bidirectional iterative process. For the $j-th$ iteration, the model first employs the Generative-Discriminative Feedback mechanism to update personality prediction results based on the current evidence utterance set $\mathcal{E}^{(j)}$; subsequently, the updated personality results are fed back to the evidence utterance identification module to adjust the evidence filtering threshold $\theta^{(j)}$ and update the evidence utterance set. The new set then enters the next iteration to repeat this process. Through this alternating update process, evidence utterances provide explicit support for personality prediction, while personality results inversely refine evidence selection, enabling both to progressively converge across multiple iterations.

The above process continues until personality results and evidence utterance sets between two adjacent iterations no longer change, then the model outputs the final personality result $\hat{\mathbf{p}}$. Through this bidirectional interaction mechanism, EPIR is able to continuously refine evidence and optimize personality results during the reasoning process, thereby achieving more precise evidence identification and more reliable personality prediction. The overall process is shown in Algorithm 1.

4.2 Graph-based Evidence Utterance Identification

To implement evidence utterance identification within the EPIR framework, we adopt a hierarchical heterogeneous graph representation commonly used in dialogue modeling, which uniformly models utterance content, speaker interaction relationships, and personality trait representations. By learning the associations between utterance nodes and personality nodes, the method filters evidence utterances that can reflect specific traits. The resulting evidence utterance set serves as explicit input to the subsequent personality prediction module, while the evidence filtering threshold is dynamically adjusted during the EPIR iterative process based on personality results.

Algorithm 1 Overall Inference Procedure

Require: Dialogue D; Large Language Model $\mathcal{M}$
Ensure: Final personality prediction $\hat{\mathbf{p}}$

```
Construct dialogue graph and compute node representations
Mine initial evidence utterances E^(0) with initial threshold θ^(0)
Obtain base personality prediction p̂^(0) via M
Initialize j ← 0
repeat
    j ← j + 1
    Generate counterfactual utterances under each trait level via M (Eq. 6)
    Decide trait levels for each evidence utterance via M-as-a-Judge scoring (Eq. 7, 8)
    Refine dialogue-level personality prediction p̂^(j) by voting aggregation (Eq. 9)
    Adjust evidence thresholds θ^(j) based on refined prediction
    Re-mine evidence utterances E^(j)
until p̂^(j) = p̂^(j−1) and E^(j) = E^(j−1)
return p̂
```

Specifically, for each input instance, the dialogue utterances and personality representations are jointly modeled as a graph, denoted as $G = \{V, E\}$, where V is the set of nodes and E denotes the edges between them. The graph contains three types of nodes, representing utterances from Speaker A and Speaker B, and five personality trait representation nodes, respectively.

During the initialization phase, all node feature vectors are generated using the pre-trained model BERT. For dialogue nodes, the corresponding text is used as input, while personality nodes take the textual descriptions of corresponding traits from the BFI-2 psychological inventory as input. All inputs are encoded by BERT, and the output [CLS] vectors are used as the initial feature embeddings of the nodes.

Regarding graph structure design, edges between nodes include three types of relationships: intra-speaker edges, inter-speaker edges, and personality node edges. As shown in Figure 2, for adjacent utterances from the same speaker, edges are established in chronological order to explicitly encode temporal dependencies and semantic continuity within the sequence. For inter-speaker edges, connections are only established between nodes corresponding to utterances where a speaker switch occurs, modeling the reply relationships and cross-speaker information transfer. For personality node edges, each personality node is connected to all target speaker utterance nodes, forming the basis for mining evidence relationships between specific utterances and traits. Additionally, all personality nodes are fully connected to each other to align the representation spaces, preventing feature vector outliers. All edges are undirected.

We employ a two-layer GCN to capture both the similarities among text nodes and the associations between texts and personalities, utilizing local graph structural information to enrich the feature contained in node vectors, thereby constructing global node representations $\mathbf{H}$ for evidence utterances:

$$\mathbf{H} = \widetilde{\mathbf{A}} \cdot ReLU(\widetilde{\mathbf{A}}\mathbf{X}\mathbf{W_1})\mathbf{W_2}, \tag{3}$$

where $\mathbf{A}$ is the adjacency matrix representing graph edge information and $\widetilde{\mathbf{A}}$ denotes its normalized form, $\mathbf{X}$ denotes the initial node embeddings, and $\mathbf{W}_1$, $\mathbf{W}_2$ are trainable parameters.

For the global node representations, the method computes the cosine similarity between the feature vectors of all utterance nodes from the target speaker, $\mathbf{H}_{tar}$, and the feature vectors of the five personality nodes, $\mathbf{H}_{per}$, and identifies utterance as an evidence utterance for trait t based on whether the similarity exceeds the threshold $\theta^{(j)}$, similar to Equation 1:

$$\mathcal{E}_t^{(j)} = \{u_i \mid s(u_i, t) > \theta^{(j)},\ u_i \in D_s\}, \tag{4}$$

$$s(u_i, t) = \cos(\mathbf{H}_{tar}, \mathbf{H}_{per}) = \frac{\mathbf{H}_{tar}^{\top}\mathbf{H}_{per}}{\|\mathbf{H}_{tar}\|\,\|\mathbf{H}_{per}\|}, \tag{5}$$

where $\theta^{(j)}$ is adjusted based on the results of the personality prediction module.

During the training phase, binary cross-entropy loss is employed for supervised learning, optimizing the parameters of the graph and the upstream BERT encoder. During EPIR inference, the model directly uses the trained model to identify evidence utterances for the current iteration, which serve as explicit input to the subsequent personality prediction module.

4.3 Explicit Evidence-Aligned LLM Personality Prediction

When evidence utterances are used as the prompt input for the LLM, due to the model's overconfidence, it is difficult to achieve an accurate assessment of the fine-grained personality level. For example, when faced with unexpected expenses,

a speaker may produce short-term, frequent, and exaggerated complaints such as "If I had known it was so expensive, I would have just hired a nanny for you", "Can we still drop out of school now?", which lead the LLM to conclude that the speaker has a high level of neuroticism. However, such complaints are behaviors that the ordinary speaker would exhibit in this context, and have nothing to do with neurotic trait. Therefore, how to guide the model to pay attention to this domain-relevant knowledge is the key challenge.

To address this issue, we design a Generative-Discriminative Feedback (GDF) mechanism as shown in Steps 6, 7, and 8 of Algorithm 1. It first generates possible speaker behaviors at high, low, and cannot judge levels for a specified trait. Then, a discriminator evaluates the original utterance to determine which level of trait it most closely resembles, providing feedback to help LLM learn relevant knowledge and re-evaluate.

Specifically, the Generator leverages a counterfactual reasoning mechanism, and utilizes the personality embedding capability of LLMs to simulate the expressions of evidence utterances under different trait levels based on the dialogue history. For example, for the $i-th$ evidence utterance u_i, generator sequentially generates the $i-th$ utterance content under three trait levels (i.e. high, low, and cannot judge) for each trait based on the preceding $i-1$ utterances in the dialogue:

$$u_i' = Generator(u_1 \dots u_{i-1}, trait, level). \quad (6)$$

The Discriminator adopts a LLM-as-a-Judge approach to score the similarity between the original evidence utterance and different candidate utterances, quantifying their differences in the current dialogue context. For the $i-th$ evidence utterance, the discrimination process is expressed as:

$$Sim_{i,level,trait} = Score_{LLM}(u_i, u_i'), \quad (7)$$

$$p_i = \arg\max_{level}(Sim_{i,level,trait}). \quad (8)$$

where a higher similarity score reflects greater semantic consistency between the candidate utterance and the original evidence utterance. Therefore, the trait level corresponding to the generated utterance exhibiting maximal similarity is regarded as the local personality trait level result p_i for the current evidence utterance.

In the aggregation update component, the local prediction results from all evidence utterances under the same trait are aggregated, and a voting strategy is adopted to obtain the new prediction result for the current dialogue on trait t. Specifically, if the personality trait level $\hat{p}_t^{(j-1)}$ from the previous iteration is cannot judge, we retain it; otherwise, the new prediction result is used to update it:

$$\hat{p}_t^{(j)} = \begin{cases} \hat{p}_t^{(j-1)}, \text{if } \hat{p}_t^{(j-1)} = Cannot\ judge, \\ \arg\max_{c \in \mathcal{P}} \sum_{i=1}^{m} \mathbb{I}(p_i = c), \text{otherwise.} \end{cases} \quad (9)$$

where $\mathcal{P} = \{High, Low, Cannot\ judge\}$ is the set of trait level labels, $\mathbb{I}(\cdot)$ is the indicator function that equals 1 when the condition in parentheses is true and 0 otherwise, and m is the number of evidence utterances for the corresponding personality trait in the current dialogue. The updated personality prediction result constitutes the output of the personality prediction module for the current iteration.

5 Experiment Setup

5.1 Dataset

We evaluate the proposed framework on two personality recognition datasets, PersonalityEvd (Sun et al., 2024) and PANDORA (Gjurković et al., 2021). PersonalityEvd is built from the Chinese emotional dialogue corpus CPED (Chen et al., 2022), providing three-level Big Five labels (i.e., high, low, and cannot judge) and corresponding evidence utterance annotations, while PANDORA is an English Reddit-based personality dataset, for which we construct binary Big Five labels (i.e., high, low, excluding cannot judge) and post-level evidence annotations to support evidence-aware personality recognition. More details about the dataset and the annotation process are provided in Appendix D.

5.2 Baseline

We select three categories of representative methods from existing personality recognition research as baselines: (i) content-based methods: TextCNN (Kampman et al., 2018) and BERT (Devlin et al., 2019); (ii) relation-based methods: HC-GNN (Fu et al., 2024), DSIG (Xie et al., 2025), SAH-GCN (He et al., 2025), KEHG (Song et al., 2025), DEN (Zhu et al., 2025), and Semi-PerGCN (Zhu et al., 2024); (iii) LLM-based methods:

Model	Evidence utterance F1 score (%)						std.dev
	Avg.	OPN	CON	EXT	AGR	NEU	(pp)
GPT	75.92	83.49	86.17	60.33	73.85	75.76	9.05
Claude	67.53	74.72	75.49	58.53	68.65	60.26	7.07
DeepSeek	77.28	90.67	86.97	74.61	72.53	61.63	10.47
CoPE	75.94	-	-	-	-	-	-
Qwen	80.19	87.59	87.23	77.07	74.49	74.57	5.97
Graph	**83.10**	**91.22**	**88.31**	**79.41**	**75.48**	**81.06**	5.81

Table 1: Evidence Utterance Identification Performance on PersonalityEvd.

Method	Personality Accuracy (%)					
	Avg.	OPN	CON	EXT	AGR	NEU
GPT	60.33	72.18	62.53	46.28	60.33	60.33
w/ Evi	62.48	71.90	63.36	49.04	65.84	62.26
DeepSeek	63.52	79.83	67.31	57.22	62.01	51.25
w/ Evi	65.67	77.41	66.12	58.95	62.53	63.36

Table 2: Effect of Ground-Truth Evidence Utterance Injection on Personality Prediction Accuracy for PersonalityEvd.

GPT-4-Turbo (OpenAI, 2025), Claude-4-Sonnet (Anthropic, 2025), DeepSeek-V3.2 (DeepSeek-AI et al., 2025), ETM (Bi et al., 2025), and EmoPerso (Shen et al., 2025a), as well as LoRA (Hu et al., 2022)-fine-tuned CoPE (Sun et al., 2024) and Qwen3-8B (Team, 2025). Details of each baseline and evaluation metrics are provided in Appendix A.1 and A.2.

5.3 Implementation Details

We implement the model in PyTorch and train it with AdamW for 50 epochs, using BERT-base-chinese for PersonalityEvd and BERT-base-uncased for PANDORA to generate node embeddings and employing 256-dimensional GCN layers. Additional implementation details, including hyperparameter choices and LLM inference settings, are reported in Appendix A.3.

6 Experiment Results

We first investigate evidence-related effects on PersonalityEvd, which provides high-quality manually annotated evidence utterances, followed by an overall evaluation of personality recognition performance across different data settings with PersonalityEvd and PANDORA.

6.1 Impact of Evidence Utterance Injection on Personality Prediction Performance

We compare the personality prediction accuracy of GPT and DeepSeek before and after injecting ground-truth evidence utterances on PersonalityEvd. As shown in Table 2, after evidence utterance injection, GPT's average accuracy increases from 60.33% to 62.48%, and DeepSeek's average accuracy increases from 63.52% to 65.67%, with relative improvement ratios of 3.56% and 3.38%, respectively. The results demonstrate that explicitly introducing evidence utterances can effectively guide LLMs to focus on utterances with genuine personality diagnostic value, reducing interference from irrelevant background information and improving overall recognition accuracy.

6.2 Evidence Utterance Identification Performance Across Models

Table 1 compares the performance of LLMs and our graph model on PersonalityEvd for evidence utterance identification. It can be observed that the graph model approach achieves the best F1 scores across all five personality traits, with an average F1 of 83.10%, representing a 3.63% improvement over the second-best model Qwen. This confirms that compared to LLMs, the graph model better captures the structural information of different speakers and contextual interactions in dialogues, enhancing evidence utterance identification performance. At the same time, while achieving superior performance, the lightweight graph model introduces only about one-thousandth additional latency, as shown in Appendix C.

6.3 Performance of EPIR on Personality Prediction

We compare the proposed model with all baseline models for personality prediction, using the accuracy of each trait as well as the average accuracy across all traits as evaluation metrics. All experiments were repeated five times with random seeds, and the averaged results are reported in Table 3.

On PersonalityEvd, among different types of methods, content-based models showed relatively low performance, with average accuracies of 55%–58%. BERT improved over TextCNN but only marginally, indicating a limitation of modeling personality solely from text semantics. Relation-

Method	Personality Accuracy (%)						std.dev
	Avg.	OPN	CON	EXT	AGR	NEU	(percent point, pp)
PersonalityEvd							
TextCNN	54.85	70.52	**71.07**	51.52	39.39	41.77	13.64
BERT	57.99	71.28	65.05	53.29	47.40	52.94	8.79
HC-GNN	50.78	50.83	50.61	50.15	48.76	53.50	1.54
DSIG	56.20	71.07	66.94	50.69	47.38	44.90	10.70
SAH-GCN	58.35	70.80	70.80	56.20	48.76	45.18	10.77
KEHG	61.27	70.25	70.25	59.78	50.97	55.10	7.84
DEN	54.77	69.42	60.61	56.20	45.73	41.87	9.99
GPT	60.33	72.18	62.53	46.28	60.33	60.33	8.28
Claude	60.61	68.60	64.19	53.44	60.33	56.47	5.39
DeepSeek	63.52	**79.83**	67.31	57.22	62.01	51.25	9.73
Qwen	66.34	79.61	69.42	62.53	60.88	59.23	7.49
ETM	58.51	72.45	63.36	59.50	49.04	48.21	9.11
EmoPerso	62.09	72.73	66.39	64.74	60.06	46.56	8.76
CoPE	66.45	75.48	68.87	63.39	63.36	**61.15**	5.18
EPIR (GPT)	63.69	71.35	63.09	60.88	65.79	57.35	4.72
EPIR (DeepSeek)	**67.46**	74.93	68.26	**65.45**	**68.60**	60.06	4.83
PANDORA							
TextCNN	50.33	48.34	49.52	52.84	50.55	50.41	1.48
BERT	52.41	51.37	53.73	52.77	54.69	49.52	1.82
Semi-PerGCN	51.13	52.10	51.00	50.63	50.77	51.14	0.52
KEHG	50.94	51.51	51.59	48.93	50.92	51.73	1.04
DEN	52.28	53.28	51.51	51.14	52.69	52.77	0.81
DeepSeek	56.75	53.14	**58.67**	58.30	**57.56**	56.09	2.01
Qwen	53.05	53.28	54.32	51.81	52.84	52.99	0.81
ETM	54.13	53.51	51.59	55.87	54.02	55.65	1.56
EmoPerso	55.06	54.91	56.97	56.75	54.24	52.40	1.69
EPIR (DeepSeek)	**57.86**	**56.46**	56.09	**60.15**	56.83	**59.78**	1.74

Table 3: Overall Personality Prediction Results.

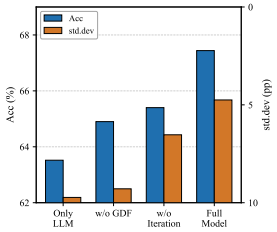
based models performed better, with KEHG achieving 61.27%, a 5.66% improvement over BERT, showing the benefit of structural modeling. LLM-based models showed substantial accuracy improvements, with average accuracies ranging from 58% to 66%. The best baseline model CoPE reached 66.45%, reflecting their strong ability to capture high-level semantic and personality cues. Among all baselines, the proposed method with DeepSeek achieved the best result at 67.46%, surpassing the second-best CoPE by 1.52%, and improving over GPT or DeepSeek alone by 5.57% and 6.20%, respectively, demonstrating its effectiveness.

Similarly, on PANDORA, EPIR (DeepSeek) achieved the highest average accuracy of 57.86%, yielding a relative improvement of 1.96% over the best baseline. This result provides complementary validation of the proposed method in a different linguistic and cultural setting, supporting its broader robustness and applicability.

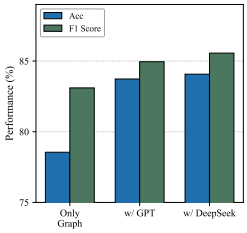
In addition, the proposed model can also reduce performance variance across different personality traits, yielding more balanced predictions on PersonalityEvd. As shown in Table 3, except for HC-GNN, which performs poorly with accuracies all around 0.5, most models achieve the highest accuracy on Openness, followed by Conscientiousness, while the remaining traits show noticeably lower performance. The standard deviation of accuracy across the five traits exceeds 7.5pp for most models, with only Claude and CoPE achieving lower values of 5.39pp and 5.18pp, respectively. This imbalance arises because the dataset is derived from scripted dialogues, where OPN and CON are often consistently emphasized to shape character identities. However, other traits fluctuate with situational factors, leading to ambiguous decision boundaries and frequent misclassification due to LLM overconfidence. By leveraging the Generative–Discriminative Feedback mechanism, the proposed method effectively mitigates overconfidence-induced errors, reducing the accuracy variance to 4.72pp, and leading to reductions of 43.00% and 50.37% compared to using the original LLMs alone.

6.4 Ablation Study

To demonstrate the effectiveness of different components in the EPIR framework, we ablate the bidirectional iterative process and the Generative-Discriminative Feedback mechanism on Personali-



(a) Ablation analysis of personality prediction.



(b) Ablation analysis of evidence identification.

Figure 3: Ablation Results. (a) and (b) show results for personality prediction and evidence utterance identification, respectively.

tyEvd.

Figure 3a presents the ablation results for different EPIR components, where "w/o Iteration" and "w/o GDF" denote the removal of the bidirectional iterative process and the Generative-Discriminative Feedback mechanism, respectively. Compared to the LLM baseline, introducing either component individually brings performance improvements, indicating that both the bidirectional iterative process and the GDF mechanism contribute positively to personality prediction. Removing the bidirectional iterative process and the GDF mechanism leads to average accuracy decreases of 3.02% and 3.77%, respectively, with cross-personality dimension stability declining by 37.47% and 95.58%. This reveals the importance of both components to the proposed method and further validates the contribution of the GDF mechanism to cross-dimension performance stability. The complete model achieves the best results, indicating synergistic gains between the two components.

Figure 3b compares evidence utterance identification performance between using only the graph model and receiving iterative feedback signals from different LLMs, validating the feedback role of personality prediction results on evidence utterance identification within the EPIR framework.

Results show that feedback signals from the personality prediction module can effectively improve evidence utterance identification performance. Compared to using only the graph model, iterative feedback from GPT and DeepSeek personality prediction results achieves performance improvements of 2.23% and 2.96%, respectively, validating the promoting effect of personality prediction results on evidence utterance refinement.

7 Conclusion

In this paper, we first reframe personality recognition as an Evidence-Personality Iterative Reasoning problem, which systematically explores the interplay between evidence utterance identification and personality prediction. The experimental results show that the graph model can more accurately locate the evidence utterances and the bidirectional iterative mechanism contributes to improved personality prediction. The full model achieves accuracies of 67.46% on PersonalityEvd and 57.86% on PANDORA, demonstrating the effectiveness of the proposed method.

Limitations

The proposed method achieves iterative reasoning between the graph module and LLM module by leveraging their respective classification outputs. Although its effectiveness has been validated across multiple experiments, several limitations still remain. First, the current framework integrates the predictions of each module in a unified manner. More advanced encoder-decoder or mixture-of-expert structure could be explored to further enhance module interaction. In addition, to maintain a clear focus on the core research questions of the effectiveness of the proposed framework, this work does not conduct an exhaustive exploration of all possible LLM variants or fine-tuning strategies, although these may lead to further performance gains. These directions will be explored in future work.

Ethical Statement

Personality recognition can enable sensitive profiling and may be misused for surveillance, manipulative targeting, or discriminatory decision-making (e.g., hiring, insurance, or credit scoring). Errors in trait inference may also lead to unfair treatment or psychological harm, particularly in high-stakes contexts. Our experiments are conducted on two research datasets, PersonalityEvd, derived from CPED, and PANDORA, based on Reddit data. Models trained on these datasets may be affected by biases arising from dataset construction, platform characteristics, linguistic differences, and annotation procedures, and may not generalize well across populations, domains, or languages. We therefore recommend restricting use to research and other low-stakes settings with meaningful human oversight, avoiding integration into high-impact deci-

sion pipelines, and conducting additional evaluations of fairness, robustness, and privacy risks prior to any real-world deployment.

Acknowledgments

This work was partially supported by the National Natural Science Foundation of China under Grant No. 72293575 and 62372454.

References

AI Anthropic. 2025. Claude 4 sonnet.

Joshua Evan Arijanto, Steven Geraldy, Cyrena Tania, and Derwin Suhartono. 2021. Personality prediction based on text analytics using bidirectional encoder representations from transformers from english twitter dataset. *International Journal of Fuzzy Logic and Intelligent Systems*, 21(3):310–316.

Christiane Attig, Daniel Wessel, and Thomas Franke. 2017. Assessing personality differences in human-technology interaction: An overview of key self-report scales to predict successful interaction. pages 19–29.

Weihong Bi, Feifei Kou, Lei Shi, Yawen Li, Haisheng Li, Jinpeng Chen, and Mingying Xu. 2025. Leveraging the dual capabilities of llm: Llm-enhanced text mapping model for personality detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(22):23487–23495.

Egon Brunswik. 1952. *The Conceptual Framework of Psychology*, volume 1 of *International Encyclopedia of Unified Science*. University of Chicago Press, Chicago.

Yirong Chen, Weiquan Fan, Xiaofen Xing, Jianxin Pang, Minlie Huang, Wenjing Han, Qianfeng Tie, and Xiangmin Xu. 2022. Cped: A large-scale chinese personalized and emotional dialogue dataset for conversational ai. *arXiv preprint arXiv:2205.14727*.

DeepSeek-AI, Aixin Liu, Aoxue Mei, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenhao Xu, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, and 245 others. 2025. Deepseek-v3.2: Pushing the frontier of open large language models. *Preprint*, arXiv:2512.02556.

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

Yahui Fu, Haiyue Song, Tianyu Zhao, and Tatsuya Kawahara. 2024. Enhancing personality recognition in dialogue by data augmentation and heterogeneous conversational graph networks. *ArXiv*, abs/2401.05871.

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

Matej Gjurković, Vanja Mladen Karan, Iva Vukojević, Mihaela Bošnjak, and Jan Snajder. 2021. 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. Association for Computational Linguistics.

Wenjing Han, Yirong Chen, Xiaofen Xing, Guohua Zhou, and Xiangmin Xu. 2023. Speaker-aware hierarchical transformer for personality recognition in multiparty dialogues. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Haijun He, Bobo Li, Yiyun Xiong, Li Zheng, Kang He, Fei Li, and Donghong Ji. 2025. Heuristic personality recognition based on fusing multiple conversations and utterance-level affection. *Information Processing Management*, 62(1):103931.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Linmei Hu, Hongyu He, Duokang Wang, Ziwang Zhao, Yingxia Shao, and Liqiang Nie. 2024. Llm vs small model? large language model based text augmentation enhanced personality detection model. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'24/IAAI'24/EAAI'24. AAAI Press.

He Jun, Liu Peng, Jiang Changhui, Liu Pengzheng, Wu Shenke, and Zhong Kejia. 2021. Personality classification based on bert model. In *2021 IEEE International Conference on Emergency Science and Information Technology (ICESIT)*, pages 150–152.

Onno Kampman, Elham J. Barezi, Dario Bertero, and Pascale Fung. 2018. Investigating audio, video, and text fusion methods for end-to-end automatic personality prediction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 606–611, Melbourne, Australia. Association for Computational Linguistics.

Thomas N Kipf and Max Welling. 2017. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations*.

Simone Leonardi, Diego Monti, Giuseppe Rizzo, and Maurizio Morisio. 2020. Multilingual transformer-based personality traits estimation. *Information*, 11(4).

Tian Ma, Kaiyu Feng, Yu Rong, and Kangfei Zhao. 2025. From post to personality: Harnessing llms for mbti prediction in social media. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, CIKM '25, page 5011–5015, New York, NY, USA. Association for Computing Machinery.

OpenAI. 2025. Gpt-4-turbo-2024-04-09.

Donald Redelmeier, Umberin Najeeb, and Edward Etchells. 2021. Understanding patient personality in medical care: Five-factor model. *Journal of General Internal Medicine*, 36.

Lingzhi Shen, Xiaohao Cai, Yunfei Long, Imran Razzak, Guanming Chen, and Shoaib Jameel. 2025a. Emoperso: Enhancing personality detection with self-supervised emotion-aware modelling. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, CIKM '25, page 2577–2587, New York, NY, USA. Association for Computing Machinery.

Lingzhi Shen, Yunfei Long, Xiaohao Cai, Guanming Chen, Imran Razzak, and Shoaib Jameel. 2025b. Less but better: Parameter-efficient fine-tuning of large language models for personality detection. In *2025 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8.

Yuxuan Song, Qiudan Li, Yilin Wu, David Jingjun Xu, and Daniel Dajun Zeng. 2025. Knowledge-enhanced hierarchical heterogeneous graph for personality identification with limited training data. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(2):1529–1537.

Christopher J Soto and Oliver P John. 2017. The next big five inventory (bfi-2): Developing and assessing a hierarchical model with 15 facets to enhance bandwidth, fidelity, and predictive power. *Journal of personality and social psychology*, 113(1):117.

Lei Sun, Jinming Zhao, and Qin Jin. 2024. Revealing personality traits: A new benchmark dataset for explainable personality recognition on dialogues. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 19988–20002, Miami, Florida, USA. Association for Computational Linguistics.

Tommy Tandera, Derwin Suhartono, Rini Wongso, and Yen Lina Prasetio. 2017. Personality prediction system from facebook users. *Procedia Comput. Sci.*, 116(C):604–611.

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

Qwen Team. 2025. Qwen3: Think deeper, act faster.

Liping Wang, Fenyu Hu, Shu Wu, and Liang Wang. 2021. Label-informed graph structure learning for node classification. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, CIKM '21, page 3488–3492, New York, NY, USA. Association for Computing Machinery.

Yuwei Xia, Ding Wang, Qiang Liu, Liang Wang, Shu Wu, and Xiao-Yu Zhang. 2024. Chain-of-history reasoning for temporal knowledge graph forecasting. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 16144–16159, Bangkok, Thailand. Association for Computational Linguistics.

Mengyu Xiang, Yuxuan Song, Qiudan Li, Shu Wu, and Daniel Dajun Zeng. 2025. Dynamic detection of sarcasm topic-target pairs via llm-based knowledge alignment. In *Companion Proceedings of the ACM on Web Conference 2025*, WWW '25, page 1422–1425, New York, NY, USA. Association for Computing Machinery.

Yunhe Xie, Rui Mao, Wei Li, Atika Qazi, and Erik Cambria. 2025. A discourse structure- and interlocutor-guided network for dialogue act recognition and sentiment classification. *IEEE Transactions on Affective Computing*, 16(4):2918–2930.

Di Xue, Lifa Wu, Zheng Hong, Shize Guo, Liang Gao, Zhiyong Wu, Xiaofeng Zhong, and Jianshan Sun. 2018. Deep learning-based personality recognition from text posts of online social networks. *Applied Intelligence*, 48(11):4232–4246.

Tao Yang, Jinghao Deng, Xiaojun Quan, and Qifan Wang. 2023a. Orders are unwanted: dynamic deep graph convolutional network for personality detection. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'23/IAAI'23/EAAI'23. AAAI Press.

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

Tao Yang, Feifan Yang, Haolan Ouyang, and Xiaojun Quan. 2021. 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. Association for Computational Linguistics.

Tal Yarkoni. 2010. Personality in 100,000 words: A large-scale analysis of personality and word use among bloggers. *Journal of Research in Personality*, 44(3):363–373.

Haein Yeo, Taehyeong Noh, Seungwan Jin, and Kyungsik Han. 2025. PADO: Personality-induced multi-agents for detecting OCEAN in human-generated texts. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 5719–5736, Abu Dhabi, UAE. Association for Computational Linguistics.

Bo Zhang, Yi Ming Li, Jian Li, Jing Luo, Yonghao Ye, Lu Yin, Zhuosheng Chen, Christopher J Soto, and Oliver P John. 2022. The big five inventory–2 in china: A comprehensive psychometric evaluation in four diverse samples. *Assessment*, 29(6):1262–1284.

Yin Zhang, Rong Jin, and Zhi-Hua Zhou. 2010. Understanding bag-of-words model: a statistical framework. *International journal of machine learning and cybernetics*, 1(1):43–52.

Haohao Zhu, Xiaokun Zhang, Junyu Lu, Youlin Wu, Zewen Bai, Changrong Min, Liang Yang, Bo Xu, Dongyu Zhang, and Hongfei Lin. 2025. Enhancing textual personality detection toward social media: Integrating long-term and short-term perspectives. *IEEE Transactions on Audio, Speech and Language Processing*, 33:4563–4573.

Yangfu Zhu, Yue Xia, Meiling Li, Tingting Zhang, and Bin Wu. 2024. Data augmented graph neural networks for personality detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(1):664–672.

Appendix

A Experiment Setup Details

A.1 Baseline Details

In this section, we introduce the details of the baselines.

A.1.1 Content-based Methods

- TextCNN (Kampman et al., 2018): This method adopts a tri-modal CNN that fuses audio, visual, and textual features from video clips for personality prediction. We use its text-channel CNN as a baseline.

- BERT (Devlin et al., 2019): This Method uses a pretrained BERT-base-chinese for PersonalityEvd and bert-base-uncased for PANDORA to extract global semantic features from dialogues and then applies a two-layer MLP to output personality predictions.

A.1.2 Relation-based Methods

- HC-GNN (Fu et al., 2024): This method proposes a graph to model speaker relations and enhances diversity via data interpolation, then applies multi-layer attention to fuse heterogeneous features for personality prediction.

- DSIG (Xie et al., 2025): This method constructs a heterogeneous graph over utterances, dialogue, and speakers, and improves structural awareness via a dynamic graph and pretraining on dialogue discourse parsing.

- SAH-GCN (He et al., 2025): This method builds an utterance-level heterogeneous graph focusing on target speaker, and adopts emotion-enhanced contrastive learning to mitigate semantic similarity issues.

- KEHG (Song et al., 2025): This method constructs a graph over posts, words, and multiple external knowledge, and models associations via global multi-view graph encoding to obtain aggregated personality trait predictions.

- DEN (Zhu et al., 2025): This method models long-term traits using a LIWC-based psychological graph and short-term states from individual posts, then fuses them through bidirectional graph interaction.

- Semi-PerGCN (Zhu et al., 2024): This method constructs heterogeneous personality graphs and combines LIWC-based multi-view graph augmentation with graph consistency learning for semi-supervised personality detection.

A.1.3 LLM-based Methods

We evaluate several mainstream LLMs, including GPT-4-Turbo (OpenAI, 2025), Claude-4-Sonnet (Anthropic, 2025), and DeepSeek-V3.2 (DeepSeek-AI et al., 2025), under a zero-shot inference setting to assess their performance on the task of dialogue-based personality prediction. As a competitive fine-tuning-based approach, we also include Qwen3-8B (Team, 2025), which is parameter-efficiently fine-tuned with LoRA (Hu et al., 2022). We further consider two recent LLM-enhanced personality detection methods. ETM (Bi et al., 2025) enhances user representations with LLM embeddings and aligns them with multi-perspective personality-label interpretations through contrastive learning. EmoPerso (Shen et al., 2025a) combines LLM-based data augmentation with joint personality-emotion modeling, cross-attention, and self-taught reasoning. Finally, we adopt CoPE (Sun et al., 2024), a baseline tailored to conversational settings that constructs chain-of-thought data and combines it with fine-tuning for dialogue-based personality prediction.

A.2 Evaluation Metrics

We employ F1-score and accuracy as the primary metrics. Specifically, for evidence utterance identification, we report the F1-score for each Big-Five trait and the average F1-score across all five traits. For personality recognition, we report the accuracy for each trait and the average accuracy across the five traits. Additionally, to further assess performance stability across personality traits, we report the standard deviation (std.dev) over the five traits.

A.3 Implementation Details

The model is implemented using PyTorch and trained for 50 epochs by AdamW with an initial learning rate of 2e-5 for BERT and 1e-4 for the GCN and a dropout ratio of 0.5. We use BERT-base-chinese for PersonalityEvd and BERT-base-uncased for PANDORA to generate node embeddings. The hidden dimension of all GCN layers is 256. All hyperparameters are tuned on the validation set.

Each personality trait is associated with an independent evidence-selection threshold. A two-layer

Method	Personality Macro-F1 (%)						std.dev
	Avg.	OPN	CON	EXT	AGR	NEU	(percent point, pp)
PersonalityEvd							
TextCNN	31.57	27.65	27.98	32.96	37.62	31.61	3.65
BERT	33.51	35.15	30.69	31.38	38.42	31.89	2.90
HC-GNN	23.45	27.51	27.70	24.21	15.07	22.75	4.60
DSIG	39.31	41.21	37.28	34.23	43.45	40.36	3.22
SAH-GCN	32.39	30.23	30.12	37.00	30.31	34.30	2.79
KEHG	32.90	27.51	29.14	39.23	33.81	34.79	4.19
DEN	41.08	41.98	39.28	42.44	45.04	36.66	2.87
GPT	49.87	54.85	49.50	38.95	48.57	57.51	6.39
Claude	52.39	53.15	56.76	42.02	54.48	55.53	5.32
DeepSeek	52.34	52.55	52.45	46.88	59.41	50.41	4.09
Qwen	53.02	**57.16**	45.85	46.99	59.28	55.82	5.51
ETM	41.28	37.50	38.06	37.26	48.30	45.27	4.60
EmoPerso	53.89	52.79	**57.26**	**52.93**	59.54	46.91	4.34
EPIR (GPT)	53.88	53.83	48.84	51.22	57.99	57.51	3.54
EPIR (DeepSeek)	**54.16**	45.90	49.78	49.67	**67.77**	**57.67**	7.81
PANDORA							
TextCNN	46.61	47.86	47.98	50.26	41.69	45.26	2.93
BERT	51.64	50.46	53.07	52.33	53.54	48.82	1.76
Semi-PerGCN	42.65	44.38	36.51	47.07	44.96	40.36	3.76
KEHG	37.80	36.24	34.03	44.91	37.06	36.79	3.71
DEN	50.92	52.23	50.36	49.43	51.03	51.57	0.97
DeepSeek	52.40	39.13	52.31	58.16	**57.17**	55.21	6.93
Qwen	52.37	52.20	54.06	50.93	52.20	52.48	1.00
ETM	52.48	**52.93**	48.29	53.86	52.04	55.28	2.35
EmoPerso	53.40	52.43	**56.40**	55.68	53.39	49.11	2.59
EPIR (DeepSeek)	**54.41**	45.62	51.17	**59.86**	56.54	**58.88**	5.33

Table 4: Overall Personality Prediction Results of Macro-F1.

Trait	Ours Mean $\pm$ SD	95% CI	Strongest Baseline	Δ	p	Significance
OPN.	74.93 ± 0.52	$[74.29, 75.57]$	75.48	-0.55	0.0379	*
CON.	68.26 ± 0.45	$[67.70, 68.83]$	68.87	-0.61	0.0201	*
EXT.	65.45 ± 0.60	$[64.71, 66.20]$	63.39	$+2.06$	0.0008	***
AGR.	68.60 ± 1.23	$[67.07, 70.12]$	63.36	$+5.24$	0.0003	***
NEU.	60.06 ± 0.00	$[60.06, 60.06]$	61.15	-1.09	–	
Avg.	$\mathbf{67.46 \pm 0.31}$	$[\mathbf{67.07}, \mathbf{67.85}]$	**66.45**	$\mathbf{+1.01}$	**0.0010**	***

Table 5: Statistical significance analysis over five runs against the strongest baseline.

MLP learns the mapping from personality predictions to the corresponding trait-specific evidence-selection thresholds. During training, ground-truth personality labels are provided through teacher forcing to guide the learning of threshold adjustments. During iterative inference, the ground-truth labels are replaced by the personality predictions obtained from the preceding iteration, allowing the evidence-selection thresholds and evidence sets to be dynamically refined.

The personality recognition module is implemented via API-based LLM inference with a frozen LLM. In practice, generation for different traits and levels is executed via parallel and incremental API calls to reduce latency. During EPIR inference, the iterative process stops when both the dialogue-level personality prediction and the evidence set remain unchanged between two consecutive iterations or when the maximum number of four iterations is reached. For LLM inference, we set the temperature to 0.7 for the Generator, and 0.0 for the Discriminator.

B Supplementary Experiment

B.1 Performance of Macro-F1

we report the Macro-F1 results in Table 4. EPIR (DeepSeek) achieves the highest average Macro-F1, 54.16% on PersonalityEvd and 54.41% on PANDORA, outperforming the strongest baseline with relative improvement ratios of 0.50% and 1.89%, respectively. The consistent improvements in both accuracy and Macro-F1 further demonstrate the effectiveness and robustness of the proposed method for personality recognition.

B.2 Statistical Significance Analysis

To evaluate the robustness of the observed performance gains, we report 95% confidence intervals

Model	Task	Parameter Count	Computational Cost	End-to-end Delay
BERT	Per	102.28M	386.40 GFLOPs	5.06 ms
DeepSeek	Per	685B	–	11.05 s
DeepSeek	Evi	685B	–	6.19 s
Graph	Evi	104M	386.41 GFLOPs	6.73 ms
EPIR (DeepSeek)	Per + Evi	685B + 104M	386.41 GFLOPs + LLM API	17.32 s

Table 6: Computational cost and latency comparison. In this table, Per and Evi denote Personality Prediction and Evidence Utterance Identification

Statistic	Value
# Speakers	72
# Dialogues	1,924
# Utterances	32,673
Avg. dlg per spk	26.72
Avg. utt per dlg	16.98
Avg. length of utt	16.73
Avg. tgt spk utt per dlg	8.43
Avg. evi utt IDs per dlg	4.26
Train speakers	51
Validation speakers	7
Test speakers	14
Train dialogues	1,380
Validation dialogues	181
Test dialogues	363

Table 7: Dataset statistics of PersonalityEvd. In this table, spk, dlg, and utt stand for speaker, dialogue, and utterance, respectively; tgt and evi denote target and evidence, respectively.

Statistic	Value
# Users	1,355
# Posts	23,035
Avg. posts per usr	17.00
Avg. length of post	39.34
Avg. evi posts per usr–trait	4.25
Training users	948
Validation users	136
Test users	271

Table 8: Dataset statistics of PANDORA. In this table, usr and evi denote user and evidence, respectively.

Interval	Dialogue Count	Proportion
$[6, 10]$	137	7.10%
$[11, 15]$	775	40.30%
$[16, 20]$	528	27.40%
$[21, 25]$	315	16.40%
$[26, 30]$	160	8.30%
$[31, 43]$	9	0.50%

Table 9: Distribution of dialogue lengths in PersonalityEvd.

(CIs) and statistical significance tests over five runs against the strongest baseline on PersonalityEvd. The proposed framework achieves an average accuracy of 67.46 ± 0.31 (mean $\pm$ standard deviation), with a 95% CI of $[67.07, 67.85]$, while the strongest baseline achieves an average accuracy of 66.45. Notably, the lower bound of the confidence interval already exceeds the baseline result. A one-sample t-test yields $t(4) = 7.21$ and $p = 0.0010$, and all five runs consistently outperform the baseline. Detailed results are presented in Table 5.

The approximately 1.0-point gain in average accuracy exhibits a clear directional pattern. At the level of individual personality traits, the proposed framework achieves statistically significant improvements on Extraversion ($+2.06$, $p < 0.001$) and Agreeableness ($+5.24$, $p < 0.001$), while showing slight decreases on Openness, Conscientiousness, and Neuroticism ($\Delta \leq 1.1$), all within approximately one standard deviation. Specifically, substantial and statistically significant gains on two traits, together with only minor variations on the remaining three, indicate that the overall performance improvement is not attributable to noise.

C Computational Cost and Latency

We report the parameter count, computational cost, and latency per sample of the proposed framework. The graph-based evidence mining module is lightweight and adds only marginal computation beyond the base LLM call. In comparison, using DeepSeek for evidence utterance identification introduces an average latency of 6.19 s per dialogue, whereas our graph-based module adds only 6.73 ms while delivering higher accuracy. Furthermore, the extra latency introduced by EPIR is only 6.27s per dialogue, which is approximately the same as the cost of performing evidence utterance identification through one additional LLM iteration. The generation and discrimination processes are implemented with parallel and incremental LLM calls to further reduce computational overhead. Overall, the average number of iterations per sample is 2.13, and the average inference time is 17.32s per dialogue on a single NVIDIA A100.

Trait	Dialogue Count	# Evidence utterances			Coverage (%)	
		Mean	Median	Max	Mean	Median
OPN	440	3.6	3	17	21.30	18.50
CON	616	3.9	3	18	23.30	20.00
EXT	916	4.9	4	23	28.30	25.00
AGR	1,336	3.9	3	22	23.50	20.00
NEU	1,030	4.7	4	28	27.50	25.00

Table 10: Per-dialogue evidence utterance counts and their global proportion among all dialogue utterances.

Location Interval	OPN	CON	EXT	AGR	NEU
$[0.0, 0.2)$	16.20	14.10	18.90	16.60	14.90
$[0.2, 0.4)$	18.50	17.60	18.90	17.80	18.00
$[0.4, 0.6)$	22.20	20.40	20.20	19.80	20.70
$[0.6, 0.8)$	21.80	22.10	20.60	20.90	22.20
$[0.8, 1.0]$	21.20	25.70	21.30	24.90	24.20

Table 11: Location distribution of Evidence utterances in the Dialogue.

D Dataset Details

D.1 Annotation Procedure of PersonalityEvd

PersonalityEvd (Sun et al., 2024) is a dialogue-based explainable personality recognition benchmark built on CPED, providing both dialogue-level personality labels and corresponding evidence utterance annotations. The evidence annotations were obtained through a rigorous multi-stage manual annotation and verification procedure involving psychology-trained annotators and multiple rounds of expert review, with additional quality assessments confirming their reliability and interpretability. Given this careful annotation and quality-control process, we directly adopt the released evidence annotations to preserve the original benchmark setting without introducing additional annotation variance. Detailed annotation procedures and evaluation results are provided in the original paper (Sun et al., 2024).

D.2 Dataset Statistics

We provide detailed statistics of PersonalityEvd in Table 7. The dataset is split according to speakers, ensuring that no speaker appears in more than one data split.

We further provide detailed statistics of PANDORA in Table 8

D.3 Distribution of Dialogue and Evidence Utterances

We further report the distribution of dialogue lengths and trait-specific evidence utterances.

Table 9 presents the distribution of utterance counts within each dialogue, most dialogues contain between 11 and 20 utterances

Table 10 presents the per-dialogue distribution of evidence utterance counts, along with the overall proportion of evidence utterances across the entire dialogue set.

Table 11 reports the location distribution of evidence utterances within dialogues.

D.4 Impact of Dialogue Length

We further analyze the relationship between dialogue length and performance on evidence utterance identification and personality recognition. The correlation results are reported in Table 12.

Evidence utterance identification shows a weak negative correlation with dialogue length, whereas personality recognition remains relatively stable across dialogues of different lengths. Overall, dialogue length has a limited effect on the performance of EPIR.

D.5 PANDORA Preprocessing and Annotation

For our experiments, we preprocess PANDORA and construct post-level evidence annotations to support evidence-aware personality recognition. We first retain 1,568 users with complete Big Five annotations. We then remove duplicate, empty and non-English posts, retaining posts containing no fewer than six and no more than 200 words. Since PANDORA contains a large number of posts for each user, we retain users with at least 17 posts and randomly sample 17 posts per user. This number is close to the average dialogue length of PersonalityEvd, thereby maintaining a comparable input scale across the two datasets. This process results in 1,355 users and 23,035 posts. Finally, the median score of the retained users is used to divide

Task	Pearson		Spearman	
	r	p	ρ	p
Evidence utterance identification	-0.12	0.022	-0.04	0.400
Personality recognition	0.07	0.170	0.07	0.185

Table 12: Correlation between dialogue length and task performance.

Method	Avg.	OPN	CON	EXT	AGR	NEU	std.dev
GPT + CoT + SC	57.85	71.35	63.09	50.69	55.92	48.21	8.46
GPT + Plan	60.33	72.18	62.53	46.28	60.33	**60.33**	8.28
DeepSeek + CoT + SC	59.83	72.18	68.32	49.31	61.43	47.93	9.79
DeepSeek + Plan	63.52	**79.83**	67.31	57.22	62.01	51.25	9.73
EPIR (GPT)	63.69	71.35	63.09	60.88	65.79	57.35	4.72
EPIR (DeepSeek)	**67.46**	74.93	**68.26**	**65.45**	**68.60**	60.06	4.83

Table 13: Comparison with alternative prompting strategies on PersonalityEvd. CoT, SC, and Plan denote chain-of-thought reasoning, self-checking, and task planning, respectively.

Iteration	Personality accuracy (%)
0	62.75
1	65.40
2	67.16
3	67.44
4	67.44

Table 14: Personality recognition performance from single inference across iterations.

each personality trait into High and Low levels.

For evidence annotation, a psycholinguistic grounding was introduced following Yarkoni (Yarkoni, 2010), combined with the LIWC dictionary. Correlation coefficients between Big Five traits and LIWC categories were used to compute evidence scores for each comment. Specifically, LIWC category proportions were first extracted for each comment and then standardized using z-score normalization. The evidence score for trait t is computed as:

$$signed_{c,t} = \sum_{k} r_{t,k} \cdot z_{c,k}, \tag{10}$$

where $r_{t,k}$ denotes the correlation with trait t and LIWC category k. The scores are further normalized to $[0, 1]$, and comments above the median score for each trait are labeled as evidence comments. Evidence posts are further required to be polarity-consistent with the corresponding user-level personality labels.

E Alternative Prompting Strategies

We further compare EPIR with alternative prompting strategies, including chain-of-thought reasoning (CoT) and self-checking (SC). The existing prompting strategy based on task planning is denoted as Plan. The results on PersonalityEvd are reported in Table 13.

Among alternative prompting strategies, EPIR achieves higher average accuracy and higher cross-trait stability with both GPT and DeepSeek. These results support that the performance gains mainly arise from the proposed iterative framework rather than from prompt complexity.

F Analysis of Iterative Performance

We further report personality recognition performance across different iterations. The results from single inference are presented in Table 14.

The results indicate that most performance gains are obtained in the first two iterations, with each iteration generally improving over the previous one in terms of overall personality recognition performance. During the iterative process, 252 samples experienced changes in predicted personality labels, among which 150 samples showed performance improvements, 53 showed degradation, and 49 remained unchanged.

The degraded cases are mainly concentrated in Openness. Since this trait is highly dependent on event context and interlocutor intent, irrelevant information in the initial evidence set may introduce local bias during iterative refinement. Nevertheless, the consistent improvement in overall accuracy indicates that the iterative process improves performance for most affected samples.

G Cross-Model Discriminator Analysis

We further evaluate GDF using different discriminator models. GPT-4-Turbo is adopted as an independent discriminator to evaluate candidates generated by DeepSeek, and is compared with the original DeepSeek self-discrimination setting. The results

(a) Overall performance			
Metric	**DeepSeek self**	**GPT judge**	**Δ (GPT−DeepSeek)**
Accuracy	0.4702	0.4654	−0.005
Macro-F1	0.4585	0.4351	−0.023

(b) Trait-level F1 scores				
Trait	**DeepSeek self**	**GPT judge**	**ΔF1**	**Holm p**
Openness	0.379	0.377	−0.002	0.162
Conscientiousness	0.394	0.409	+0.015	3×10^{-5}
Extraversion	0.363	0.346	−0.017	0.125
Agreeableness	0.516	0.450	−0.066	1×10^{-14}
Neuroticism	0.453	0.463	+0.009	8×10^{-4}

Table 15: Performance with different discriminators. (a) reports overall performance, and (b) reports trait-level F1 scores with Holm-corrected p-values.

ID	Speaker	Utterance	Role in the case
U1	Speaker A	Dear Miaomiao, you have to take me to the radio station to visit him.	–
U2	Speaker B	Why do you watch him broadcast?	–
U3	Speaker A	I have to declare my ownership.	–
U4	Speaker B	We have rules at our radio station.	Initial rule enforcement
U5	Speaker B	No one else is allowed to come in.	Initial rule enforcement
U6	**Speaker B**	**Bringing someone in with you is fine, too.**	**Core evidence**
U7	**Speaker B**	**But we need to lay down a few ground rules.**	**Core evidence**
U8	**Speaker B**	**You can only watch for 5 minutes. And look, but don't touch.**	**Core evidence**

Table 16: Case study of iterative evidence refinement for Agreeableness recognition. Bold utterances denote the final evidence selected by EPIR.

are reported in Table 15.

The two discriminator settings achieve comparable overall accuracy, with no significant difference under the McNemar test ($p = 0.41$) and a Cohen's κ of 0.52. At the trait level, neither discriminator consistently outperforms the other. The results indicate that GDF maintains generally consistent discriminative behavior across different LLM families.

H Case Study

We further add case study as shown in Table 16. The case analysis further reveals how the proposed method works. The example illustrates a scenario related to Agreeableness recognition, where speaker A requests access to a broadcasting station, and the target speaker B, as an employee of the broadcasting station, needs to respond to this request involving access to a restricted area. Faced with diverse emotional signals and complex contextual information in the dialogue, the graph-based evidence utterance identification module, which associates utterances with personality traits, initially produces a set of redundant and noisy candidate evidence utterances. In this case, all last six utterances are treated as evidence utterances, including explicit rejections of the entry request and rule enforcement-related utterances. This leads to misinterpretation by the model, resulting in the initial personality recognition output incorrectly classifying the target speaker as having low Agreeableness.

The proposed Evidence-Personality Iterative Reasoning framework (EPIR) improves the accuracy of evidence utterance identification, guiding the graph model to focus on core evidence utterances (U6, U7, U8) that truly reflect the target trait. Although the speaker initially rejects the entry request to the broadcasting station, as friends, the speaker later agrees to relax the restriction in utterances 6, 7, and 8, allowing speaker A to temporarily enter the broadcasting station. Meanwhile, with the help of the proposed Generative-Discriminative Feedback (GDF) mechanism, the model further verifies that the target speaker's exceptional behavior reflects high empathy and flexibility. This ultimately corrects the personality recognition result to high Agreeableness.

I Prompt Template

Generator Prompt

Role: You are a dialogue generation expert and need to simulate and generate the next line of dialogue based on the context.

Task Background: In the Big Five personality traits assessment system, {trait_desc}
Please strictly role-play as a speaker with **{level} {trait_name}**, and assume the persona of {character_name}.
The next sentence is considered evidence of the low {trait_name} trait of "{speaker}". Please generate a dialogue that conforms to the expression of this personality trait based on the context.

Task Description: Based on the following dialogue context, simulate and generate a short sentence that "{speaker}" might say next, ensuring it remains consistent with the context. The generated sentence must:
1. Be consistent with the dialogue context, fitting the dialogue scene and character relationships.
2. Consistent with "{speaker}'s" speaking style and tone in the dialogue.
3. Reflect the personality traits of "{trait_name}".
4. Be logically coherent and naturally fluent.
5. Generate only one sentence; do not generate multiple sentences.

Dialogue Context: {context_text}

Please generate the next sentence spoken by "{speaker}". Output the generated content directly.

Discriminator Prompt

###Role: You are a dialogue analysis expert. Based on the context of a dialogue, you need to determine which statement best reflects the speaker's personality traits, first assigning a consistency score (GPTScore), and making a final selection.

Input
Dialogue Context: {context_text}; Original Evidence Sentence: The original statement spoken by {speaker} is: {original_content}; Candidate Sentences: A (High {trait_name}): {generated_high}; B (Low {trait_name}): {generated_low}; C (Neutral): {generated_neutral}

Judging Task (Must be performed step-by-step)
Step 1: Calculate GPTScore (0-100) You need to give scores to A/B/C respectively. Calculate its consistency score (GPTScore) with the original evidence sentence in the current dialogue context.
The GPTScore represents whether the candidate sentence reproduces the expression of the original sentence in this context + the true intent + the strength level of {trait_name}.

Scoring Dimensions (Total score 100, must be summed according to the following dimensions):
1) Similarity of Expression (0-20): Whether the tone, wording, politeness/aggression, etc., are close to the original sentence.
2) Consistency of {trait_name} Level (0-40): Whether the intensity of {trait_name} reflected in the candidate sentence is consistent with the original sentence (not consistent with labels A/B/C, but consistent with the "level inferred from the original sentence").
3) Consistency of Semantics and Intent (0-40): In this context, whether the candidate sentence expresses the same appeal/purpose/stance as the original sentence.
> Note: Scoring must be "relatively compared," A/B/C There should be a reasonable gap between the three scores; avoid assigning very similar scores to all three.

Step 2: Output the final decision based on the GPTScore
- Select the option with the highest GPTScore as the final decision (A/B/C).
- If the highest scores are tied or the difference is ≤ 2 points:
1) Prioritize the option with the higher consistency (0-40) in the "{trait_name} level" dimension;
2) If still indistinguishable, choose C (neutral/not obvious) as a conservative decision.

Output format (must be strictly followed; do not output any extra content)
The first part outputs the GPTScore, and the second part outputs the final selection letter:
GPTScore: A: <0-100 integer>; B: <0-100 integer>; C: <0-100 integer>
Decision: <A/B/C>

Initial Prompt

Based on the following {{dialogue content}}, strictly follow the steps below to analyze the personality trait dimensions revealed in {character_name}'s speech, and determine the value of the "level" attribute for this dimension as "high"/"low"/"cannot judge" along with the corresponding supporting evidence.
Personality Analysis Step One: Contextual Understanding
Core Objective: Understand the background information of the current dialogue, focus on {character_name}'s words and actions, and analyze and collect evidence that may be used to understand personality traits. Ensure that subsequent personality trait judgments are based on a deep understanding of the dialogue context, character motivations, and emotional state, rather than taking things out of context. Note that you should record and save the {{contextual understanding}} content, as it will be used as a basis for subsequent judgments and included in the final output.
Please strictly follow these sub-steps to complete the contextual understanding:
Summarize the core issue, focusing on the core content that {character_name} cares about most from the perspective of the protagonist.
Based on the core issue and context, deduce the "deeper intention/subtext" ({character_name}'s true purpose) beyond the literal meaning.
Identify the identities of the speakers (e.g., children, friends, etc.) and determine the nature of the relationship, the social situation, and the interaction style (active, passive, dominant, commanding, compliant, confrontational, etc.) from {character_name}'s perspective.
Capture {character_name}'s emotional state and its changes, focusing on linking emotional changes to the core issue and interaction pattern to form a comprehensive understanding of the current dialogue context.
Personality Analysis Step Two: Extracting Behavioral Intentions
Core Objective: Based on the contextual understanding in Step One, extract the target character's behavioral intentions from their speech and summarize them using {{behavioral description}} and {{core intention}}.
Please strictly follow these sub-steps to complete the behavioral intent extraction:
Understand the content in context and find the core of the entire dialogue (event, purpose, conflict, etc.);
Describe what {character_name} did in neutral and objective language, and save it as {{behavioral description}};
From {character_name}'s perspective, analyze the subtext in the context to infer {character_name}'s purpose in saying this, and save it as {{core intent}}.
Step 3 of Personality Analysis: Big Five Personality Mapping
Core Objective: Systematically map each {{behavioral description}} and {{core intent}} obtained in Step 2 to the five dimensions of the Big Five personality traits and determine its "level" as "high" or "low".
Please strictly follow these sub-steps to complete the Big Five personality mapping:
Treat the given BFI-2 scale as a "description library." Determine which dimension and trait level the behavioral description and core intention match, and record the matching results. For each description, note the dimension and corresponding trait level.
{BFI-2 scale};
For each matching result, re-examine and judge whether the current behavioral description and core intention clearly demonstrate a match between character_name and personality trait descriptions. Note that a judgment can only be made if the description clearly matches the scale content or matches multiple descriptions under a single personality trait;
Record the judgment results. For personality dimensions where no level judgment is made, mark them as "Cannot be judged."
Step 4 of Personality Analysis: Judgment Result
The results of the Big Five personality mapping in Step 3 are organized into the following JSON format as {{judgment Result}}:
{{
"openness level": "",
"conscientiousness level": "",
"extraversion level": "",
"agreeableness level": "",
"neuroticism level": "",
}}
The final output JSON format is as follows:
{{
Contextual Understanding: {{contextual Understanding}};
Behavioral Intent: {{behavioral Description}} and {{core Intent}};
Judgment Result: {{judgment Result}}
}}
The current {{dialogue content}} is as follows:
{dialogue}