Dialogue-oriented Interpretable Personality Recognition via Evidence-Guided Bidirectional Iterative Optimization

Anonymous ACL submission

Abstract

Identifying personality from dialogue can improve interpretability and adaptability for human–computer interaction and psychological assessment. Existing research focuses on modeling emotional trajectories and interaction patterns from entire dialogue, failing to predict personality from the specific evidence, which may serve as key clues for reasoning and enabling accuarate personality prediction. How to establish an adaptive iterative mutual reinforcing mechanism between evidence and personality is a key challenge. This paper proposes a Generative–Discriminative Feedback Refinement mechanism for dialogue-based personality prediction, it constructs a hierarchical dialogue graph to jointly model the speaker's role, contextual dependencies, and heterogeneous interaction relationships for evidence utterance mining. Then, the generator simulates evidence utterances at different trait levels, and the discriminator derives a consistency-based judgement between the generated and original utterances to refine the initial LLM-based prediction. And the updated prediction is fed back to the graph model via bidirectional iterative optimization, improving interpretability and overall performance. Experimental results on public dataset demonstrate that the proposed method achieves the best performance over the state-of-the-art model.

1 Introduction

With the rapid development of online communication platforms, people communicate and interact with others through dialogue, expressing their attitudes and emotions. During the process, the emotional trajectories and interaction patterns of different participants reveal their underlying personality traits. Identifying personality from dialogue automatically can improve interpretability and adaptability for downstream applications, such as enhancing the intelligence of human–computer interaction systems and increasing the reliability of social behavior analysis and psychological state inference (Redelmeier et al., 2021; Attig et al., 2017).

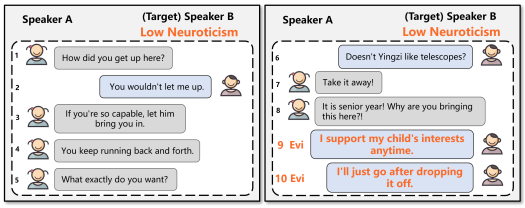

Figure 1: Example of evidence utterance mining and evidence-based reasoning for personality prediction.

Existing studies primarily focus on uncovering social media users' personality traits from their posts. Some researches adopted pre-trained language models to learn language patterns (Xue et al., 2018), emotional expressions (Ren et al., 2021), and semantic features (Devlin et al., 2019). D-DGCN (Yang et al., 2023a) and KEHG (Song et al., 2025) captured underlying structural relationships in text based on graph neural networks. Large language models (LLMs) (V Ganesan et al., 2023; Yang et al., 2023b) have demonstrated strong zero-shot capabilities and reasoning abilities. Recently, personality identification based on dialogue data has garnered increasing research attention (Bhandari et al., 2025). Heterogeneous graph networks were proposed to capture contextual influences (Fu et al., 2024) and fuse speaker features from multiple dialogues (He et al., 2025). To enhance interpretability of personality recognition, CoPE (Sun et al., 2024) first introduced evidence utterances that reflect speakers' emotional fluctuations and interaction patterns, and fine-tuned large models. However, they ignored the associations among evidence utterances, sperkers'role, scenario atmosphere and personality, which may serve as key clues for reasoning and enabling accuarate personality prediction.

For example, as shown in Figure 1, in a tense

and conflictual interaction, B, the ex-husband of A, bought a gift for their child, but A repeatedly and intensely questioned him (1, 5, 8). In the face of A's intense criticism and questioning, B remained calm and rational, clearly expressing support for his child's interests (utterances 9 and 10). These utterances reflect B's emotional stability and explain the reasons why identify him as low neuroticism.

Based on the above analysis, modelling the relationship between evidence and personality in dialogue scenario faces several key challenges: First, to achieve reliable evidence sentence mining, it is necessary to jointly model the speaker's role, contextual dependencies, and heterogeneous interaction relationships. Second, it is crucial to improve the LLMs personality inference by considering the Big Five Inventory and evidence utterances. Finally, the coupling of evidence utterance mining and personality prediction may lead to cumulative errors and mutual misguidance. Thus, how to establish an adaptive iterative mutual reinforcing mechanism is important.

To address these challenges, we propose a Generative–Discriminative Feedback Refinement mechanism for dialogue-based personality prediction, enabling bidirectional optimization between evidence utterance mining and personality inference. Specifically, a hierarchical dialogue graph is first constructed to capture connections between utterances and personality traits, thereby extracting evidence utterances that support both personality prediction and interpretability. Then, a Generative–Discriminative mechanism enhances LLM-based personality prediction by evaluating the consistency between personality level-specific generated content and the original evidence utterances. Finally, the refined prediction is fed back to update the graph model's representations, enabling iterative mutual reinforcement and improving robustness in complex dialogues. The main contributions are as follows:

- To the best of our knowledge, this paper is the first to propose an evidence-guided collaborative optimization method for dialogue-based interpretable personality prediction.
- The proposed framework utilizes a Generative-Discriminative Feedback Refinement mechanism to jointly optimize graph and LLM module using the other module's outputs, achieving interpretable personality prediction.
- Experimental results demonstrate that the proposed method achieves the best performance in personality prediction, with a 1.72% improvement over the state-of-the-art model.

2 Related Work

Traditional personality prediction methods primarily rely on explicit linguistic features, such as LIWC (Tausczik and Pennebaker, 2010) and bag-of-words models (Zhang et al., 2010). With the rapid advancement of deep learning, a range of neural architectures, including CNN (Xue et al., 2018), LSTM (Tandera et al., 2017), and Transformer (Leonardi et al., 2020), have been extensively adopted for personality prediction and have yielded obvious improvements over earlier approaches. In particular, pretrained language models such as BERT (Devlin et al., 2019) provide more effective feature extraction than traditional methods (Arijanto et al., 2021; Jun et al., 2021) and reach better performance. However, because these approaches often struggle to model complex structured relations, graph neural networks have therefore been introduced to capture the associations underlying text, achieving improved results (Song et al., 2025; Yang et al., 2023a; Yang et al., 2021). More recently, LLMs have shown strong potential for personality prediction due to their zero-shot capability and chain-of-thought reasoning (V Ganesan et al., 2023; Yang et al., 2023b). Some methods leverage LLMs to extract personality-related features from multiple perspectives (Hu et al., 2024; Yeo et al., 2025), while others fine-tune LLMs in this specific domain to improve performance (Ma et al., 2025; Shen et al., 2025; Sun et al., 2024). However, feature extraction alone may not fully capture complex personality traits, and fine-tuning LLMs is usually computationally intensive, limiting their practicality in real-world applications.

Dialogues, as a common form of daily interaction, have attracted increasing attention from researchers in personality prediction. With the release of personality conversational datasets such as CPED (Chen et al., 2022), personality recognition in conversations has emerged as a more practical and important research paradigm. Methods like HC-GNN (Fu et al., 2024) and SH-Transformer (Han et al., 2023) model inter-speaker relations and utterance-level textual context to extract diverse personality cues from dialogues, and SAH-GCN (He et al., 2025) introduces utterance-level affective information via contrastive learning to further

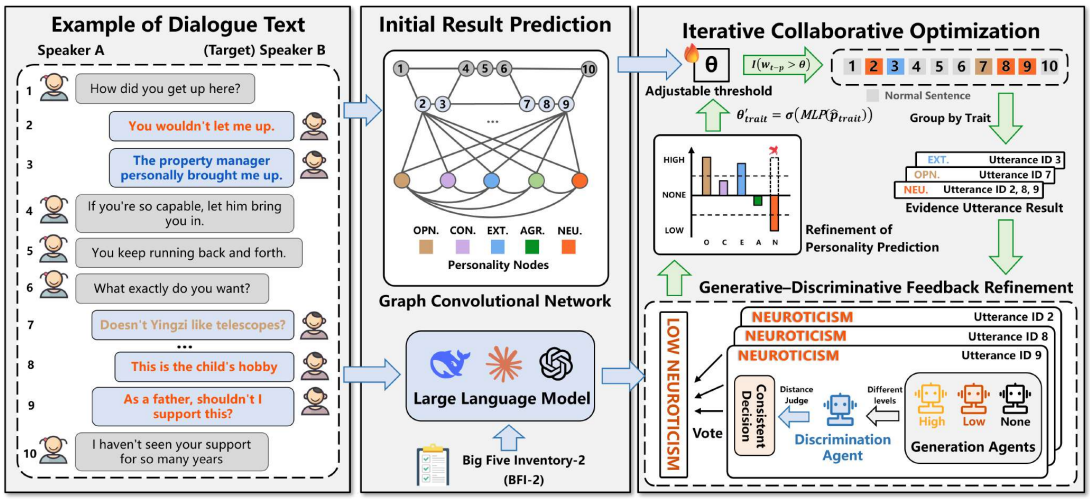

Figure 2: Overall architecture of the proposed model.

capture fine-grained semantic differences. More recent work, such as CoPE (Sun et al., 2024), constructs new inference dataset and improves LLMs on dialogue-based personality recognition through fine-tuning. However, most existing methods adopt an end-to-end paradigm and overlook intermediate evidence in the prediction process, which limits both their accuracy and interpretability. Besides, LLM-based approaches may also suffer from performance instability due to their overconfidence.

Different from the above works, we introduce personality evidence utterance mining as an intermediate task, and adopt a Generative–Discriminative Feedback Refinement mechanism to enable bidirectional co-evolution between the graph model and the LLM at both the mechanistic and parameter levels. This design achieves high-performance and interpretable personality prediction.

3 Methodology

The architecture of the proposed model, which consists of two main components: a hierarchical heterogeneous graph–based evidence utterance mining module and an explicit evidence-aligned LLM personality prediction module, is shown in Figure 2. The graph module employs a graph convolutional network (GCN)(Kipf and Welling, 2017) to identify utterances that strongly reflect personality traits and feeds the results as prompts to the LLM. Here, the evidence utterance mining for a specific trait is formulated as a sequence-level binary classification, judging each utterance for showing clear personality indications or not. Then, the personality prediction module uses a Generative–Discriminative Feedback Refinement mechanism to refine the LLM's initial predictions based on mined evidence. This task is formulated as a five-dimensional, three-class classification for each user, assessing each Big Five trait as high, low, or indeterminate. The LLM's outputs are also fed back to update the graph model's parameters. By mutually influencing each other in a cyclical optimization process, the two modules achieve simultaneous improvements in both personality prediction performance and interpretability.

3.1 Graph-based Evidence Mining

In dialogue scenarios, mining evidence utterances reflecting speakers' affective dynamics and communication patterns allows model predictions to be grounded in semantically meaningful text segments, thereby enhancing the interpretability and transparency of personality classification results. Motivated by this insight, the proposed method constructs a hierarchical heterogeneous graph over dialogue data modeling speakers' roles and interactions to identify key evidence utterances, improving both prediction performance and interpretability.

Specifically, for each input dialogue instance, the dialogue utterances and personality representations are jointly modeled as a hierarchical heterogeneous 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, $\{V_a, V_b, V_p\}$, corresponding to utterances from speaker A, utterances from speaker B, and representation nodes of the five personality traits, respectively.

For the constructed graph, we employ a two-layer GCN to capture both the similarities among text nodes and the associations between texts and personalities. By leveraging local graph structural information, the GCN enriches the feature representations encoded in node embeddings, thereby producing global node representations $\mathbf{H}$ for personality evidence utterance mining:

$$\mathbf{H} = \widetilde{\mathbf{A}} \cdot ReLU(\widetilde{\mathbf{A}}\mathbf{X}\mathbf{W_1})\mathbf{W_2}, \quad (1)$$

$$\widetilde{\mathbf{A}} = \mathbf{D}^{-\frac{1}{2}}(I + \mathbf{A})\mathbf{D}^{-\frac{1}{2}}, \quad (2)$$

$$[\mathbf{D}]_{ii} = \sum_j [\mathbf{A}]_{ij}, \quad (3)$$

where $\mathbf{A}$ denotes the adjacency matrix encoding the graph connectivity, $\mathbf{X}$ represents the initial node embeddings, $\mathbf{W}_1$ and $\mathbf{W}_2$ are trainable parameters.

For the embedding matrix $\mathbf{X}$, the feature vectors of all nodes are generated using a pre-trained BERT model. Specifically, dialogue nodes take the corresponding texts as input, while personality nodes use the textual descriptions of the corresponding traits from the Big Five Inventory-2 (BFI-2)(Soto and John, 2017; Zhang et al., 2022) 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.

For the adjacency matrix $\mathbf{A}$, the edges among nodes mainly encode three types of relations: within speakers, between speakers, and personality-related edges. As shown in Figure 2, utterances from the same speaker are sequentially connected according to their speaking order, modeling the temporal dependency and logical continuity of the utterance sequence. Inter-speaker edges are constructed only when a speaker switch occurs (e.g., speaker B replies with utterance $i+1$ immediately after speaker A's utterance i), in order to capture reply relations and cross-speaker information flow in the dialogue. For edges associated with personality nodes, each of the five personality nodes is connected to all utterance nodes from the target speaker, which serves as the basis for mining evidence relations between specific utterances and target personality traits. In addition, all personality nodes are also fully connected with each other to align the representation space across different personality dimensions and to prevent feature vectors from becoming outliers. All edges are undirected, so an edge between nodes i and j is denoted as $[\mathbf{A}]_{ij} = [\mathbf{A}]_{ji} = 1$, and 0 otherwise.

After obtaining the global node representations $\mathbf{H}$, the proposed method computes the cosine similarity between the feature vectors of all utterance nodes from the target speaker, denoted as $\mathbf{H}_{tar}$, and those of the five personality nodes, denoted as $\mathbf{H}_{per}$. An utterance is then identified as an evidence utterance for a given personality trait if the corresponding similarity score exceeds a predefined threshold θ, which is initially set to 0.5 and is subsequently adjusted based on feedback from the personality prediction module.

The predicted results y_{pred} are then compared with the ground-truth labels $\hat{y}$ with binary cross-entropy as the loss function to optimize the parameters of the graph and the upstream BERT:

$$L = -y_{pred} log(\hat{y}) - (1 - y_{pred})\, log(1 - \hat{y}). \quad (4)$$

By continuously updating the model parameters until convergence, the evidence utterance mining results for the current iteration are obtained. These results will serve as auxiliary prompt information for the personality prediction module, guiding and refining the subsequent classification process.

3.2 Generative–Discriminative Enhanced Personality Prediction

Building upon the evidence utterance information that captures users' interaction behaviors and trait-relevant signals, a Generative–Discriminative Feedback Refinement mechanism is then proposed. It focuses on evidence indicative of distinct personality traits to constrain and revise the initial personality predictions $\hat{p}$ produced by the LLM, thereby enhancing both accuracy and interpretability.

Specifically, this mechanism reformulates the original personality level prediction task as determining which personality level yields the strongest consistency with the evidence utterance, thereby converting the classification decision into a simpler distance comparison problem. The mechanism consists of three components: a Generator, a Discriminator, and a Feedback loop. The Generator produces different texts corresponding to all personality levels based on the extracted evidence utterance, while the Discriminator infers the specific personality level represented by original utterance based on the differences between it and the generated texts. Finally, the Feedback loop aggregates

all the outputs of both the Generator and the Discriminator to re-estimate the speaker's personality traits and correct the basic prediction results.

For the generation component, an utterance Generator is constructed to simulate dialogue content conditioned on different levels of personality traits. In this process, the Generator adopts a Counterfactual Reasoning mechanism and leverages the LLM's capability for personality alignment to simulate alternative evidence utterance formulations under different trait levels using the dialogue history. Specifically, for the t-th utterance u_t in current dialogue, the Generator considers five personality traits and, for each trait in turn, simulates the t-th utterance content $u_t^{'}$ under three personality level conditions—high, low, and not exhibiting the trait—based on the preceding t-1 utterances:

$$u_t' = Generator(u_1 \dots u_{t-1}, trait, level). \quad (5)$$

Subsequently, the Discriminator is employed to analyze the generated content for each trait and selects the most reliable personality level based on its deviation from the original utterance. Specifically, the Discriminator adopts an LLM-as-a-Judge approach, in which a LLM assigns scores to quantify the distance between statements generated under different trait levels and the original evidence utterance. The trait level associated with the simulated utterance with the shortest distance is then taken as the prediction p_t for t-th evidence utterance:

$$Dist_{t,level,trait} = Score_{LLM}(u_t, u_t'), \quad (6)$$

$$p_t = \arg\min_{level} (Dist_{t,level}). \quad (7)$$

Finally, the Feedback loop aggregates the trait level prediction results produced in the Discrimination component for all evidence utterances. It then applies a voting scheme to obtain an updated trait level for current dialogue under the corresponding personality trait. For those origin $\hat{p}$ whose level is not labeled as unknown, the Feedback loop further refines them using this newly updated level, yielding the final personality distribution of the personality prediction module:

$$\hat{p}_{trait} = \begin{cases} \hat{p}_{trait}, \text{if } \hat{p}_{trait} = unknown, \\ \arg\max \sum_{i=1}^{m} I\left(p_i = p_j\right), \text{otherwise.} \end{cases} \quad (8)$$

where $I(\cdot)$ denotes the indicator function, which takes value 1 if the condition in parentheses holds and 0 otherwise. m is the total number of evidence utterances in the dialogue.

3.3 Iterative Evidence-Personality Optimization

To achieve dynamic mutual reinforcement between the two tasks while avoiding cumulative errors and mutual misguidance, the proposed method performs iterative optimization of the two modules according to the following strategy to simultaneously improving model performance and interpretability.

First, the initial predictions of both evidence utterance mining task and personality prediction task are obtained, which serve as supervisory signals for generating optimization cues for the other module. For the evidence utterance mining task, the predicted results y_{pred} are determined based on the comparison between the node similarity score and the threshold θ described in Section 3.1.

For the personality prediction task, a base agent is first constructed to generate the basic results for personality prediction. The prompt is augmented with the content of the BFI-2 into the prompt, and perform personality prediction over the original dialogue in a zero-shot setting:

$$\hat{p} = BasicAgent(Dialogue_{raw}, BFI-2), \quad (9)$$

where $\hat{p}$ denotes the predicted result of personalities and serves as the basic prediction result for LLM.

Leveraging the initial predictions from the two tasks, the proposed method first utilizes the Generative–Discriminative Feedback Refinement mechanism to refine the LLM's personality predictions and obtain updated results. Then, it further uses the newly generated predictions to adjust the evidence utterance decision threshold θ in the evidence utterance mining module, providing feedback that improves the extraction of evidence utterances:

$$\theta'_{trait} = \sigma(\mathrm{MLP}(\hat{p}_{trait})). \quad (10)$$

With the threshold updated, the graph model performs inference again, and the newly mined evidence utterances are then used to further refine the personality prediction module. The above procedure is repeated until both modules stabilize, yielding predictions that are both reliable and interpretable, along with the corresponding evidence.

4 Experiment Setup

4.1 Dataset

We evaluate our framework on PersonalityEvd (Sun et al., 2024), a dataset with evidence utterance annotations built from the Chinese emotional dialogue corpus CPED (Chen et al., 2022). Grounded in the BFI-2, this dataset covers five traits—Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism—with trait level annotations per dialogue, and provides evidence utterance IDs. Following the original setting, we split the data into train/val/test set with a 7:1:2 ratio.

4.2 Baseline

We compare the proposed model with three groups of baseline methods: (1) Content-based models, (2) Relation-based models, and (3) LLM-based models, which are detailed as follows:

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 model to extract global semantic features from dialogues and then applies a two-layer MLP to output personality predictions.

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.

LLM-based Methods

We evaluate several mainstream LLMs, including GPT-4-Turbo (OpenAI, 2025), Claude-4-Sonnet (Anthropic, 2025), and Deepseek-V3 (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) for comparison. In addition, we adopt CoPE (Sun et al., 2024) as a baseline tailored to conversational settings. It constructs chain-of-thought data and combines it with fine-tuning, representing a leading approach for dialogue-based personality prediction.

4.3 Implementation Details

The experiments were conducted on a single NVIDIA A100 GPU (40GB PCIe). 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. BERT-base-chinese is used to generate node embeddings. The hidden dimension of all GCN layers is 256. For LLM inference, we set the temperature to 0.7 for the Generator, and 0.0 for the Discriminator. All hyperparameters are tuned over the validation set to obtain the optimized results.

5 Experiment Results

5.1 Personality Prediction Performance

We compared 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 1.

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-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

Model	Personality Accuracy (%)						std.dev
	Avg.	OPN	CON	EXT	AGR	NEU	(percent point, pp)
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	50.69	55.10	7.92
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
CoPE	66.45	75.48	68.87	63.39	63.36	**61.15**	5.18
Ours + GPT	63.69	71.35	63.09	60.88	65.79	57.35	4.72
Ours + Deepseek	**67.44**	74.38	68.60	**65.01**	**69.15**	60.06	4.75

Table 1: Overall results of the proposed and baseline models in the personality prediction task.

Model	Personality F1 score (%)						std.dev (pp)
	Avg.	OPN	CON	EXT	AGR	NEU	
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	-	-	-	-	-	-
Ours	**85.56**	**92.34**	**90.08**	**83.98**	**78.03**	**83.37**	5.11

Table 2: Overall results of the proposed and baseline models in the evidence utterance mining task.

average accuracies ranging from 60% 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 67.44%, surpassing the second-best CoPE by 1.72%, and improving over GPT or Deepseek alone by 5.57% and 6.17%, respectively, demonstrating its effectiveness.

In addition, the proposed model can also reduce performance variance across different personality traits, yielding more balanced predictions. As shown in Table 1, 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 Refinement mechanism, the proposed method effectively mitigates overconfidence-induced errors, reducing the accuracy variance to 4.72pp, and lead to reductions of 42.00% and 51.18% compared to using the original LLMs alone.

5.2 Evidence Utterance Mining Performance

To evaluate the effectiveness of the proposed method in enhancing the interpretability of personality prediction, we further compare it with different LLM-based strategies on the evidence utterance mining task. The results are reported in Table 2. The proposed method achieves the highest F1 score of 85.56%, outperforming the second-best model Deepseek by 10.71%, demonstrating its reliability in mining personality-related evidence utterances and its effectiveness for interpretable personality prediction. Moreover, consistent with the personality prediction task, our method also alleviates performance imbalance across personality traits in evidence utterance mining, reducing the standard deviation of F1 scores across the five traits by 27.72% compared to the second-best result. These findings indicate that the proposed approach not

Method	Acc	F1
GPT w/o graph	60.33	59.99
GPT w/ graph	63.69	64.30
Deepseek w/o graph	63.52	63.90
Deepseek w/ graph	67.44	68.34

Table 3: LLM personality prediction results.

Method	Acc	F1
only graph	78.55	83.10
w/ GPT	83.73	84.95
w/ Deepseek	84.07	85.56

Table 4: Graph evidence utterance mining results.

only reaches higher overall performance but also exhibits greater cross-trait stability and robustness, substantially improving its practical applicability and interpretability reliability.

5.3 Impact of Bidirectional Optimization

To examine the effectiveness of the proposed bidirectional iterative optimization strategy, we further compare the performance of each module before and after being refined by the other module. Specifically, Table 3 reports the performance of two LLMs on the personality prediction task when used independently and when collaboratively optimized with the proposed graph module. Incorporating the graph module leads to accuracy improvements of 5.57% and 6.17%, and F1 score gains of 7.18% and 6.95% for GPT and Deepseek, respectively, demonstrating the significant benefit of inferring personality traits based on reliable evidence utterances. Table 4 presents the results for the evidence utterance mining task, comparing the stand alone graph-based classifier with versions refined by GPT and Deepseek. Compared to using the graph model alone, introducing GPT and Deepseek yields accuracy improvements of 6.59% and 7.03%, and F1 score increases of 2.23% and 2.96%, respectively, indicating that feedback from personality prediction can also enhance the reliability of evidence mining. Overall, these bidirectional reinforcing improvements validate the effectiveness of the proposed framework.

5.4 Case Study

Figure 3 illustrates a scenario for Agreeableness recognition, where Speaker B must respond to a request to enter a restricted broadcasting station. Facing this dialogue's diverse emotional signals and complex contextual conditions, because the graph model associates utterances with personality

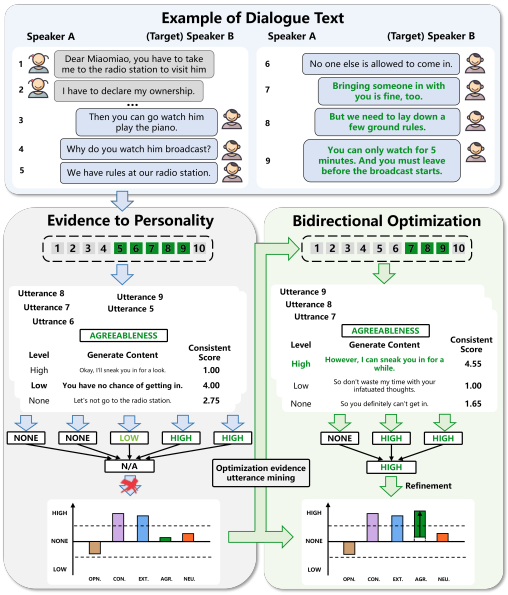

Figure 3: Example of iterative optimization for refining personality prediction on Agreeableness.

traits only during the evidence utterance mining, so that the initial mining produces a noisy set of candidate evidence utterances, which makes it difficult to resolve conflicting cues in such a setting. And initial personality prediction also occurs misunderstanding. In contrast, the LLM's assessment of the personality trait level can be used to improve the precision of evidence utterance mining, guiding the graph model to focus on the core evidence that genuinely reflects the trait, the compromising behaviors (U7, U8, U9). In the subsequent Generative-Discriminative Feedback Refinement, the model confirms Speaker B's exception-making behavior exhibits high empathy and flexibility. As a result, the final prediction is corrected to High Agreeableness, highlighting the framework's ability to capture personality traits in complex contexts via iterative refinement, and demonstrating the effectiveness of bidirectional iterative enhancement between evidence and personality.

6 Conclusion

This paper proposes a Generative–Discriminative Feedback Refinement mechanism for dialogue-based personality prediction. A bidirectional iterative optimization framework is constructed with a graph model for evidence utterance mining and a Feedback Refinement mechanism for personality prediction. Extensive experiments on public datasets validate the effectiveness of the proposed framework.

Limitations

The proposed method collaboratively optimizes 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, nor does it include a detailed quantitative analysis of model size and computational cost. These aspects 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 a research benchmark (PersonalityEvd, derived from CPED), and we do not claim deployment readiness. Models trained on such data may reflect dataset-and-annotation-specific biases and may not generalize 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 decision pipelines, and conducting additional evaluations of fairness, robustness, and privacy risks prior to any real-world deployment.

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.

Pranav Bhandari, Nicolas Fay, Michael J Wise, Amitava Datta, Stephanie Meek, Usman Naseem, and Mehwish Nasim. 2025. Can LLM agents maintain a persona in discourse? In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 29213–29229, Suzhou, China. Association for Computational Linguistics.

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.

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, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

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.

Zhancheng Ren, Qiang Shen, Xiaolei Diao, and Hao Xu. 2021. A sentiment-aware deep learning approach for personality detection from text. *Inf. Process. Manag.*, 58(3):102532.

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

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.

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.

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.

A Appendix

A.1 Generator and Discriminator Prompt

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>