2025 International Conference on Intelligent Computing
July 26-29, Ningbo, China
https://www.ic-icc.cn/2025/index.php

DKE: LLM-Based Domain Knowledge Enhancement for Comprehensible Personality Detection

Yutian Zhang[1,2], Conghui Zheng[1,2], and Li Pan[1,2] (✉)

[1] School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, Shanghai 200240, China
panli@sjtu.edu.cn
[2] Shanghai Key Laboratory of Integrated Administration Technologies for Information Security, Shanghai 200240, China

Abstract. Personality detection aims to identify personality traits of individuals based on social media posts. Many existing methods map words to psycholinguistic categories by utilizing Linguistic Inquiry and Word Count (LIWC) to introduce the domain knowledge required for personality detection. However, the categories of LIWC are static and lack a direct connection to personality detection, thereby limiting the model's ability to effectively leverage domain-specific knowledge. Trained on massive amounts of data, large language models (LLMs) possess extensive world knowledge, especially domain-specific knowledge that is beneficial to comprehending personality taxonomies. Inspired by this, we propose an LLM-based domain knowledge enhanced model to capture the implicit psycholinguistic knowledge in posts, achieving more accurate and comprehensible personality detection. Specifically, to leverage the LLM's comprehensive knowledge base, we first input the posts into the LLM to obtain personality judgments and corresponding rationales, which are derived based on the core characteristics of each MBTI dimension. To better incorporate personality-related knowledge, the proposed model then conducts feature interactions between the rationales and the posts, generating text representations that better reflect personality traits. After that, the model adaptively adjusts the weights of the interactive features and aggregates them with the semantic features to form the final representations, based on which personality detection is performed. Experimental results on real-world datasets demonstrate the proposed model effectively improves the quality of user personality representation and outperforms baseline methods.

Keywords: Large Language Model · Personality Detection · Knowledge Enhancement.

1 Introduction

Personality refers to the dynamic integration of a person's behavior patterns, including conscious behaviors and unconscious behavior patterns [1]. Personality traits are defined by long-established personality theories such as the Myers-Briggs Type Indicator (MBTI) taxonomy [2]. Traditional questionnaire-based personality assessment methods are both time-consuming and laborious. The rise of social media has generated a

vast number of posts that track users' behavior, providing new possibilities for personality detection [3]. Furthermore, it was suggested that analyzing a person's writing provides a more objective assessment of one's personality than the traditional questionnaires [4]. Nowadays, personality detection aims to detect the personality traits of users based on their social media posts, which has expanded to massive applications such as recommendation systems [5], dialogue systems [6], and psychological treatments [7].

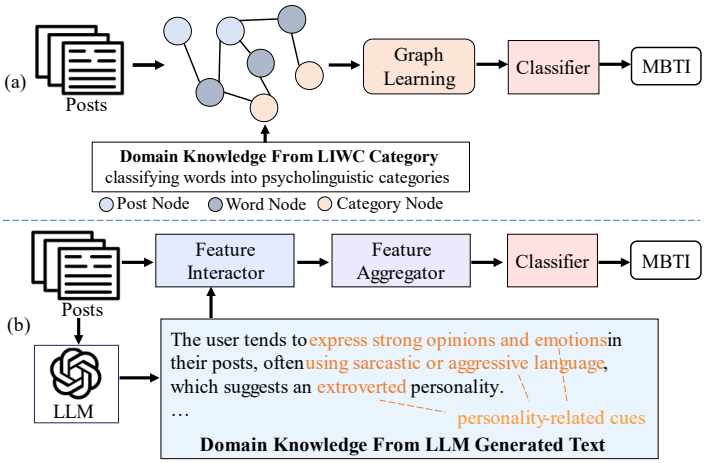


Fig. 1. Comparison with existing domain-knowledge enhanced personality detection methods. (a) Existing methods represent post features as nodes in a graph neural network and incorporate psycholinguistic knowledge from LIWC. (b) Our proposed method leverages an LLM to generate personality-related cues to supplement the required domain knowledge.

The existing personality detection studies [8] are mainly divided into traditional machine learning and deep learning methods. Traditional machine learning methods in personality detection heavily rely on manually designed features, which is not only time-consuming but also prone to missing valuable insights [9]. Recently, deep learning methods have dominated personality detection research because of their outstanding ability to learn textual representations from data automatically. Most existing deep learning methods directly aggregate semantic information from multiple posts to depict an overall personality profile for each user. However, they ignore the importance of domain-specific knowledge, which limits their performance. The input posts cover various topics, and some may offer very few personality clues. Directly aggregating these posts based on their contextual representations may inevitably introduce noise [10]. Moreover, the difficulty of collecting labeled data makes the user personality representation unsatisfactory [11]. To address the aforementioned challenges, existing methods have proposed explicitly introducing domain knowledge into the models to capture critical personality cues [10]. As shown in Fig. 1(a), existing methods enhance personality detection by introducing psycholinguistic knowledge of Linguistic Inquiry and Word Count (LIWC) [12], which is a tool based on a fixed, static dictionary for text analysis that maps words to psychologically meaningful categories. However, the static classification may ignore the ambiguity of words in different contexts and many general

categories are not directly related to personality, offering limited guidance for personality detection. For example, in two posts with the texts “love being stuck in traffic.” and “love you very much.” the words ‘I’ and ‘love’ appear in both. One uses sarcasm, where ‘love’ means dislike, while the other genuinely expresses affection. However, their LIWC classification remains the same, unaffected by context.

Large language models present new opportunities for addressing the aforementioned challenge, as their training on vast amounts of data endows them with extensive world knowledge [13], including knowledge related to personality classification. PsyCoT [14] found through experiments that directly using LLMs for personality detection is less effective than fine-tuned small language models (SLMs). Although LLMs fail in direct personality detection, their extensive world knowledge and understanding of personality theories enable them to provide more comprehensive insights into users’ personalities. Moreover, the strong comprehension abilities of LLMs allow them to recognize various rhetorical devices in posts, such as sarcasm and metaphor, thereby facilitating a more accurate analysis of users’ personalities. Additionally, some researchers have leveraged LLMs to enhance the classification capabilities of SLMs, achieving promising performance [15, 16]. Inspired by these methods, this paper is dedicated to utilizing large language models to generate user personality analyses from posts and leveraging this as domain knowledge to interact and integrate with SLM features, thereby addressing the issue of insufficient domain knowledge in personality detection with the SLM. As shown in Fig. 1 (b), our method employs LLM to generate personality traits implied in the posts based on the characteristics of different personality dimensions, supplementing the psychological knowledge of personality detection.

In this paper, we propose an LLM-based domain knowledge enhanced model for comprehensible personality detection. The domain knowledge required for personality detection is related to the psychological definition of personality, with a different focus on different personality dimensions in the same text. Therefore, domain knowledge varies across different dimensions. To leverage the LLM’s comprehensive knowledge base, we first input the posts into the LLM to obtain personality judgments and corresponding rationales, which are derived based on the characteristics of each MBTI dimension. The rationales given by LLM include psychological knowledge and the user’s personality tendencies underlying posts, which is the domain knowledge that small models lack. Subsequently, a dual cross-attention mechanism is leveraged to incorporate domain knowledge into the small model, generating interactive features that contain richer personality-related information. Moreover, to mitigate the effect of incorrect analysis by the LLM, a rationale credibility evaluator is introduced to adaptively adjust the weight of the interactive features. In addition, we also utilize the attention mechanism to obtain the semantic representation of posts, ensuring that the features implicit in the semantic information are not overlooked. Finally, a more accurate and comprehensible personality representation is obtained by adaptively aggregating the interactive features and the original semantic features.

The main contributions of this paper are as follows:

- We propose an LLM-based model to enhance domain knowledge in personality detection, which leverages the LLM to obtain more synthetical personality analysis and

use it to guide the representation of the small model, thereby facilitating a more comprehensible representation of user personalities.
- To alleviate the impact of uncertainty in LLM judgments, our proposed model adaptively adjusts the weights of the features obtained from interactions with the generated text through a rationale credibility evaluator and aggregates them with the original semantic features, achieving a more reliable personality representation.
- Experimental results on real-world datasets demonstrate the proposed model outperforms baseline methods, which shows the effectiveness of our method.

2 Related Work

2.1 Personality Detection

Traditional machine learning methods rely on manually designed features, such as using LIWC to extract psycholinguistic features or employing the Bag-of-Words model [17] to obtain statistical text features. Traditional methods extract features and then use support vector machines(SVM) [18] and extreme gradient boosting(XGBoost) [19] for classification, but they are unable to effectively represent the diverse social media texts. Due to the significant performance improvement of deep learning over traditional machine learning, some deep learning methods, such as BiLSTM [20] and AttRCNN [3], have been applied to personality detection.

Recently, pre-trained language models have demonstrated strong capabilities in text semantic representation, which has also accelerated the development of personality detection. Some works focus on improving text semantic aggregation to better represent users' personality traits after employing pre-trained models such as BERT [21] for text representation. Transformer-MD [22] proposed a dimension attention mechanism to generate a trait-specific representation for each personality dimension and put together information in different posts to depict a personality profile for each user. D-DGCN [23] uses a deep graph convolutional network to try to piece together information from multiple posts into an overall user profile. However, domain knowledge such as psycholinguistic knowledge is not fully utilized in these methods, resulting in limited model performance. Some methods recognize the importance of domain knowledge in personality detection. For example, TrigNet [10] injects structural psycholinguistic knowledge from LIWC by constructing a heterogeneous tripartite graph and PS-GCN [24] integrating psychological knowledge based on the LIWC dictionary and sentiment semantic features through Graph Convolution Networks. However, these methods rely on static dictionaries for domain knowledge, which fail to directly capture the relationship between text semantics and personality traits, thereby limiting their effectiveness. Different from these methods, our method generates user-specific personality analysis by leveraging the massive knowledge of the LLM, and the psychological knowledge contained in it can enhance domain information.

2.2 Domain Knowledge Enhancement

Many text classification tasks have benefited from supplementing external knowledge. Shahabikargar et al. [25] leverage the domain knowledge and expertise in cognitive science to build a domain-specific knowledge base for mental health disorder concepts and patterns and leverage database knowledge to improve the identification of mental disorder patterns. Guo et al. [26] develop a depression lexicon based on domain knowledge of depression and a sentiment lexicon, and experimental results indicate that the depression domain lexicon features improve classification performance and fusing these features based on their correlations can further enhance prediction effectiveness. Furthermore, FinSent-KDR [27] utilizes a better domain-specific word embedding representation to improve the performance of the financial sentiment analysis model. Therefore, it is crucial to incorporate domain knowledge into the text classification related to psychology and sentiment. Given the vast knowledge storage of LLMs, some methods introduce domain knowledge through LLMs. For example, ARG [16] leverages the LLM to generate analyses from the perspectives of textual descriptions and commonsense to enhance fake news detection. OKI [28] collaboratively leverages open-domain knowledge by two LLM-based agents and demonstrates superior performance compared to some solid baselines. Different from these studies, personality detection is a multi-document multi-label classification task that requires obtaining a comprehensible personality profile of a user from multiple topic-independent posts [10].

3 Method

According to previous studies [10, 23], personality detection can be defined as a multi-document multi-label classification task. Given a set of posts $P = \{p_1, p_2, \ldots, p_N\}$ from a certain user, where $p_i = \{w_{i1}, w_{i2}, \ldots, w_{iL}\}$ represents the i-th post with L tokens, the personality detection model needs to predict the user's personality traits $Y = \{y_1, y_2, \ldots, y_T\}$, where each $y_t \in \{0, 1\}$ and T is the number of personality traits. The value of T is defined as 4 when using the MBTI taxonomy, whereas it is set to 5 under the Big Five taxonomy. Since the dataset used for validation is the MBTI dataset, T is set to 4. The prediction of each personality dimension can be regarded as an individual classification task [23]. Therefore, we conduct personality detection for each dimension separately using the same framework. As shown in Fig. 2, our method is divided into three parts: representation, feature interaction, and classification.

3.1 Representation

We conducted a study on the performance of LLMs in personality detection. Yang et al. [14] demonstrate that directly using ChatGPT[1] for personality detection performs poorly, and although the chain-of-thought designed from psychological questionnaires improves performance, the resulting performance is still inferior to that of a carefully

[1] https://chat.openai.com/

designed small model. Our experiments show that directly using Llama[2] for personality detection yields poor performance, with specific results documented in the experiment section. We input the posts into the Llama, prompting it to provide the personality judgments of users across the four dimensions of the MBTI as well as the rationales for those judgments. Through analysis of the output texts from Llama, we found that Llama is easily influenced by users' writing characteristics on social media. The language used on Twitter shows specific stylistic characteristics in terms of the use of exclamation marks, sentence structure, and grammar, thus making everyone seem more extroverted than they are [29]. This may be the rationale for the misjudgment of the LLMs. Additionally, the disorderliness of posts can lead the model to perceive users as more spontaneous and less inclined toward an organized lifestyle, thereby increasing the likelihood of the model classifying the user as having a perceiving personality.

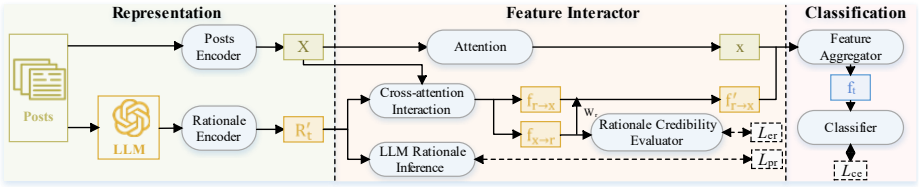


Fig. 2. The architecture of the DKE model. In the DKE, the posts and LLM rationales are respectively encoded into X and R_t'. The rationales and the original posts engage in comprehensive feature interaction through cross-attention interaction, LLM rationale inference, and rationale credibility evaluator to obtain the interactive features $f_{r\to x}'$. The interactive features are finally aggregated with the semantic features of the posts to obtain the user personality representation f_t for personality detection.

As illustrated in Fig. 3, the LLM incorrectly classifies users who reply to others on social media as extroverted, when in fact, an introverted person may also frequently engage in replies on social media. Although the LLM misjudged users' personalities, the provided rationales align well with the psychological characteristics associated with these personality dimensions. As shown in Fig. 3, the personality characteristics generated by the LLM align with the psychological definition of "ENFP". This suggests that the LLM possesses domain knowledge from psychology, which is precisely what small models lack. Therefore, we propose leveraging the judgments and rationales generated by LLMs to enhance the domain knowledge of the small model. To facilitate the extraction of judgments and rationales, the prompt template requires the model's output to follow a format similar to the one shown in Fig. 3. According to previous research [22], the posts are unordered, and the input sequence of the posts affects the generalization ability of the models. Therefore, we incorporated information about the disorder of the posts into the prompt template, as follows:

You are an AI assistant who specializes in text analysis. You will complete a text analysis task. The task is as follows: according to a set of posts written by an author, please predict the MBTI personality of the author and give the rationales. Please answer according to the following format: {format}. Please do not provide any other text

[2] https://llama.meta.com/

besides the template. The following is the text posted by the user and posts are separated by "|||". The posts are not in order, so you do not know the order in which they were posted: {posts}.

We use two separate BERT models as the posts and rationale encoders to obtain semantic representations. To facilitate batch processing and parallel computation, the user posts are padded to the same length. For a specific user, the user's N posts are encoded using the posts encoder to obtain their embeddings, resulting in a matrix $P_i \in \mathbb{R}^{L_{max} \times d}$, where L_{max} is the padded length of the post, and d is the embedding dimension. The embeddings of all N posts are stacked into a matrix $X \in \mathbb{R}^{N \times L_{max} \times d}$. The matrix X represents the combined embeddings of the user's posts, where each post is represented as a sequence of embeddings.

Prompt

You are an AI assistant who specializes in text analysis. You will complete a text analysis task. The task is as follows: …

Input

'to go back to topic, here is my opinion about the game i posted in another forum :) maybe little SPOILERS ahead! …'

Output

The personality of the user in the I/E dimension is E. The reason is that the author frequently engages in conversations, responds to others, and shares their thoughts and opinions, **indicating a preference** for extraverted interactions.
The personality of the user in the S/N dimension is N. The reason is that the author often uses imaginative and hypothetical scenarios, makes connections between seemingly unrelated things, and explores abstract ideas, **suggesting a preference** for intuitive thinking.
The personality of the user in the T/F dimension is F. The reason is that the author frequently expresses emotions, empathizes with others, and values personal relationships, **indicating a preference** for feeling-based decision-making.
The personality of the user in the J/P dimension is P. The reason is that the author tends to be flexible, adaptable, and spontaneous, often changing topics or responding to others in a playful and carefree manner, **suggesting a preference** for perceiving.

Fig. 3. Example of using LLM for personality detection: LLM incorrectly predicts a user with ISFJ personality as ENFP. The rationales given by the LLM contain the behavioral tendencies of the corresponding personality type, which is of significance to personality detection.

To save API resources and time, all posts are input into the LLM simultaneously, and the rationales are generated based on all of the user's posts. The rationale text of dimension t is encoded by the rationale encoder to obtain its semantic representation $R_t \in \mathbb{R}^{L_{max} \times d}$. To facilitate feature interaction, the rationale text is replicated N times to match the same dimensions as the posts, creating a new matrix $R_t' \in \mathbb{R}^{N \times L_{max} \times d}$.

3.2 Feature Interactor

The steps of the feature interaction module are designed to provide rich interactions between the post and the rationales, utilizing the domain knowledge underlying the

rationales to guide the model's attention and improve personality detection. To achieve this goal, DKE consists of three modules, which are described in detail as follows:

Cross-attention Interaction: To achieve comprehensive information exchange between posts and rationales, we introduce a dual cross-attention mechanism as the cross-attention interaction module to encourage feature interactions. The cross-attention can be described as:

$$CA(Q, K, V) = softmax\left(Q' \cdot K'^{T} / \sqrt{d}\right)V' \tag{1}$$

where $Q' = W_Q Q$, $K' = W_K K$, and $V' = W_V V$. d is the dimensionality. Given representations of the posts X and the rationale matrix R'_t of dimension t, the process is:

$$f_{r \to x} = AvgPool(CA(R'_t, X, X)) \tag{2}$$

$$f_{x \to r} = AvgPool(CA(X, R'_t, R'_t)) \tag{3}$$

where $AvgPool(\cdot)$ is the average pooling over the token representations outputted by cross-attention to obtain the interactive representation f.

LLM Rationale Inference: Understanding the judgment implied by the generated rationale is essential for fully leveraging the domain knowledge behind the rationales. To achieve this, we design the LLM rationale inference module, which aims to predict the LLM's judgment of the user's personality based on the given rationale. This is expected to enhance the understanding of the rationale texts. We input the rationale representation R_t of dimension t into the LLM rationale inference module, which is parameterized using a multi-layer perceptron (MLP):

$$\hat{m}_r = sigmoid(MLP(R_t)),\ L_r = CE(\hat{m}, m_r) \tag{4}$$

where m_r and $\hat{m}_r$ are respectively the LLM's claimed judgment and the prediction of this module. The loss L_r is a cross-entropy loss $\mathrm{CE}(\hat{y}, y) = -y \log(\hat{y}) - (1-y) \log(1-\hat{y})$.

Rationale Credibility Evaluator: The text generated by the LLM is not always accurate, and over-relying on its output can lead to incorrect features being assigned high weights. To enable the model to adaptively determine the trustworthiness of the rationales, we design a rationale credibility evaluator. Through this module, we assess the credibility of the LLM's outputs and adjust their weights to improve personality representation. This process consists of two stages: evaluation and reweighting. For evaluation, we input the vector $f_{x \to r}$ into the rationale credibility evaluator (parameterized by an MLP) to predict its credibility u_r. Based on the assumption that rationales provided with correct judgments are more trustworthy, we use judgment correctness as the label for rationale credibility. To match the dimension of $f_{x \to r}$, u_r is also extended accordingly, and the loss L_e is averaged over N:

$$\hat{u}_r = sigmoid(MLP(f_{x \to r})),\ L_e = \frac{1}{\mathrm{N}} CE(\hat{u}_r, u_r) \tag{5}$$

In the weight adjustment phase, we input vector $f_{x \to r}$ into the MLP to obtain a weight w_r, which is used to reweight the interactive features $f_{r \to x}$. The procedure is as follows:

$$f'_{r\to x} = w_r \cdot f_{r\to x} \tag{6}$$

After that, $f'_{r\to x}$ is averaged over N to obtain the final interactive features $f'_{r\to x}$. We also use an attention mechanism to obtain the semantic representation of the original post, and then average the representations of all posts to obtain the overall semantic representation, denoted as x.

3.3 Prediction

Based on the outputs from the last step, we now aggregate the semantic vector of the posts x and interactive features $f'_{r\to x}$ for the final prediction. For a user's personality label $y_t \in \{0,1\}$ in the dimension t, we aggregate these vectors with different weights:

$$f_t = w_x^t \cdot x + w_r^t \cdot f'_{r\to x} \tag{7}$$

where w_x^t and w_r^t are learnable parameters ranging from 0 to 1. f_t is the fusion vector, which is then fed into the MLP classifier for final prediction of personality:

$$L_{ce} = CE(MLP(f_t), y_t) \tag{8}$$

The overall loss function is the weighted sum of the loss terms described above:

$$L = L_{ce} + \beta_1 L_e + \beta_2 L_r \tag{9}$$

where β_1 and β_2 are hyperparameters used to control the weights of different losses.

4 Experiments

4.1 Datasets

Following previous studies [10, 22, 23], we choose the Kaggle[3] and Pandora[4] MBTI datasets for evaluations. Both datasets are based on the MBTI taxonomy, which classifies personality types into four characteristics: Introversion vs. Extroversion (I/E), Sensing vs. Intuition (S/N), Thinking vs. Feeling (T/F), and Perception vs. Judging (P/J). The Kaggle dataset has 8,675 users with 45-50 social media posts each, and the Pandora dataset has 9,067 users with dozens to hundreds of social media posts each. Consistent with prior studies [10, 22, 23], we remove all words that match any personality label from the posts to prevent information leaks. Table 1 shows the statistics of the two datasets. Since the two datasets are severely imbalanced, we employ the Macro-F1 metric to measure the performance. Following previous work [23], the data is split in a 60-20-20 ratio for training, validation, and testing.

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

Table 1. Statistics of the Kaggle and Pandora datasets.

Dataset	Types	Train	Validation	Test
Kaggle	I / E	4011 / 1194	1326 / 409	1339 / 396
	S / N	727 / 4478	222 / 1513	248 / 1487
	T / F	2410 / 2795	791 / 944	780 / 955
	P / J	3096 / 2109	1063 / 672	1082 / 653
Pandora	I / E	4278 / 1162	1427 / 386	1437 / 377
	S / N	610 / 4830	208 / 1605	210 / 1604
	T / F	3549 / 1891	1120 / 693	1182 / 632
	P / J	3211 / 2229	1043 / 770	1056 / 758

4.2 Baselines

We compare our model with several baselines as follows.

SVM [18] and **XGBoost** [19]: These methods first concatenate all of a user's posts into a single document and then employ SVM or XGBoost for classification, using features extracted through bag-of-words models.

BiLSTM [20]: GloVe is used to generate word embeddings, after which a Bi-directional LSTM encodes each post, with the averaged post representation serving as the user's representation for personality detection.

AttRCNN [3]: It uses a hierarchical deep neural network, combining the At-tRCNN structure and a variant of Inception, to learn deep semantic features from user posts and integrates them with statistical linguistic features to predict personality.

BERTconcat [21]: This method simply concatenates the posts of a user into a document and then inputs it into BERT.

Transformer-MD [22]: Transformer-MD leverages memory tokens from Transformer-XL to capture user information across posts and incorporates a dimension attention mechanism for trait-specific representations in multi-trait personality detection.

TrigNet [10]: TrigNet integrates structural psycholinguistic knowledge from LIWC into a heterogeneous graph and employs a flow graph attention network (GAT) to efficiently learn and aggregate post information for personality detection.

D-DGCN [23]: DDGCN firstly encodes each post using a domain-adapted BERT, and then aggregates the posts in a disorderly manner by a dynamic deep graph network.

PS-GCN [24]: PS-GCN employs Bi-LSTM for sentence sentiment representation, P-GCN and S-GCN to capture psycholinguistic dependencies and syntactic structure, and utilizes attention mechanisms to enhance feature relevance, thereby improving personality detection accuracy.

MvP [9]: MvP leverages a Multi-view MoE network to analyze user posts from various perspectives, employs User Consistency Regularization to integrate these perspectives, and optimizes training through a multi-task joint learning strategy.

LLAMA: We applied the 'llama3-70b-8192' version of Llama from Groq[5], and set the temperature to 0 to make the outputs largely deterministic for the inputs.

4.3 Implementation Details

All the deep learning models are implemented in PyTorch and trained with Adam. We employed 'bert-base-uncased' from BERT as our posts encoder and rationale encoder. The learning rate for DKE is set to 5e-4 on Kaggle and 7e-4 on Pandora, with a batch size of 128. Following previous works [10], We limit each user to a maximum of 50 posts and limit each post and rationale text in both datasets to a maximum length of 70 words. For the LLM, considering the cost of querying, we use the 'llama3-70b-8192' model from Groq to generate personality judgments and rationales and we set the temperature as 0 to ensure deterministic outputs. The weight of loss functions L_e, L_r in the DKE are grid searched.

Table 2. Overall results of our DKE and baseline models in Macro-F1 (\%) score.

Dataset	Kaggle					Pandora				
	I/E	S/N	T/F	P/J	Avg	I/E	S/N	T/F	P/J	Avg
SVM	53.34	47.75	76.72	63.03	60.21	44.74	46.92	65.37	56.32	53.34
XGBoost	56.67	52.85	75.42	65.94	62.72	45.99	48.93	63.51	55.55	53.50
BiLSTM	57.82	57.87	69.97	57.01	60.67	48.01	52.01	63.48	56.12	54.91
AttRCNN	59.74	64.08	78.77	66.44	67.26	48.55	56.19	64.39	57.26	56.60
$BERT_{concat}$	58.33	53.88	69.36	60.88	60.61	54.22	49.15	58.31	53.14	53.71
Transformer-MD	66.08	69.10	79.19	67.50	70.47	55.26	58.77	69.26	60.90	61.05
TrigNet	69.54	67.17	79.06	67.69	70.87	56.69	55.57	66.38	57.27	58.98
D-DGCN	69.52	67.19	80.53	68.16	71.35	59.98	55.52	70.53	59.56	61.40
PS-GCN	70.52	65.73	70.51	67.13	68.47	-	-	-	-	-
MvP	67.68	**69.89**	80.99	68.32	71.72	60.08	**56.99**	69.12	61.19	61.85
Llama	70.48	63.04	77.70	51.96	65.80	56.02	49.14	68.14	55.01	57.08
DKE	**72.19**	67.46	**81.01**	**68.63**	**72.32**	**62.68**	56.52	**70.62**	**61.91**	**62.93**

4.4 Overall Results

The overall results are presented in Table 2. We categorize the compared models into two groups: one consisting of existing small models based on machine learning or deep learning, and the other comprising the LLM, Llama. We extract the judgments from the text generated by Llama to calculate the Macro-F1 score. We can observe that the proposed DKE outperforms all baseline models in terms of average Macro-F1 score on the two datasets, with scores of 72.32% and 62.93%, respectively. Additionally, our model also obtains the highest macro-average F1 scores on three personality dimensions,

[5] https://groq.com/

which demonstrates the superiority of our approach. Compared to Llama, our method improves by 6.52% on the Kaggle dataset and 5.85% on the Pandora dataset. The effectiveness of our method can be attributed to the use of rationales generated by the LLM, which serve to guide the small model in constructing more precise user personality representations. This subsequently facilitates the identification of implicit personality traits underlying posts. We also observed that our approach did not achieve optimal performance on the S/N dimension, which may be due to the fact that this dimension reflects individual differences in perceiving and gathering information. As a result, it may be inherently difficult to infer this trait based solely on textual content. Nevertheless, our method outperforms other domain knowledge-enhanced approaches, such as TrigNet [10] and PS-GCN [7], demonstrating superior ability in leveraging domain knowledge for representation. Since we utilize the open-source large language model LLaMA, which currently offers free API access, it does not significantly increase the cost of model training. On the small model, we mitigate the impact of potential misjudgments from the large language model by adaptively adjusting the weights of interaction vectors. This demonstrates the practicality and reliability of our approach.

Table 3. Results of ablation study on Macro-F1 on the Kaggle dataset.

Methods	Kaggle				
	I/E	S/N	T/F	P/J	Avg
$DKE_{w/o\ rationale\ credibility\ evaluator}$	71.02	61.64	79.91	66.72	69.82
$DKE_{w/o\ LLM\ rationale\ inference}$	**72.31**	64.97	80.31	66.07	70.92
$DKE_{w/o\ Attention}$	70.80	64.10	79.89	64.11	69.73
$DKE_{w/o\ All}$	66.31	63.50	79.07	64.86	68.44
DKE_{concat}	68.92	61.52	79.96	62.48	68.22
DKE	72.19	**67.46**	**81.01**	**68.63**	**72.32**

4.5 Ablation Study

To verify the importance of each component in our DKE model, we conducted a series of ablation experiments on the Kaggle dataset, and the results are shown in Table 3. First, we remove the rationale credibility evaluator, LLM rationale inference, and attention respectively to observe the impact of these three parts on the overall performance of the model. We found that removing the LLM rationale inference part has the smallest impact, which indicates that optimizing the textual representation of the rationales provides limited gain. Additionally, we observed that removing either the rationale credibility evaluator or the attention mechanism led to nearly identical performance, suggesting that both the semantic representation and the interaction representation of the posts are equally crucial to the model's effectiveness. After removing the three parts separately, the performance of the model generally decreased, which shows that our model benefits from the existing architectural design. "Without all" refers to the removal of all modules related to the rationales generated by the LLM. The decrease in performance after this removal is the largest, indicating that our method enhances

the detection performance of the small model through the output of the LLM. Additionally, when we replaced the feature aggregation component with direct concatenation and averaging, we observed a notable decline in performance. This suggests that the importance of the post semantic vector and the interaction vector varies across different posts. In cases where the large language model's judgment is less reliable, the weight assigned to the original post's semantic vector should be increased.

Table 4. Llama performances on Kaggle testing set.

Methods	Kaggle				
	I/E	S/N	T/F	P/J	Avg
Llama	70.48	63.04	77.70	51.96	65.80
$\text{Llama}_{\text{cot}}$	69.84	63.25	78.52	51.00	65.65
$\text{Llama}_{\text{3 shot}}$	67.30	60.45	72.57	48.40	62.18
DKE	**72.19**	**67.46**	**81.01**	**68.63**	**72.32**

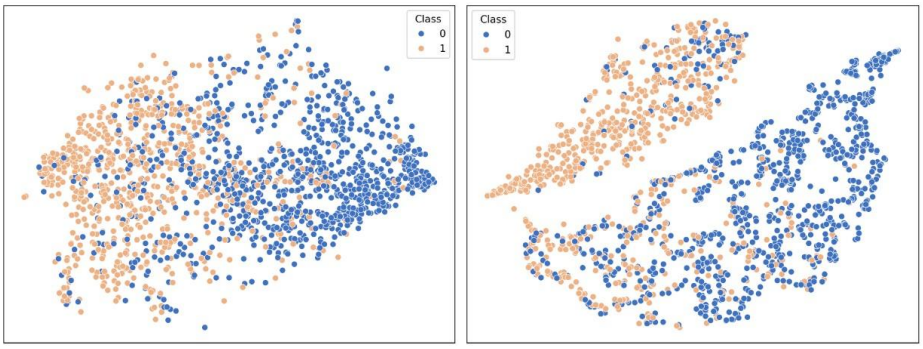
4.6 LLM Performance

To analyze the performance of LLM in the personality detection task, we conducted a series of experiments using Llama. We used the Kaggle test set to evaluate the capabilities of Llama in three settings: zero-shot, CoT, and few-shot. For the few-shot setting, we randomly selected 3 examples from the training set. Due to the input length limitation, only some posts in the sample were given, and user posts that exceeded the input limit were discarded. As shown in Table 2, the performance of Llama lags behind most of the fine-tuned small models. As shown in Table 4, classification performance does not improve in either the CoT or few-shot settings. This suggests that existing methods struggle to improve the performance of LLMs in personality detection, making it unsuitable to apply LLMs directly to this task. Leveraging the knowledge acquired by LLMs to improve the detection performance of small models is a promising direction.

We also consider the performance of other LLMs in personality detection. PsyCoT [14] utilizes ChatGPT for personality detection and finds that directly applying LLMs for personality detection yields unsatisfactory results. While using psychological questionnaires as a chain-of-thought shows a certain degree of improvement, it still fails to surpass the performance of small models. Our approach outperforms existing small models. Additionally, limited by API resources, PsyCoT only tested 200 samples. This indicates that API access restrictions limit the research potential of LLMs in personality detection. In this context, leveraging the open-source LLaMA model for personality detection offers a distinct advantage, as it is more accessible than proprietary models such as ChatGPT, thereby avoiding the limitations imposed by restricted availability and closed-source architectures.

4.7 Visualization Analysis

To more intuitively illustrate the improvement of our model in user personality representation, we applied dimensionality reduction on high-dimensional data to obtain visualized results. Since D-DGCN is the previous state-of-the-art method and its code is publicly available, we chose D-DGCN as our comparison model. Specifically, we employ t-SNE [30] to reduce the dimension of the representations learned by D-DGCN and our proposed model in the T/F trait of Kaggle's test set to 2 and visualize the effect. As shown in Fig. 4, the representations obtained by our model exhibit a more distinct clustering effect in the data distribution, indicating that our method has improved user personality representation. This improvement is due to the incorporation of domain knowledge into the model, which results in personality representations that are more aligned with the personality traits defined in psychological theories. In psychological terms, the two personalities within the same dimension are opposites, making it easier to identify representations that align more closely with the personality definition.



(a) D-DGCN (b) DKE

Fig. 4. Visualization results of the produced user representations of the T/F dimension, in which (a) is the D-DGCN, (b) is our DKE.

5 Conclusion

In this paper, we propose an LLM-based domain knowledge enhanced model for comprehensible personality detection. The model leverages the powerful semantic understanding and vast knowledge base of the LLM to generate personality analysis rationales for users, serving as external domain knowledge to enhance the personality detection of the small model. Additionally, the model enables rich information interaction between rationales and posts through dual cross-attention and adaptively integrates interactive features with the original semantic features. Finally, the model obtained a more comprehensible personality representation, with a greater distinction between different personality traits. The improved interpretability stems from the alignment between learned representations and psychological personality theory, yielding representations that better reflect personality traits. Experimental results on two datasets

demonstrate that our model outperforms all baseline models. Distilling knowledge from LLMs to enhance the knowledge of small models is a promising direction for future work.

Acknowledgments. This work was supported by the National Natural Science Foundation of China under Grants 62172278 and 62302303.

References

1. Kernberg, O.F.: What Is Personality? Journal of Personality Disorders. 30, (2016).
2. Myers, I.B.: Introduction to type: A description of the theory and applications of the Myers-Briggs Type Indicator. Consulting Psychologists Press (1987).
3. Xue, D., Wu, L., Hong, Z., Guo, S., Gao, L., Wu, Z., Zhong, X., Sun, J.: Deep learning-based personality recognition from text posts of online social networks. Applied Intelligence. 48, 4232–4246 (2018).
4. Stachl, C., Pargent, F., Hilbert, S., Harari, G.M., Schoedel, R., Vaid, S., Gosling, S.D., Bühner, M.: Personality research and assessment in the era of machine learning. European Journal of Personality. 34, 613–631 (2020).
5. Shen, T., Jia, J., Li, Y., Ma, Y., Bu, Y., Wang, H., Chen, B., Chua, T.-S., Hall, W.: Peia: Personality and emotion integrated attentive model for music recommendation on social media platforms. In: Proceedings of the AAAI conference on artificial intelligence. pp. 206–213 (2020).
6. Wen, Z., Cao, J., Yang, R., Liu, S., Shen, J.: Automatically select emotion for response via personality-affected emotion transition. In: Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. pp. 5010–5020 (2021).
7. Bagby, R.M., Gralnick, T.M., Al-Dajani, N., Uliaszek, A.A.: The role of the five-factor model in personality assessment and treatment planning. Clinical Psychology: Science and Practice. 23, 365 (2016).
8. Mehta, Y., Majumder, N., Gelbukh, A., Cambria, E.: Recent trends in deep learning based personality detection. Artificial Intelligence Review. 53, 2313–2339 (2020).
9. Zhu, H., Zhang, X., Lu, J., Yang, L., Lin, H.: Integrating Multi-view Analysis: Multi-view Mixture-of-Expert for Textual Personality Detection. In: CCF International Conference on Natural Language Processing and Chinese Computing. pp. 359–371. Springer (2024).
10. Yang, T., Yang, F., Ouyang, H., Quan, X.: 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). pp. 4229–4239 (2021).
11. Zhu, Y., Xia, Y., Li, M., Zhang, T., Wu, B.: Data Augmented Graph Neural Networks for Personality Detection. In: Proceedings of the AAAI Conference on Artificial Intelligence. pp. 664–672 (2024).
12. Pennebaker, J.W.: Linguistic inquiry and word count: LIWC 2001, (2001).
13. Xi, Y., Liu, W., Lin, J., Cai, X., Zhu, H., Zhu, J., Chen, B., Tang, R., Zhang, W., Yu, Y.: Towards open-world recommendation with knowledge augmentation from large language models. In: Proceedings of the 18th ACM Conference on Recommender Systems. pp. 12–22 (2024).
14. Yang, T., Shi, T., Wan, F., Quan, X., Wang, Q., Wu, B., Wu, J.: PsyCoT: Psychological Questionnaire as Powerful Chain-of-Thought for Personality Detection. In: Findings of the Association for Computational Linguistics: EMNLP 2023. pp. 3305–3320 (2023).

15. Zhao, H., Chen, H., Ruggles, T.A., Feng, Y., Singh, D., Yoon, H.-J.: Improving text classification with large language model-based data augmentation. Electronics. 13, 2535 (2024).
16. Hu, B., Sheng, Q., Cao, J., Shi, Y., Li, Y., Wang, D., Qi, P.: Bad actor, good advisor: Exploring the role of large language models in fake news detection. In: Proceedings of the AAAI Conference on Artificial Intelligence. pp. 22105–22113 (2024).
17. Zhang, Y., Jin, R., Zhou, Z.-H.: Understanding bag-of-words model: a statistical framework. International journal of machine learning and cybernetics. 1, 43–52 (2010).
18. Cui, B., Qi, C.: Survey analysis of machine learning methods for natural language processing for MBTI Personality Type Prediction. Final Report Stanford University. (2017).
19. Amirhosseini, M.H., Kazemian, H.: Machine Learning Approach to Personality Type Prediction Based on the Myers-Briggs Type Indicator R. Multimodal Technologies and Interaction, 4, Article 9. (2020).
20. Tandera, T., Suhartono, D., Wongso, R., Prasetio, Y.L.: Personality Prediction System from Facebook Users. Procedia Computer Science. 116, 604–611 (2017).
21. Jiang, H., Zhang, X., Choi, J.D.: 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. pp. 13821–13822 (2020).
22. Yang, F., Quan, X., Yang, Y., Yu, J.: Multi-document transformer for personality detection. In: Proceedings of the AAAI conference on artificial intelligence. pp. 14221–14229 (2021).
23. Yang, T., Deng, J., Quan, X., Wang, Q.: Orders are unwanted: dynamic deep graph convolutional network for personality detection. In: Proceedings of the AAAI Conference on Artificial Intelligence. pp. 13896–13904 (2023).
24. Liu, W., Sun, Z., Wei, S., Zhang, S., Zhu, G., Chen, L.: PS-GCN: Psycholinguistic graph and sentiment semantic fused graph convolutional networks for personality detection. Connection Science. 36, 2295820 (2024).
25. Shahabikargar, M., Beheshti, A., Khatami, A., Nguyen, R., Zhang, X., Alinejad-Rokny, H.: Domain Knowledge Enhanced Text Mining for Identifying Mental Disorder Patterns. In: 2022 IEEE 9th International Conference on Data Science and Advanced Analytics (DSAA). pp. 1–10. IEEE (2022).
26. Guo, Z., Ding, N., Zhai, M., Zhang, Z., Li, Z.: Leveraging domain knowledge to improve depression detection on Chinese social media. IEEE Transactions on Computational Social Systems. 10, 1528–1536 (2023).
27. Agarwal, B.: Financial sentiment analysis model utilizing knowledge-base and domain-specific representation. Multimedia Tools and Applications. 82, 8899–8920 (2023).
28. Xie, A., Zhu, F., Han, J., Hu, S.: Integrating Open-domain Knowledge via Large Language Model for Multimodal Fake News Detection. In: 2024 27th International Conference on Computer Supported Cooperative Work in Design (CSCWD). pp. 1917–1922. IEEE (2024).
29. Štajner, S., Yenikent, S.: Why is MBTI personality detection from texts a difficult task? In: Proceedings of the 16th conference of the European chapter of the association for computational linguistics: main volume. pp. 3580–3589 (2021).
30. Van der Maaten, L., Hinton, G.: Visualizing data using t-SNE. Journal of machine learning research. 9, (2008).