Revealing Personality Traits: A New Benchmark Dataset for Explainable Personality Recognition on Dialogues

Lei Sun [1]
leisun@ruc.edu.cn

Jinming Zhao [2,*]
zhaojinming1@gmail.com

Qin Jin [1,*]
qjin@ruc.edu.cn

[1]School of Information, Renmin University of China, [2]Independent Researcher

Abstract

Personality recognition aims to identify the personality traits implied in user data such as dialogues and social media posts. Current research predominantly treats personality recognition as a classification task, failing to reveal the supporting evidence for the recognized personality. In this paper, we propose a novel task named **Explainable Personality Recognition**, aiming to reveal the reasoning process as supporting evidence of the personality trait. Inspired by personality theories, personality traits are made up of stable patterns of personality state, where the states are short-term characteristic patterns of thoughts, feelings, and behaviors in a concrete situation at a specific moment in time. We propose an explainable personality recognition framework called **Chain-of-Personality-Evidence** (CoPE), which involves a reasoning process from specific contexts to short-term personality states to long-term personality traits. Furthermore, based on the CoPE framework, we construct an explainable personality recognition dataset from dialogues, **PersonalityEvd**. We introduce two explainable personality state recognition and explainable personality trait recognition tasks, which require models to recognize the personality state and trait labels and their corresponding support evidence. Our extensive experiments based on Large Language Models on the two tasks show that revealing personality traits is very challenging and we present some insights for future research. Our data and code are available at https://github.com/Lei-Sun-RUC/PersonalityEvd.

1 Introduction

Personality, a characteristic way of thinking, feeling, and behaving (Roberts, 2009), has a great impact on our lives, well-being, and health. Therefore, identifying a person's personality has great potential in many real-world applications, such as

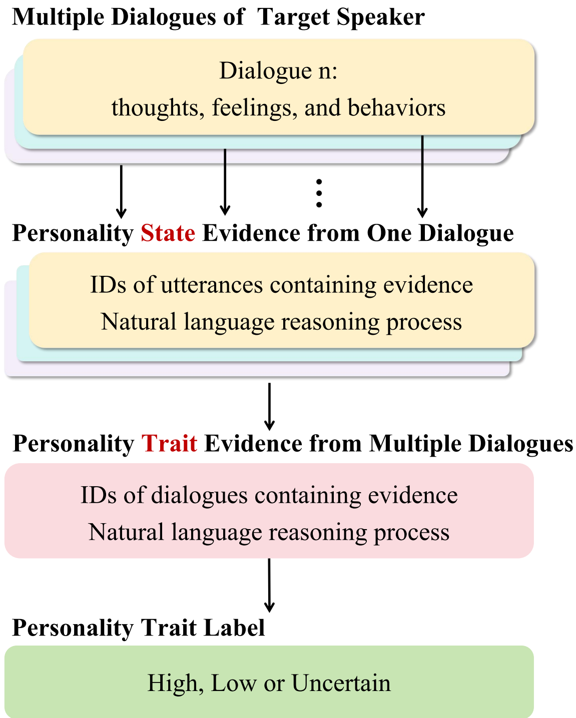

Figure 1: Chain-of-Personality-Evidence (CoPE) framework illustrates the reasoning process for revealing supporting evidence of personality traits.

human-computer interaction (Attig et al., 2017), psychological diagnosis and regulation (Claridge and Davis, 2013; Redelmeier et al., 2021), and job candidate screenings (Liem et al., 2018; Caldwell and Burger, 1998). Traditional personality recognition methods typically depend on self-reported results from designed questionnaires. Such an approach is not only time-consuming but also necessitates the cooperation of the subjects. Therefore, Automatic Personality Recognition (APR) has attracted increasing attention in recent years, which aims to predict one's personality based on user data. To support the APR research, various personality datasets have been proposed, such as the FriendsPersona dataset (Jiang et al., 2020) based on dialogues, the Essays I (Pennebaker and King,

[†]Corresponding Authors.

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 19988–20002
November 12-16, 2024

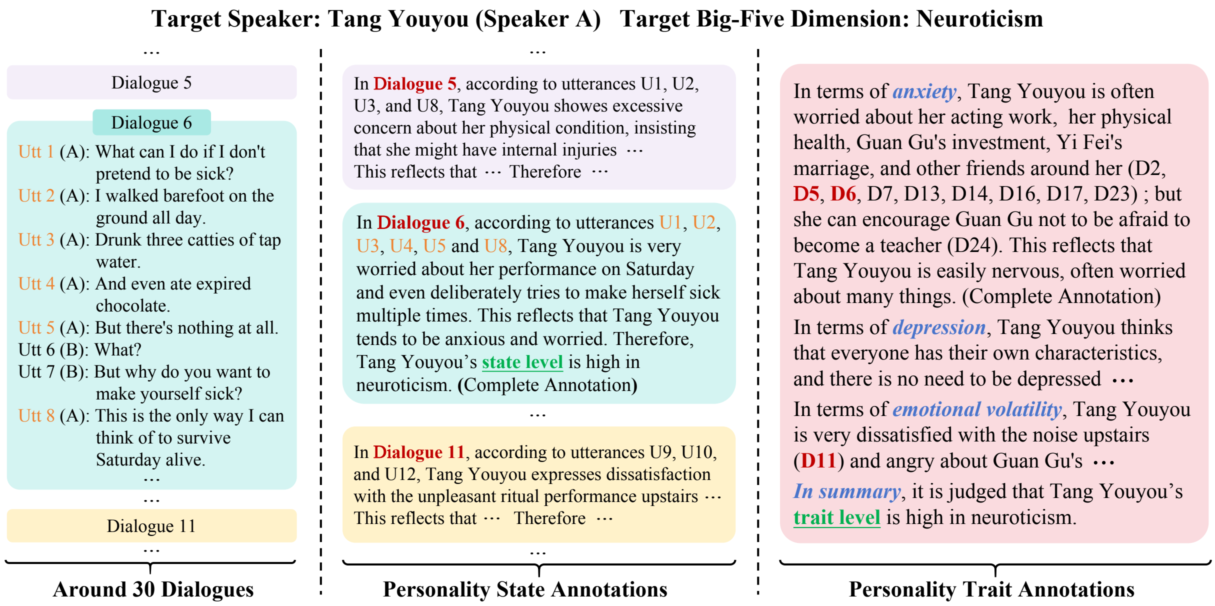

Figure 2: A Speaker's explainable personality annotation of the Neuroticism dimension from the Big-Five Personality Model, including dialogue-level personality state and speaker-level personality trait labels with corresponding evidence (natural language reasoning process). U# denotes evidence utterances and D# denotes evidence dialogues. The neuroticism dimension contains anxiety, depression, and emotional volatility three facets in the BFI-2 scale. For personality trait evidence annotation, we annotate the natural language reasoning process for each facet.

1999) based on stream-of-consciousness essays, the PAN-2015 dataset (Rangel et al., 2015) based on Twitter data, and the YouTube Vlogs dataset (Biel and Gatica-Perez, 2012) based on YouTube videos.

However, previous works have mainly focused on recognizing personality trait labels and fail to reveal the supporting evidence of the personality traits, making the model's predictions uninterpreted. Therefore, in this work, we propose a novel task to reveal personality traits named explainable personality recognition. Most theories of personality suggest that personality traits are made up of enduring patterns of personality states, where the personality states are short-term characteristic patterns of thoughts, feelings, and behaviors in a concrete situation at a specific moment in time (Fleeson, 2001; Fleeson and Jayawickreme, 2015). Inspired by these theories, we propose an explanation framework called **Chain-of-Personality-Evidence** (CoPE) shown in Figure 1, which is a reasoning process to reveal the supporting evidence for personality explanation. The reasoning process goes from specific thoughts, feelings, and behaviors to short-term personality states to long-term stable personality traits. Due to the repetitions in one's state being critical to capture trait (Roberts, 2009), we first analyze the short-term personality state and then reveal the real stable personality trait based on all one's short-term state patterns. When revealing one's personality from dialogues, a dialogue can be seen as a short-term specific context containing thoughts, feelings, and behaviors that reflect one's personality states, and then we can obtain one's personality traits by analyzing personality states from multiple dialogues. To achieve clear and reasonable explanations, personality state evidence is comprised of evidence utterance IDs and natural language reasoning process, and trait evidence is comprised of evidence dialogue IDs and corresponding natural language reasoning process.

Furthermore, based on the proposed CoPE framework, we construct an explainable personality dataset, **PersonalityEvd**, which consists of 72 speakers and about 2000 dialogues from Chinese TV series. So each speaker is involved in around 30 dialogues. We also provide a translated English version of the dialogues. As shown in Figure 2, we annotate not only dialogue-level personality state and speaker-level personality trait labels but also detailed corresponding reasoning processes to justify these labels.

Based on the proposed PersonalityEvd dataset, we propose two sub-tasks: 1) **E**vidence grounded **P**ersonality **S**tate **R**ecognition (**EPR-S**), which requires the model to predict the state label and provide prediction evidence from each dialogue. 2)

Evidence grounded **P**ersonality **T**rait **R**ecognition (**EPR-T**), which requires the model to predict the trait label and provide prediction evidence from one's multiple dialogues. Both tasks are highly challenging, especially for the trait-level task which exists conflicts between short-term states, interactions with different interlocutors, the long context from lots of dialogues, etc.

We further establish a strong baseline with powerful Large Language Models (LLMs) and conduct extensive experiments on the explainable personality tasks. Human evaluation results show that current LLMs are far from humans in personality understanding. We analyze the experimental results and find that introducing supporting evidence helps improve the performance of personality recognition. We also discover that analyzing the state evidence as an intermediate result contributes to the EPR-T task, which also proves the necessity of introducing the EPR-S task. We hope our insights can offer inspiration for further exploration.

The main contributions of this work include: (1) We propose a personality explanation framework called **Chain-of-Personality-Evidence** (**CoPE**) based on personality theories, which reveals a detailed reasoning process as supporting evidence of personality traits. (2) We manually construct a high-quality supporting dataset, **PersonalityEvd**, based on dialogues to support explainable personality recognition tasks. (3) We introduce two personality recognition tasks and propose a LLM-based strong baseline method. We conduct extensive experimental results and present some insights for future research.

2 Related Work

2.1 Personality Theories

Various personality theories have been developed to categorize, interpret, and understand human personality, including the Cattell Sixteen Personality Factor (16PF; (Cattell and Mead, 2008)), the Hans Eysenck's psychoticism, extraversion, and neuroticism (PEN; (Eysenck, 2012)), Myers-Briggs Type Indicator (MBTI; (Briggs, 1976)), the Big-Five Model (McCrae and John, 1992) and so on. Among them, the most frequently used personality models are the Big-Five Model and the MBTI model. The Big-Five model measures personality through five dipolar scales: **O**penness, **C**onscientiousness, **E**xtraversion, **A**greeableness, and **N**euroticism. We utilize the Big-Five model in this work. The MBTI model lays out a binary classification based on four distinct functions (**E**xtraversion/**I**ntroversion, **S**ensing/I**N**tuition, **T**hinking/**F**eeling, **J**udgment /**P**erception).

2.2 Automatic Personality Recognition

Automatic personality recognition, as an important topic in computational psycho-linguistics, focuses on determining one's personality from a variety of data sources, such as dialogues, Twitter, Facebook, and YouTube (Mushtaq and Kumar, 2022). For text modality, essays are a popular mode of text and the corresponding datasets such as Essays II (Tausczik and Pennebaker, 2010). Rangel et al. (2015) proposes the PAN-2015 corpus collected from Twitter in four languages: English, Spanish, Italian, and Dutch. The TwiSty dataset (Verhoeven et al., 2016) is also a multilingual twitter stylometry corpus for personality profiling. MyPersonality dataset (Park et al., 2015) comprises status updates of over 66,000 Facebook users. There are also several dialogue-based datasets, such as the FriendsPersona dataset (Jiang et al., 2020) from Friends TV Show, Story2Personality Dataset (Sang et al., 2022) from movie scripts, PANDORA (Gjurković et al., 2021) and MBTI9k (Khan et al., 2020) dataset sourced from Reddit posts. Some datasets with other modalities have been proposed as well, such as audio (Polzehl et al., 2010; Mohammadi and Vinciarelli, 2012), visual (Cristani et al., 2013), and multimodal modalities (Sanchez-Cortes et al., 2013; Ponce-López et al., 2016; Escalante et al., 2017; Celiktutan et al., 2017; Palmero et al., 2021). However, no personality evidence to justify the label has been considered in all previous works.

3 The PersonalityEvd Dataset

3.1 Dialogue Selection

We build our PersonalityEvd dataset by leveraging the CPED corpus (Chen et al., 2022), which is a large-scale Chinese emotional dialogue dataset containing more than 12K dialogues of 392 speakers from 40 TV shows. We first remove dialogues with less than 10 utterances because these dialogues may be incomplete or contain insufficient information. We then randomly select 72 speakers and 30 dialogues for each speaker to form a candidate set with 2,160 dialogues in total, based on which we annotate our evidence grounded dataset.

3.2 Annotation Content

We employ the Big Five Inventory-2 scale (BFI-2) (Soto and John, 2017; Zhang et al., 2022) as our theoretical foundation, which contains 60 short and easy-to-understand phrases. Every Big-Five dimension has 12 characteristic items, of which 6 items belong to the high level and the remaining 6 items belong to the low level. More detailed information can be found in Appendix A.

Following previous personality recognition works (Rangel et al., 2016; Jiang et al., 2020), we define three levels for the Big-Five personality dimensions: *high, low, uncertain*. *high* / *low* means that the target speaker in the dialogue exhibits a high or low level of characteristics in the corresponding Big-Five dimension, while *uncertain* means that the level of the target Big-Five personality dimension cannot be judged according to the dialogue. In addition to personality trait labels, we also annotate state labels, which can reflect the potential personality tendencies of the speaker.

As for the personality state evidence, it consists of three parts: *evidence utterance IDs*, *utterance summaries*, and *personality characteristics*. *Utterance summaries* and *personality characteristics* together constitute natural language reasoning process of personality states. For the *evidence utterance IDs*, we only choose the utterances from the target speaker, although other speakers in the dialogue may provide some background information. We annotate the most relevant utterances that could reflect the personality, such as utterances within personality keywords. The *utterance summaries* are the concrete thoughts, feelings, or behaviors summarized into natural language, according to the evidence utterance IDs. The *personality characteristics* are the features that the *utterance summaries* reflect, which are from the 60 personality description items in the BFI-2 psychological questionnaire. Since the 60 items of the BFI-2 are still limited and cannot cover various situations, we allow annotators to make slight changes to the items to adapt to the dialogue context while keeping the meaning consistent with the original descriptions. Finally, we organize the three parts and the state label into the form of Chain-of-Thought (CoT) (Wei et al., 2022; Ho et al., 2023) using an overall description template structure, such as "according to utterances ...(evidence utterance IDs)..., ...(utterance summaries)... This reflects ...(personality characteristics)... Therefore ...(state label)... ".

As for the personality trait evidence, we formulate it as a combination of three-faceted descriptions of the target Big-Five dimension. For example, the neuroticism dimension includes anxiety, depression, and emotional volatility based on the BFI-2 psychological questionnaire. Each facet description also contains three parts, and they play similar roles as state evidence components: *evidence dialogue IDs*, *dialogue summaries*, and *personality characteristics*. *Evidence dialogue IDs* is enclosed in parentheses after the corresponding *dialogue summaries*. We format them using similar templates, such as "In terms of (facet), ...(dialogue summaries)... This reflects ...(personality characteristics)...". Finally, we also organize the description of these three facets and the trait label into a CoT format, as shown in Figure 2.

3.3 Annotation Process of Personality States

The annotation of state evidence is highly difficult, as it requires an analysis that encompasses the five dimensions of the Big Five personality theory. To reduce annotation difficulty, improve annotation speed, and ensure high quality, we take two steps: first using GPT-4-Turbo (Achiam et al., 2023) to pre-annotate, and then performing manual correction.

3.3.1 GPT-4 Pre-Annotation

GPT-4 exhibits strong performance on various tasks, including the personality prediction and explanation task (Ji et al., 2023). As there are five dimensions of the Big-Five Model, we handle these dimensions separately, which means that a dialogue will be taken as input five times to obtain five Big-five dimension results. Due to the limited input token length of GPT-4, we use GPT-4 to analyze five dialogues at once. Detailed prompts are described in Appendix B. Because of the powerful natural language understanding capability, reasoning ability, and world knowledge, GPT-4 provides a good basis for further correction.

3.3.2 Human Correction

Annotators. Since the results of GPT-4 still contain mistakes, we screen 12 undergraduates majoring in psychology who have adequate knowledge about the Big-Five Personality Model to perform further correction. We first organize a special workshop to train them on our tasks. Each participant needs to complete trial annotations. Then feedback is provided to participants to correct their mistakes.

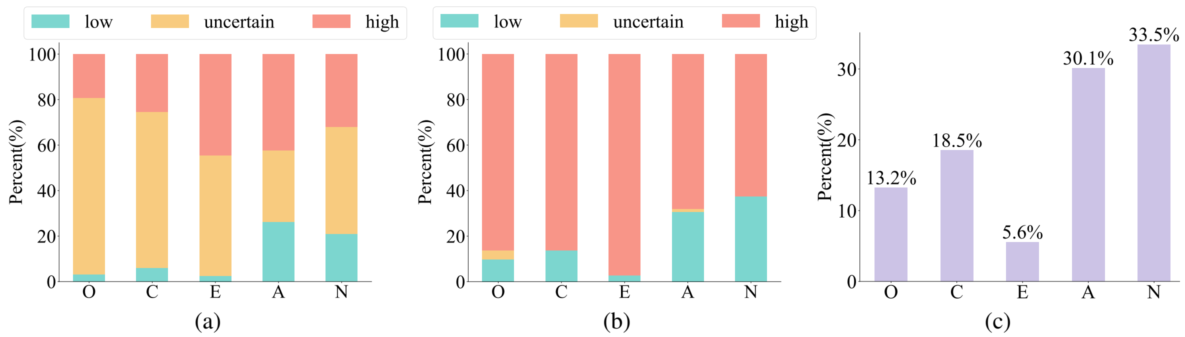

Figure 3: (a) The distribution of state labels. (b) The distribution of trait labels. (c) The ratio of the state label different from the trait label. (O: openness, C: conscientiousness, E: extraversion, A: agreeableness, N: neuroticism)

Only when they achieve satisfactory performance on the trial dialogues can they start annotating the main dataset.

Annotation Guidelines. Each sample is first annotated by one annotator. The annotator first understands the dialogue context and then makes corrections based on the results predicted by GPT-4. We ask annotators to analyze obvious personalities and not to over-interpret or speculate on the dialogue. For evidence utterance IDs, we ask the annotator to annotate the most relevant sentence IDs; for utterance summaries and personality characteristics, if the analysis of GPT-4 is reasonable and comprehensive, the annotator needs to simplify the content to highlight the key points. If the analysis is wrong, the annotator needs to re-annotate. Approximately 30% of the samples were re-annotated manually, greatly improving the efficiency of our dataset construction.

There are two new quality inspectors who are graduate students with more specialized knowledge in personality theory, ensuring higher accuracy in the annotations. They give the annotator suggestions for modifications to help the annotator refine the annotation. If there is a disagreement between the quality inspector and the annotator, they will discuss to reach a consensus. We remove samples where there is no consensus. The two quality inspectors review all the samples to ensure the annotation quality.

Final Check. At the final stage, we, the authors, manually review, filter, and correct all the annotations again. We design several strategies to filter and correct the unsatisfactory annotations: 1) We remove dialogues whose state labels of the five Big-Five personality dimensions are all *uncertain* labels, as these dialogues cannot provide evidence information and thus don't match the focus of our dataset. The number of these dialogues is small and does not affect our dataset scale. 2) We remove dialogues with contradictory personality information. Although it is reasonable that one shows contradictory personalities, this will make it difficult to annotate the state label. 3) We correct the language errors, such as typos, misnomers, etc.

3.4 Annotation Process of Personality Traits

Annotators. After completing the personality state annotations, we recruit another 6 undergraduates majoring in psychology to finish the personality trait annotations. We train them using similar training steps as the state labeling for trait annotation.

Annotation Guidelines. Each sample is first labeled individually by 3 annotators. Annotators are asked to score and write the evidence according to the BFI-2 scale based on about 30 dialogues and previous state annotations. To maintain the consistency of state and trait annotations, annotators are allowed to modify the state annotation if they feel that the original annotation is unreasonable. Some dialogues are difficult to understand because the dialogue scenes are difficult to infer, involve many characters, etc., and different annotators may interpret the dialogue differently.

After getting individual annotations, the 3 annotators will reach a consensus through discussion, including the score and corresponding evidence. To obtain the trait label, we calculate the average score of 12 BFI-2 items corresponding to each dimension and convert it to binary labels using the median split.

Final Check. In the end, we manually review and correct all trait annotations again: 1) we update the state annotations modified in this stage into final results. 2) We also check for language errors.

(a) High Openness (b) Low Openness (c) High Conscientiousness (d) Low Conscientiousness

Figure 4: The word clouds of personality **states** reasoning process on openness and conscientiousness dimensions.

Statistics	**Number**
# Dialogues	1,924
# Speakers	72
# 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
Avg. evi dlg IDs per spk	11.96
SD. dlg per spk	2.47
SD. utt per dlg	5.36
SD. length of utt	4.76
SD. tgt spk utt per dlg	4.01
SD. evi utt IDs per dlg	2.87
SD. evi dlg IDs per spk	5.88

Table 1: **PersonalityEvd** dataset statistics. (utt, tgt, spk, dlg, evi, SD refer to utterance, target, speaker, dialogue, evidence, standard deviation.)

3.5 Dataset Statistics

Dialogue Statistics. Table 1 presents the statistics of our constructed PersonalityEvd dataset. It contains 1,924 dialogues of 72 speakers, which means that every speaker is involved in approximately 30 dialogues. The average number of utterances per dialogue is 16.98, indicating long context and rich information in our dataset. The average number of utterances and dialogue IDs containing evidence is 4.26 and 11.96, respectively. The average number of target speaker utterances percentage per dialogue is about 50%, indicating the balanced involvement in dialogues.

Annotation Statistics. Figure 3 presents the distribution of two-level personality labels and the ratio of the state label when it's different from the trait label. As for the state labels, we can see that *uncertain* labels occupy a relatively large proportion due to the sparsity of personality information in dialogue. As for the trait labels, *high* labels account for a significant proportion as characters of TV series often have distinct personalities. Figure (c) shows the ratio of the state label when it's different from the trait label, which illustrates the daily variance of personality.

Figure 4 presents the top frequently used words in the natural language reasoning process of different state labels of the openness and conscientiousness dimensions, which can reflect the personality keywords of our PersonalityEvd dataset. For example, in the openness dimension, words like "different" and "curious" show a high level of openness. The word clouds of remaining three dimensions are shown in Appendix C.

Dataset splits. For the state-level data, we randomly split it into train/valid/test sets based on speakers according to the ratio of 7:1:2, ensuring that the speakers in the training set do not appear in the valid or test set. For the limited trait-level data, we randomly divide the data into 3 folds based on speakers, so each fold contains 24 speakers.

4 Proposed Tasks

We introduce two novel sub-tasks based on our dataset: Evidence grounded Personality State Recognition (**EPR-S**) and Evidence grounded Personality Trait Recognition (**EPR-T**).

4.1 The EPR-S Task

Definition There are five Big-Five personality dimensions $BF = [bf_1, bf_2, ..., bf_5]$. Each dialogue D includes m speakers $P = [p_1, p_2, ..., p_m] (m \geq 2)$. The EPR-S task aims to recognize the state label $y_s \in \{high, low, uncertain\}$ as well as generate its evidence E_s of the target speaker p_i for the target Big-Five personality dimension bf_j given one dialogue D.

Evaluation As the prediction of the EPR-S task contains four parts, we evaluate them separately. For the evaluation of evidence utterance IDs, we report the binary F1-score. For the evaluation

Task	Model	ID F1	BERT	Claude	GPT-4	Personality Accuracy					
		Avg	Avg	Avg	Avg	O	C	E	A	N	Avg
EPR-S	GLM-32k	75.42	**83.45**	3.61	3.48	74.10	68.59	63.36	55.64	61.43	64.62
	Qwen-32k	**75.94**	83.44	**3.68**	**3.56**	75.48	**68.87**	**63.39**	63.36	61.15	**66.45**
	GPT-4	71.20	76.21	3.55	3.27	**76.85**	50.41	50.96	**69.97**	**62.25**	62.09
EPR-T	GLM-32k	40.28	76.80	2.95	2.74	**81.76**	**86.11**	95.77	**73.12**	52.17	**77.78**
	Qwen-32k	**44.39**	**77.81**	**3.25**	**3.11**	74.75	75.96	**97.16**	70.41	**64.67**	76.59

Table 2: Model performance of CoT fine-tuning on two tasks. ID F1 denotes the binary F1-score of evidence utterance/dialogue IDs. BERT, Claude, and GPT-4 refer to BERTScore (F1), Claude-3-sonnet, and GPT-4-Turbo score. The Claude-3-sonnet and GPT-4-Turbo scores range from 1 to 5. CoT fine-tuning: the model is trained to generate the evidence and then the answer, as shown in Figure 2. For the EPR-T task, we report the results of 3-fold cross-validation. (O: openness, C: conscientiousness, E: extraversion, A: agreeableness, N: neuroticism)

of the natural language reasoning process, consisting of utterance summaries and personality characteristics, we use BERTScore (Zhang et al., 2019) to measure the semantic similarity between the ground truth and predicted evidence. As BERTScore is still a limited metric, we also use claude-3-sonnet-20240229 (Anthropic, 2024) and gpt-4-turbo-2024-04-09 (Achiam et al., 2023) to evaluate the semantic overlapping level from 1 to 5 (Li et al., 2024b). For the evaluation of the personality label, to be consistent with previous works (Majumder et al., 2017; Jiang et al., 2020; Guo et al., 2024), we use accuracy as the evaluation metric.

4.2 The EPR-T Task

Definition The EPR-T task aims to recognize the trait label $y_t \in \{high, low, uncertain\}$ as well as generate corresponding evidence E_t of the target speaker p for the target Big-Five personality dimension bf_i given dialogues $[D_1, D_2, ..., D_n]$.
Evaluation The prediction of the EPR-T task has a similar structure as the EPR-S task. For the evaluation of evidence dialogue IDs and trait labels, we apply the same metrics as the EPR-S task. For the natural language trait evidence, we compute the average similarity score of the three facets between the ground truth and predicted description, using the same three models mentioned above.

5 Experiments

5.1 Baselines

We evaluate three baseline LLMs on our tasks:
• **ChatGLM3-6B-32K** based on ChatGLM3-6B (Du et al., 2022; Zeng et al., 2022), further strengthens the ability to understand long texts and can better handle contexts up to 32K in length.
• **Qwen1.5-7B-Chat** is the improved version of Qwen (Bai et al., 2023), which has significant model quality improvements in chat models and supports 32K context length.
• **GPT-4-Turbo-2024-04-09** is a snapshot of GPT-4-Turbo (Achiam et al., 2023) from April 9th, 2024 with more powerful performance and lower price than GPT-4.

We denote them as **GLM-32k**, **Qwen-32k**, and **GPT-4** respectively in the following sub-sections.

Model	Fluency	Coherence	Plausibility
Ground Truth	**4.61**	**4.38**	**4.31**
GLM-32k	3.82	2.51	2.59
Qwen-32k	3.90	2.68	2.65

Table 3: Results of human evaluation on the natural language reasoning process of personality **traits**. The score range is 1 to 5.

5.2 Main Results

Tabel 2 presents the results of two tasks on the test set. We evaluate GPT-4-Turbo in the zero-shot setting and other models in the LoRA (Hu et al., 2021) fine-tuning (FT) setting. For the EPR-S task, Qwen1.5-7B-Chat achieves the best overall performance among the three models. Surprisingly, GPT-4-Turbo achieves comparable results in the zero-shot setting and even surpasses the other two models in openness, agreeableness, and neuroticism dimensions in terms of personality accuracy. For the EPR-T task, in terms of average personality accuracy, ChatGLM3-6B-32K outperforms Qwen1.5-7B-Chat, while Qwen1.5-7B-Chat surpasses ChatGLM3-6B-32K on evidence-related metrics. In summary, the scores of these models on both tasks are still low, implying significant room for further improvements.

To avoid potential bias in the GPT-4-Turbo evaluation of self-generated content, we also use Claude-3-sonnet for evaluation. In all experiments, Claude-3-sonnet scores slightly higher than GPT-4-Turbo, indicating that GPT-4-Turbo does not give higher scores to its own generated results.

5.3 Human Evaluation

We evaluate the natural language reasoning process of the traits, concerning the following criteria:

• **Fluency** measures the grammatical and formatting aspects of the sentences.

• **Coherence** measures whether the text is semantically and factually consistent with the dialogue context.

• **Plausibility** measures whether the text contains comprehensive and correct trait evidence.

We randomly selected 50 samples of 10 speakers for human evaluation. The annotation was performed by undergraduates majoring in psychology, and we assigned five evaluators to each sample. They assessed each aspect on a scale from 1 to 5, with higher scores indicating better results. Finally, we calculate the average scores of these evaluators. Results are reported in Table 3. The three scores of ground truth are very close to 5 points, so we can conclude that the quality of PersonalityEvd is guaranteed. The fluency scores of ChatGLM3-6B-32K and Qwen1.5-7B-Chat reach 3.82 and 3.90 respectively, probably because current LLMs are trained on a large corpus and have mastered human language. Both models have low coherence and plausibility scores, demonstrating the great challenge of explaining personality.

5.4 Ablation Study

Potential benefits of corresponding evidence for personality recognition. In addition to CoT fine-tuning, we provide the results of Hybrid fine-tuning, inspired by the method in (Li et al., 2024a), as shown in Table 4. We use the ChatGLM3-6B-32k model for the validation. For state recognition, the performance of Hybrid-Direct setting and Hybrid-CoT is 2.1% and 1.71% higher than that of Direct fine-tuning, respectively, proving that introducing evidence helps improve the reasoning ability of LLMs. The performance of CoT fine-tuning is declined compared to the result of Direct fine-tuning. We speculate that it is because predicting evidence increases the task difficulty, even if the CoT helps improve the model's reasoning ability.

Task	Fine-tuning	Inference	Avg accuracy
State	Direct	Direct	64.84
	CoT	CoT	64.62
	Hybrid	Direct	**66.94**
	Hybrid	CoT	66.55
Trait	Direct	Direct	75.83
	CoT	CoT	**77.78**
	Hybrid	Direct	75.53
	Hybrid	CoT	75.55

Table 4: Ablation study on GLM-32k model to prove the benefits of introducing evidence on personality recognition. Direct: the model is trained or evaluated to directly generate the answer; CoT: the model is trained or evaluated to generate the evidence and then the answer; Hybrid: the model is trained on both above tasks.

State	ID F1	BERT	Claude	GPT-4	Accuracy
None	40.28	76.80	2.95	2.74	77.78
Pred	44.18	76.99	3.31	3.11	77.99
GT	**77.09**	**81.03**	**3.69**	**3.57**	**83.52**

Table 5: Average personality trait metrics when different state clues act as inputs on GLM-32k to prove the necessity of introducing the EPR-S task. None: no state clues, just dialogues; Pred: predicted state evidence from GLM-32k; GT: ground-truth state evidence.

However, for trait recognition, the performance of both Hybrid settings is similar to that of Direct settings. We speculate that the model has not been fully trained due to the limited training data. In the CoT setting, the performance has improved, implying that improving the model's reasoning ability can overcome the increase in task difficulty.

Potential benefits of the state evidence for the EPR-T task. Table 5 presents the results when different state clues act as the input, which can prove the necessity of introducing the EPR-S task. We can find that when providing the state evidence predicted by the model itself, all metrics are slightly improved. Using the state evidence as input instead of the original dialogues can avoid handling very long contexts and reduce the difficulty of the task. However, because the current model's ability to predict state evidence is poor, the improvement is very limited. When providing ground-truth state evidence, all metrics are significantly improved, as gold references rarely contain noise information. Therefore, for the trait-level task, we believe that adopting a two-stage pipeline method is a promising direction that can be attempted in the future.

6 Conclusion

In this paper, to promote the research of explainable personality recognition, we propose a novel Chain-of-Personality-Evidence (CoPE) framework, which reveals the reasoning process from specific contexts to short-term personality states to long-term personality traits. We build an explainable personality dataset based on CoPE framework, namely PersonalityEvd, which supports two explainable personality recognition tasks (EPR-S and EPR-T) that both require the model to recognize personality as well as provide corresponding evidence. Finally, we evaluate several large language models as baselines and conduct extensive experiments on both tasks. Additional human evaluations validate the quality of our constructed dataset and our analytical experiments present insights for future work. We hope that our PersonalityEvd dataset and two novel tasks can facilitate further investigation into explainable personality analysis in the community.

Limitations

There are several limitations of our PersonalityEvd dataset and modeling. First, our dataset is in Chinese. Though we translated the dataset from Chinese into English, we think it is better to directly construct a corresponding English dataset considering data quality. Second, we acknowledge that our dataset is small-scale due to the high annotation costs of two-level personality evidence. However, we plan to expand the dataset scale in the future. Third, as our dataset is constructed around characters, standardized and psychometrically validated personality tests or self-report questionnaires could not be applied to provide objective evidence for the explanations. Last, we just essentially fine-tune the LLMs as baselines, since we are more focused on the contributions of the dataset construction and building the benchmark. We leave the development of advanced models as future work.

Ethics Statements

We choose the dialogues from the CPED dataset, which has been released to the public. There are no intellectual property disputes for our data source. Human annotation is carried out by workers we employ and we pay each worker $12 per hour, which is a fair and reasonable hourly wage in Beijing. Besides, due to the subjectivity of manual annotation, our dataset may contain biased opinions.

Acknowledgements

We thank all reviewers for their insightful comments and suggestions. This work was supported by the National Natural Science Foundation of China (No. 62072462).

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

AI Anthropic. 2024. The claude 3 model family: Opus, sonnet, haiku. *Claude-3 Model Card*.

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. In *HCI International 2017–Posters' Extended Abstracts: 19th International Conference, HCI International 2017, Vancouver, BC, Canada, July 9–14, 2017, Proceedings, Part I 19*, pages 19–29. Springer.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Joan-Isaac Biel and Daniel Gatica-Perez. 2012. The youtube lens: Crowdsourced personality impressions and audiovisual analysis of vlogs. *IEEE Transactions on Multimedia*, 15(1):41–55.

Katharine C Briggs. 1976. *Myers-Briggs type indicator*. Consulting Psychologists Press Palo Alto, CA.

David F Caldwell and Jerry M Burger. 1998. Personality characteristics of job applicants and success in screening interviews. *Personnel Psychology*, 51(1):119–136.

Heather EP Cattell and Alan D Mead. 2008. The sixteen personality factor questionnaire (16pf). *The SAGE handbook of personality theory and assessment*, 2:135–159.

Oya Celiktutan, Efstratios Skordos, and Hatice Gunes. 2017. Multimodal human-human-robot interactions (mhhri) dataset for studying personality and engagement. *IEEE Transactions on Affective Computing*, 10(4):484–497.

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*.

Gordon Claridge and Caroline Davis. 2013. *Personality and psychological disorders*. Routledge.

Marco Cristani, Alessandro Vinciarelli, Cristina Segalin, and Alessandro Perina. 2013. Unveiling the multimedia unconscious: Implicit cognitive processes and multimedia content analysis. In *Proceedings of the 21st ACM international conference on Multimedia*, pages 213–222.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. GLM: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 320–335, Dublin, Ireland. Association for Computational Linguistics.

Hugo Jair Escalante, Isabelle Guyon, Sergio Escalera, Julio Jacques, Meysam Madadi, Xavier Baró, Stephane Ayache, Evelyne Viegas, Yağmur Güçlütürk, Umut Güçlü, et al. 2017. Design of an explainable machine learning challenge for video interviews. In *2017 International joint conference on neural networks (IJCNN)*, pages 3688–3695. IEEE.

Hans Jurgen Eysenck. 2012. *A model for personality*. Springer Science & Business Media.

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

William Fleeson and Eranda Jayawickreme. 2015. Whole trait theory. *Journal of research in personality*, 56:82–92.

Matej Gjurković, 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.

Ao Guo, Ryu Hirai, Atsumoto Ohashi, Yuya Chiba, Yuiko Tsunomori, and Ryuichiro Higashinaka. 2024. Personality prediction from task-oriented and open-domain human–machine dialogues. *Scientific Reports*, 14(1):3868.

Namgyu Ho, Laura Schmid, and Se-Young Yun. 2023. Large language models are reasoning teachers. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14852–14882, Toronto, Canada. Association for Computational Linguistics.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Yu Ji, Wen Wu, Hong Zheng, Yi Hu, Xi Chen, and Liang He. 2023. Is chatgpt a good personality recognizer? a preliminary study. *arXiv preprint arXiv:2307.03952*.

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

Alam Sher Khan, Hussain Ahmad, Muhammad Zubair Asghar, Furqan Khan Saddozai, Areeba Arif, and Hassan Ali Khalid. 2020. Personality classification from online text using machine learning approach. *International journal of advanced computer science and applications*, 11(3):460–476.

Shiyang Li, Jianshu Chen, yelong shen, Zhiyu Chen, Xinlu Zhang, Zekun Li, Hong Wang, Jing Qian, Baolin Peng, Yi Mao, Wenhu Chen, and Xifeng Yan. 2024a. Explanations from large language models make small reasoners better. In *2nd Workshop on Sustainable AI*.

Zhen Li, Xiaohan Xu, Tao Shen, Can Xu, Jia-Chen Gu, and Chongyang Tao. 2024b. Leveraging large language models for nlg evaluation: A survey. *arXiv preprint arXiv:2401.07103*.

Cynthia CS Liem, Markus Langer, Andrew Demetriou, Annemarie MF Hiemstra, Achmadnoer Sukma Wicaksana, Marise Ph Born, and Cornelius J König. 2018. Psychology meets machine learning: Interdisciplinary perspectives on algorithmic job candidate screening. *Explainable and interpretable models in computer vision and machine learning*, pages 197–253.

Navonil Majumder, Soujanya Poria, Alexander Gelbukh, and Erik Cambria. 2017. Deep learning-based document modeling for personality detection from text. *IEEE Intelligent Systems*, 32(2):74–79.

Robert R McCrae and Oliver P John. 1992. An introduction to the five-factor model and its applications. *Journal of personality*, 60(2):175–215.

Gelareh Mohammadi and Alessandro Vinciarelli. 2012. Automatic personality perception: Prediction of trait attribution based on prosodic features. *IEEE Transactions on Affective Computing*, 3(3):273–284.

Sumiya Mushtaq and Neerendra Kumar. 2022. Text-based automatic personality recognition: Recent developments. In *Proceedings of Third International Conference on Computing, Communications, and Cyber-Security: IC4S 2021*, pages 537–549. Springer.

Cristina Palmero, Javier Selva, Sorina Smeureanu, Julio Junior, CS Jacques, Albert Clapés, Alexa Moseguí, Zejian Zhang, David Gallardo, Georgina Guilera, et al. 2021. Context-aware personality inference in dyadic scenarios: Introducing the udiva dataset. In

Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pages 1–12.

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

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

Tim Polzehl, Sebastian Möller, and Florian Metze. 2010. Automatically assessing personality from speech. In *2010 IEEE fourth international conference on semantic computing*, pages 134–140. IEEE.

Víctor Ponce-López, Baiyu Chen, Marc Oliu, Ciprian Corneanu, Albert Clapés, Isabelle Guyon, Xavier Baró, Hugo Jair Escalante, and Sergio Escalera. 2016. Chalearn lap 2016: First round challenge on first impressions-dataset and results. In *Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part III 14*, pages 400–418. Springer.

Francisco Rangel, Fabio Celli, Paolo Rosso, Martin Potthast, Benno Stein, Walter Daelemans, et al. 2015. Overview of the 3rd author profiling task at pan 2015. In *CLEF2015 Working Notes. Working Notes of CLEF 2015-Conference and Labs of the Evaluation forum.* Notebook Papers.

Francisco Rangel, Paolo Rosso, Ben Verhoeven, Walter Daelemans, Martin Potthast, and Benno Stein. 2016. Overview of the 4th author profiling task at pan 2016: cross-genre evaluations. In *Working Notes Papers of the CLEF 2016 Evaluation Labs. CEUR Workshop Proceedings/Balog, Krisztian [edit.]; et al.*, pages 750–784.

Donald A Redelmeier, Umberin Najeeb, and Edward E Etchells. 2021. Understanding patient personality in medical care: five-factor model. *Journal of General Internal Medicine*, 36:2111–2114.

Brent W Roberts. 2009. Back to the future: Personality and assessment and personality development. *Journal of research in personality*, 43(2):137–145.

Dairazalia Sanchez-Cortes, Oya Aran, Dinesh Babu Jayagopi, Marianne Schmid Mast, and Daniel Gatica-Perez. 2013. Emergent leaders through looking and speaking: from audio-visual data to multimodal recognition. *Journal on Multimodal User Interfaces*, 7:39–53.

Yisi Sang, Xiangyang Mou, Mo Yu, Dakuo Wang, Jing Li, and Jeffrey Stanton. 2022. MBTI personality prediction for fictional characters using movie scripts. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6715–6724, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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.

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.

Ben Verhoeven, Walter Daelemans, and Barbara Plank. 2016. Twisty: a multilingual twitter stylometry corpus for gender and personality profiling. In *Proceedings of the Tenth international conference on language resources and evaluation (LREC'16)*, pages 1632–1637.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, et al. 2022. Glm-130b: An open bilingual pre-trained model. *arXiv preprint arXiv:2210.02414*.

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.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Appendix

A Facets and Items of The BFI-2 Scale

The description of *personality statistics* source from the BFI-2 Scale (Soto and John, 2017). Figure 5 presents the facets and items of each Big-Five dimensions.

B Prompt for GPT-4 Pre-annotation

As shown in Figure 6, we design the prompt for GPT-4 pre-annotation.

C Word Clouds of Left Dimensions

We show the word clouds of the remaining three Big-Five dimensions: extraversion, agreeableness, and neuroticism in Figure 7. We filtered out stop words to highlight the key points of the natural language reasoning process.

D Implementation Details

For both tasks, we use the LoRA (Hu et al., 2021) method to fine-tune GLM-32k and Qwen-32k parameters efficiently. LoRA fine-tuning is a popular method to reduce the dependency on many training resources and yield high-quality results. We use 2*A6000 GPUs for our experiments. For the EPR-S task, the model is trained for 15 epochs with the learning rate set as 1e-4 and the batch size as 48. For the EPR-T task, we use the same training epoch and learning rate. Due to the long context of around 30 dialogues, we set the batch size as 8. To obtain more stable predictions, we make the hyper-parameter *do_sample* false to use greedy decoding at inference time.

E Case study

Figure 8 presents a case of the ablation study. In the None setting, ChatGLM3-6B-32K is almost unable to directly analyze useful personality clues from multiple dialogues, and these clues even contain contradictory content. Comparing the results under the Pred setting and the GT setting, we find that the ChatGLM3-6B-32K is relatively good at inferring trait evidence based on state evidence, but the model still faces great challenges in analyzing state evidence based on one dialogue.

Big-Five Dim	High	Low
Openness	***Intellectual Curiosity*** - Is curious about many different things. - Is complex, a deep thinker. ***Aesthetic Sensitivity*** - Is fascinated by art, music, or lterature. - Values art and beauty. ***Creative Imagination*** - Is inventive, finds clever ways to do things. - Is original, comes up with new ideas.	***Intellectual Curiosity*** - Avoids intellectual, philosophical discussions. - Has little interest in abstract ideas. ***Aesthetic Sensitivity*** - Has few artistic interests. - Thinks poetry and plays are boring. ***Creative Imagination*** - Has little creativity. - Has difficulty imagining things.
Conscientiousness	***Organization*** - Is systematic, likes to keep things in order. - Keeps things neat and tidy. ***Productiveness*** - Is efficient, gets things done. - Is persistent, works until the task is finished. ***Responsibility*** - Is dependable, steady. - Is reliable, can always be counted on.	***Organization*** - Tends to be disorganized. - Leaves a mess, doesn't clean up. ***Productiveness*** - Tends to be lazy. - Has difficulty getting started on tasks. ***Responsibility*** - Can be somewhat careless. - Sometimes behaves irresponsibly.
Extraversion	***Sociability*** - Is outgoing, sociable. - Is talkative. ***Assertiveness*** - Has an assertive personality. - Is dominant, acts as a leader. ***Energy Level*** - Is full of energy. - Shows a lot of enthusiasm.	***Sociability*** - Tends to be quiet. - Is sometimes shy, introverted. ***Assertiveness*** - Finds it hard to influence people. - Prefers to have others take charge. ***Energy Level*** - Rarely feels excited or eager. - Is less active than other people.
Agreeableness	***Compassion*** - Is compassionate, has a soft heart. - Is helpful and unselfish with others. ***Respectfuness*** - Is respectful, treats others with respect. - Is polite, courteous to others. ***Trust*** - Has a forgiving nature. - Assumes the best about people.	***Compassion*** - Feels little sympathy for others. - Can be cold and uncaring. ***Respectfuness*** - Starts arguments with others. - Is sometimes rude to others. ***Trust*** - Tends to find fault with others. - Is suspicious of others' intentions.
Neuroticism	***Anxiety*** - Can be tense. - Worries a lot. ***Depression*** - Often feels sad. - Tends to feel depressed, blue. ***Emotional Volatility*** - Is moody, has up and down mood swings. - Is temperamental, gets emotional easily.	***Anxiety*** - Is relaxed, handles stress well. - Rarely feels anxious or afraid. ***Depression*** - Stays optimistic after experiencing a setback. - Feels secure, comfortable with self. ***Emotional Volatility*** - Is emotionally stable, not easily upset. - Keeps their emotions under control.

Figure 5: Facets and Items of The BFI-2 Scale. (Dim: dimension)

```
## Role : [Master of Big Five Personality Theory]
## Profile :
- language: Chinese
- description: You are a master of the Big Five personality theory and are familiar with the specific meaning and characteristics of the five Big Five personality dimensions.
## Definition:
1. The Big Five personality theory summarizes human personality into five major dimensions: openness, conscientiousness, extraversion, agreeableness, and neuroticism
2. The definition of {target BF dim}: {BF dim definition}
## Goals :
First summary the contents that reflect the {target BF dim} level of {speaker} , and then judge the {target BF dim} level.
## Constraints :
1. Analyze one dialogue at a time
2. The requirements of [Workflow] step 2:
2.1 First analyze the behaviors, thoughts or feelings of {speaker} in the dialogue:
  - The analysis should be complete, comprehensive and detailed, and try to use the original words in the dialogue
  - Do not output what is unrelated to the {target BF dim} level
2.2 Then find the corresponding {target BF dim} characteristics that the behaviors, thoughts or feelings reflect
  - The characteristics must match the [Definition]
  - If there are no contents that match the characteristics in the [Definition], the {target BF dim} level should be judged as "uncertain"
3. The level choices of [Workflow] step 3:
 - uncertain: The dialogue does not show high or low level characteristics of {target BF dim}
 - high: The dialogue shows high-level characteristics of {target BF dim}
 - low: The dialogue shows low-level characteristics of {target BF dim}
4. The requirements of [Workflow] step 4:
 - If you judge the{target BF dim} level is "uncertain", the utterance id will directly output "none"
 - Output only the id numbers of the most relevant utterances, separated by ","
 - Do not output the utterance content
## Skills:
1. You have professional knowledge of the Big Five personality theory
2. You have a strong ability to understand dialogue context
3. You are a psychology expert
## OutputFormat:
[Dialogue 1]
- analysis:
- level:
- utterance ids:
[Dialogue 2]
...
[Dialogue 3]
...
## Workflow :
1. Analyze the participants and topics of the dialogue
2. Analyze the behaviors, thoughts or feelings that reflect the {target BF dim} level of {speaker} and corresponding characteristics
3. Find the utterance ids corresponds to the behaviors, thoughts or feelings analyzed in step 2
4. Judge the {target BF dim} level of {speaker}
5. Output according to [OutputFormat]
## Inputs :
{dialogues}
## Initialization :
Please analyze the dialogue data in [Inputs] of {speaker}, follow the [Definition], and execute according to the [Workflow].
```

Figure 6: The prompt for GPT-4 pre-annotation. (BF: Big-Five, dim: dimension)

(a) High Ext. (b) Low Ext. (c) High Agre. (d) Low Agre. (e) High Neu. (f) Low Neu.

Figure 7: The word clouds of personality state of the remaining three Big-Five dimensions. (Ext.: Extraversion, Agr.: Agreeableness, and Neu.: Neuroticism)

State Setting	Results of The Neuroticism Dimension
None (2, 1, 1, wrong)	In terms of ***anxiety***, when Qi Dasheng is questioned by Xiao Ai, she nervously asks the other party's name (19), expresses concern about Li Sanmei's leaving Beijing (20), and is disgusted with Xiao Ai's request (21); but when Li Sanmei is leaving Beijing, she shows a calm attitude (6). This shows that Qi Dasheng is relatively calm, good at dealing with pressure, and rarely feels anxious or afraid. In terms of ***depression***, Qi Dasheng's performance in all dialogues does not reflect this aspect of personality characteristics. In terms of ***emotional volatility***, Qi Dasheng is dissatisfied with Xiao Ai's request and decides to solve it himself (22). This shows that Qi Dasheng is moody and gets emotional easily. ***In summary***, it is judged that Qi Dasheng's trait level is **high** in neuroticism.
Pred (4, 3, 1, wrong)	In terms of ***anxiety***, Qi Dasheng is worried about the relationship between Yu Guo and Xia Xiaobai, his own emotional life, being deceived by others, and premature ageing (1, 3, 4, 5, 14, 15, 18, 19, 21, 22). This shows that Qi Dasheng is very nervous and often worried about many things. In terms of ***depression***, Qi Dasheng never bullies others, especially women (9). This shows that Qi Dasheng is satisfied with himself. In terms of ***emotional volatility***, Qi Dasheng angrily refutes that Yu Guo is morally corrupt (2, 23). This shows that Qi Dasheng is moody and gets emotional easily. ***In summary***, it is judged that Qi Dasheng's trait level is **high** in neuroticism.
GT (4, 4, 4, right)	In terms of ***anxiety***, Qi Dasheng is very nervous about Guo going to pick up Xia Xiaobai, Xiao Ai saying Qi Dasheng is a rat, going to see rabbits late at night, the attitude of Xiao Ai's parents when he lives with Xiao Ai, and his premature ageing (1, 3, 14, 17, 18, 20, 21). This shows that Qi Dasheng is very nervous, often worried about many things. In terms of ***depression***, Qi Dasheng believes in his own judgement that the tourism and vacation industry will develop rapidly in the future (7, 12). This shows that Qi Dasheng feels satisfied with himself. In terms of ***emotional volatility***, Qi Dasheng is not angry when Xiao Ai says Qi Dasheng is a rat (4) and prematurely ageing (19), and even advises Xiao Ai not to get angry (5). This shows that Qi Dasheng is very emotionally stable and not easily angry. ***In summary***, it is judged that Qi Dasheng's trait level is **low** in neuroticism.
Gold	In terms of **anxiety**, Qi Dasheng is worried about Yu Guo picking up Xia Xiaobai (1, 3) and is very nervous about his relationship with Xiao Ai (14, 17, 18, 20). This shows that Qi Dasheng is very nervous and often worried about many things. In terms of **depression**, Qi Dasheng is very firm in his judgement that "the tourism and holiday industry will also lead a new lifestyle" (7). This shows that Qi Dasheng feels satisfied with himself. In terms of ***emotional volatility***, Qi Dasheng almost never gets angry when facing other people's doubts and criticisms (4, 5, 19), but sometimes he is very touched by shocking scenes (21) and he gets angry when Yu Guo eats indecently (6). This shows that Qi Dasheng is relatively emotionally stable and not easily angry. ***In summary***, it is judged that Qi Dasheng's trait level is **low** in neuroticism.

Figure 8: A case of the ablation study on the potential benefits of the state evidence for the EPR-T task. The values under each setting (score1, score2, score3, correctness) refer to the GPT-4-Turbo evaluation scores of each facet and whether the trait label is correctly predicted. None: no state clues, just dialogues; Pred: predicted state evidence from GLM-32k; GT: ground-truth state evidence.