Towards Transferable Personality Representation Learning based on Triplet Comparisons and Its Applications

Kai Tang[1*], **Rui Wang** [1*], **Renyu Zhu**[2*], **Minmin Lin**[2], **Xiao Ding**[3], **Tangjie Lv**[2], **Changjie Fan**[2], **Runze Wu**[2†], **Haobo Wang**[1†],
[1]Zhejiang University [2]NetEase Fuxi AI Lab [3]Harbin Institute of Technology
{tk0819,22351017,wanghaobo}@zju.edu.cn,xding@ir.hit.edu.cn
{zhurenyu,linminmin01,hzlvtangjie,fanchangjie,wurunze1}@corp.netease.com

Abstract

Personality is an important concept in psychology that reflects individual differences in thinking and behavior, and has significant applications across various fields. Most existing personality analysis methods address this issue at the bag level, treating the entire corpus gathered from one individual as a single unit for classification. However, this paradigm presents several challenges. From the data perspective, collecting a large corpus for each individual and performing comprehensive annotations pose significant difficulties in both data collection and labeling. On the application side, concentrating on classifying the entire corpus limits its applicability in more common single-instance scenarios. To address these issues, we propose a new task paradigm in text-based personality representation learning. Specifically, we construct a triplet personality trend comparison dataset to learn single-sentence personality embeddings with desirable metric properties. This approach removes the traditional constraints on data sources, facilitating dataset expansion, and can leverage the transfer capabilities of embeddings to easily adapt to various downstream tasks. Our experiments show that the learned embeddings significantly boost performance by a relative 10% across various applications, including personality detection, personality retrieval, and emotion translation prediction. The code and dataset are available at https://github.com/zjutangk/PTCD.

1 Introduction

Personality, a key psychological concept, highlights individual differences in thoughts, feelings, and behaviors (Corr and Matthews, 2020). It is crucial as it reflects a person's true nature and influences how others perceive them (Hogan, 2017). With the development of NLP, automatic personality detection has received significant attention (Petrides and Mavroveli, 2018; Mehta et al., 2020; Yang et al., 2023a; Lynn et al., 2020) due to its wide-ranging applications such as psychological health assessment (Wilkinson and Walford, 2001), role-playing (Tu et al., 2024) and personalized recommendation systems (Hu and Pu, 2011; Kleć et al., 2023).

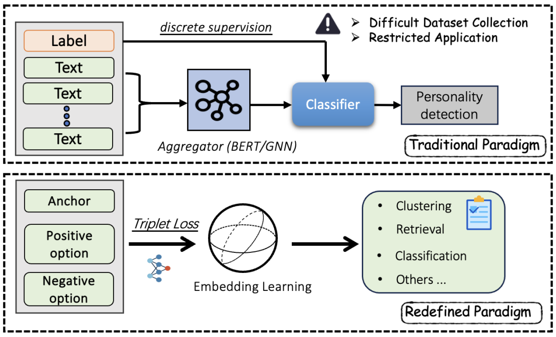


Figure 1: The diagram compares the traditional learning paradigm with our approach in terms of dataset structure and workflow setting.

To express personality in AI models, current algorithms primarily follow the traditional supervised learning paradigm (Tandera et al., 2017; Yang et al., 2021a,b; Zhao et al., 2021; Zhu et al., 2025; Yang et al., 2025). Specifically, researchers collect corpus-personality dataset to train prediction models, where each sample is composed of multiple sentences grouped into a bag, and personality labels are manually annotated in a discrete manner. This approach often involves collecting datasets from social platforms like Twitter and Reddit, where users' posts are gathered as samples[1]. Unfortunately, this multi-instance learning style (Foulds and Frank, 2010) presents several challenges: (i)-**Data Collection Difficulty**: A large amount of utterances from same individual is required, which can be hard to obtain; (ii)-**Annotation Difficulty**:

[*] Equally contribution. [†] Corresponding author.

[1] Although these platforms provide a large volume of data, it is often accompanied by redundant content and label leaks.

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 10062–10078
November 4-9, 2025 ©2025 Association for Computational Linguistics

Bag-level fine-grained annotation is highly challenging for human annotators; (iii)-**Application Difficulty**: Many downstream applications, such as dialogue systems, require single-sentence representation, which does not fit the bag-level approach. Therefore, there is an urgent need for a method to learn pretrained personality representations that balances the trade-off between sample size and annotation quality while also transferring effectively to downstream tasks.

The essence of these challenges lies in the fact that both the structure of datasets and the learning methods remain at the bag level. Technologically, transitioning to a single-sentence level for dataset construction and pipeline design could significantly facilitate dataset scale-up and task transfer. This shift, though promising, faces two critical barriers. On the one hand, single sentences often lack comprehensive and fine-grained personality traits, which makes accurate personality annotation at the sentence level inherently difficult. On the other hand, though several general-purpose sentence representation methods are available currently like BERT and GPT(Floridi and Chiriatti, 2020; Devlin et al., 2019), most of these language models are limited to semantic understanding and are not specifically trained for personality. LLMs also demonstrate strong generalization abilities in language representation. But they are more suited for generation tasks, and their application in recommendation and retrieval tasks requires further investigation. Due to the difficulty in annotating single-sentence personality supervision and the lack of pre-trained models in the personality domain, we aim to transform the form of supervision information to learn personality domain single-sentence representations.

In our work, we revisit the field of text-based personality learning and **propose a new task paradigm based on triplet comparison**. We aim to learn a continuous embedding space for personality that mirrors human perception, and then transfer it to various downstream tasks. Specifically, when humans perceive two utterances as more similar in personality traits than a third one, the embedding space should reflect this by showing a closer distance between the first two utterances compared to the third. In pursuit of our research goals, we developed the Personality-Tendency-Compare Dataset (PTCD), which contains 36,294 high-quality triplet samples extracted from a foundational corpus of 162,505 multi-domain chinese utterances. By utilizing triplet annotations, we can avoid limiting the data to a single individual's corpus and reduce the difficulty of manual labeling. This dataset is the first large-scale collection of utterances annotated with personality comparisons, setting a precedent for single-sentence personality representation learning. Based on the constructed dataset, we use triplet loss to learn a compact representation of personality with good metric properties.

Additionally, we integrate personality embeddings into various downstream tasks, including personality detection, personality-aware retrieval, and emotion translation prediction. We demonstrate their effectiveness through comprehensive experiments. Our results from the PTCD dataset show that our method not only surpasses existing embedding techniques and state-of-the-art large language models in predicting personality tendencies aligned with human preferences, but also significantly enhances performance across various domains. Specifically, we achieve an over 10% relative improvement in performance across the SOTA of all tasks, illustrating the strong transfer capabilities of our personality representation.

2 Related Work

Personality Classification. In the early stages of personality detection, (Francis and Booth, 1993) introduced the Linguistic Inquiry and Word Count (LIWC), to extract psycholinguistic features for text-based personality analysis. Based on LIWC, (Argamon et al.; Cui and Qi, 2017; Amirhosseini and Kazemian, 2020) conducted feature engineering research using traditional machine learning methods. Recent advancements in deep learning technology have significantly accelerated progress in personality classification and DNN-based methods primarily focus on feature aggregation and post-processing of encoding. About feature aggregation, (Jiang et al., 2020) aggregates at the text level by directly concatenating the corpus into a document for encoding; (Xue et al., 2018) use CNN to aggregate features of posts from social media; (Lynn et al., 2020; Wang et al., 2021) utilized a hierarchical attention network to learn user representations from posts. Several graph-based learning methods (Yang et al., 2023a; Zhu et al., 2024) have also been proposed to alleviate the misleading effects of sequential information in the corpus on model training. However, graph construction incurs additional time and computational costs during both the training and inference processes. Due to the recent focus on LLM technology in the NLP field, some works (Yang et al., 2023b; Li et al., 2024) directly use LLMs to infer personality from text, eliminating reliance on features, which are often restricted to one single downstream task and are limited by the scale of the problem.

Sentence Embedding. Sentence-level representations are crucial for various NLP tasks, as they can capture linguistic properties effectively using vector representations. Earlier methods (Kiros et al., 2015; Gan et al., 2017) used encoder-decoder architectures to predict surrounding sentences and autoencoders (Hill et al., 2016; Zhang et al., 2018) to reconstruct them. Recent advancements have employed complex transformer-based neural networks. For instance, Sentence-BERT (Reimers and Gurevych, 2019) introduced a Siamese network to enhance BERT's efficiency, while Sentence-T5 (Ni et al., 2022) utilizes the T5 model and contrastive learning to optimize embeddings for semantic similarity tasks (Conneau and Kiela, 2018). Some studies (Gao et al., 2021; An et al., 2024) in natural language inference (NLI) have applied contrastive learning to sentence embeddings, using entailed and contradicted sentences as examples. Despite these advancements, such methods rely heavily on discrete labels and have primarily focused on NLI datasets. As a result, their application in personality-related tasks remains limited due to the scarcity of reliable datasets.

3 Personality Tendency Compare Dateset

In this paper, we construct a multi-domain high-quality Personality Tendency Compare Dataset (PTCD) in triplet form to address the limitations of the traditional paradigm for personality embedding learning. Formally, personality datasets are constructed as follows:

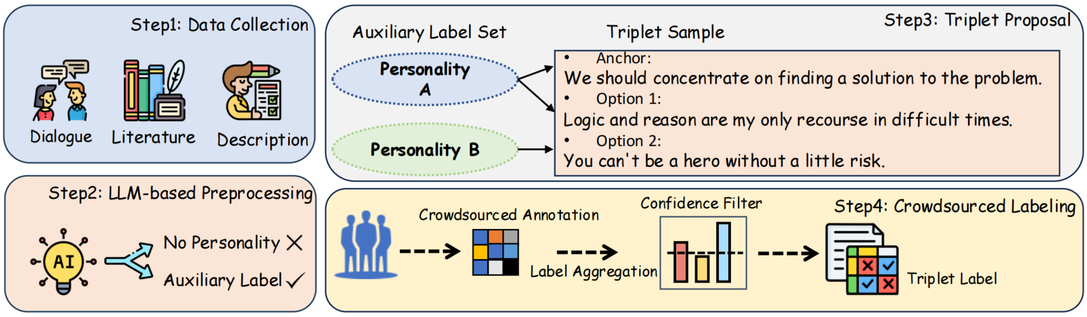
- **Traditional**: One may collect M individuals' corpus $D_M = \{U_1, U_2, ..., U_M\}$ where each individual i has N_i utterances. Scaling up per-individual textual samples faces inherent limitations in natural language processing, particularly when balancing ecological validity against privacy-preserving data collection.

- **Our Triplet Format**: Our dataset define triplets $\Gamma = \{(u_a, u_p, u_n)\}$ where anchor u_a shares contextual similarity with positive sample u_p but differs from negative sample u_n. This paradigm overcomes the traditional constraint of requiring extensive single-individual corpora through comparative triplet formulation, thereby enabling scalable construction of personality datasets.

Data Collection. To ensure the diversity of the dataset, we collected personality-related corpora from three distinct domains: dialogue, literature, and description of personality. The former two are widely prevalent across social media platforms and entertainment works, while the latter serves as a typical carrier of personality-related information. The details of different domain data are as follows: **Dialogue:** We collect open-source datasets of human conversations, including live-streaming chat log dataset LiveChat(Gao et al., 2023) and social media exchanges dataset PersonalDialog(Zheng et al., 2019, 2020). **Literature:** The Literature domain includes character dialogues from classic literature and other entertainment works (such as anime and games). Specifically, we collected dialogue data from 320 characters across three open-source datasets: ChatHaruni(Li et al., 2023), CharacterEval(Tu et al., 2024) and HPD(Chen et al., 2023). **Personality Description:** We designed a systematic approach LLM to collaboratively generate personality description corpora from character personas. Specifically, we obtained comprehensive character personas from the open-source dataset PersonalHub(Ge et al., 2024), then utilized LLM to generate corresponding personality descriptions through well-crafted prompts. About the detail of prompt design, please refer to AppendixC.

LLM-based Preprocessing. Given that we collected extensive data but aimed to construct a triplet-structured dataset (composed of three single sentences), the quality of individual sentences significantly impacts the final triplet quality. To ensure high-quality downstream generation, we implemented LLM-based preprocessing: first filtering out sentences lacking personality-related information (e.g., greetings or formalities), then assigning auxiliary labels based on MBTI taxonomy[2] to guide subsequent triplet construction. The auxiliary labeling process operates just as a coarse-grained preprocessing mechanism solely intended to guide triplet construction. Therefore, other personality taxonomies like Big 5 are also feasible, and the experiment results shown in Section 6.2.1 demonstrate the generalized effectiveness of our method

[2] Prompt details are provided in Table 14 of Appendix.



- Anchor:

We should concentrate on finding a solution to the problem.

- Option 1:

Logic and reason are my only recourse in difficult times.

- Option 2:

You can't be a hero without a little risk.

Figure 2: Overview of the construction of the PTCD dataset. The pipeline comprises four steps: (1) Corpus Collection: license-compliant corpora were collected from multiple domains; (2) LLM-based Preprocessing: LLM performed initial data curation by filtering non-personality-related utterances and annotating auxiliary labels characterizing personality tendencies; (3) Triplet Proposal: personality-relevant triplets were automatically generated based on auxiliary labels; and (4) Crowdsourced Labeling: high-quality triplet annotations were obtained through professional crowdsourcing platforms.

Domain	Dialogue	Literature	Description	All
Num	49,996	55,459	57,050	162,505

Table 1: The distribution of different domains of corpus.

Split	Train	Valid	Test	All
Num	24,195	6,050	6,049	36,294

Table 2: Details about the split of PTCD.

across multiple personality taxonomies. We do further analysis analysis between different personality taxonomies in Appendix D. After preprocessing, the distribution of different domain domains is shown in Table 1.

Triplet Proposal. During the LLM-based preprocessing, we get sentence set $\mathcal{D}_{a^i}$ for each auxiliary label a^i, where the auxiliary label set is denoted as $\mathcal{A}$. To generate triplets, we first randomly select an anchor sentence u_a belonging to the auxiliary label a^i, then sample u_p and u_n from D_{a^i} and $\cup_{a' \in \mathcal{A}/a^i} D_{a'}$ respectively. To maximize sample diversity, all random selections are in the way of sampling without replacement. During this process, we generate 40,000 triplets in all.

Crowdsourced Annotation. We leveraged the YouLing platform[3], a crowd-sourcing annotation service operated by NetEase. Each triplet was annotated by five trained labelers through the platform's workflow management system. Following annotation collection, we employed the Dawid-Skene algorithm (Dawid and Skene, 1979) to resolve label disagreements via probabilistic inference. Finally, we applied confidence thresholding to filter low-quality annotations, ensuring the final dataset met rigorous quality standards. Finally, we obtained 36,294 triplets with high annotation consistency, and the detail of split is shown in Table2.

Personality Detection Dataset Based on the collected corpus, we also develop a Chinese personality detection dataset as a by-product. This dataset includes 200 famous characters, with each character having an average of over 200 entries that clearly contain personality-related information. Compared to widely used social media datasets, our data sources are more diverse, ensuring a higher quality of the dataset serving as a benchmark.

4 Measurable Embedding Learning

To learn single-utterance personality representation with good metric properties, we conduct measurable embeddings learning (MEL) by triplet learning with a pre-trained language model BERT (Devlin et al., 2019) on the basis of PTCD dataset. Notably, we use BERT to align with other mainstream works and other pre-trained language models are equally applicable to this method.

Pretrain Setting. Integrating domain knowledge into BERT enhances its performance on personality classification tasks (Gururangan et al., 2020). Specifically, we extract single-sentence data from the triplet training set as pretraining data. We apply the masked language model (MLM) for pretraining in line with mainstream methodologies.

Triplet Learning. To learn representations with good metric properties, we train using labeled triplets from PTCD. For single utterance u_i, we

[3] https://fuxi.163.com/productDetail/zb

use BERT as encoder to obtain the corresponding normalized features $z_i = \mathrm{Normalize}(\mathrm{BERT}(\mathrm{u_i}))$. Given a triplet$(u_{\text{anchor}}, u_{\text{pos}}, u_{\text{neg}})$, our goal is to learn a set of representations $(z_{\text{anchor}}, z_{\text{pos}}, z_{\text{neg}})$ that minimize the distance between z_{anchor} and z_{pos} while maximizing their distance from z_{neg}. To be specific, we optimize the following triplet loss to learn metric representations:

$$\mathcal{L}_{tri} = \sum_{i}^{N} \max(\mathrm{d_{pos}} - \mathrm{d_{neg}} + \mathrm{margin}, 0) \quad (1)$$

where $d_{\text{pos}} = \mathrm{distance}(\mathrm{z_{anchor}}, \mathrm{z_{pos}})$ and $d_{\text{neg}} = \mathrm{distance}(\mathrm{z_{anchor}}, \mathrm{z_{neg}})$. Since the features used here are normalized, we directly use the L2 distance to calculate the distance. margin is a hyperparameter that controls the strictness of loss constraints. A larger margin leads to clearer feature separation but makes model convergence more challenging. Through measurable embedding learning on triplet dataset, we get personality encoder M_P that solves the problem of single-sentence personality representation. Leveraging its transferability, we can apply it to various downstream tasks.

5 Downstream Application

In this section, we explore the methodologies utilized to leverage learned representations in diverse downstream applications.

5.1 Direct Personality Detection

Personality detection is the most direct application of personality representation, and the community has conducted extensive research on this issue. Using this task as an example, we explore the specific application methods of personality embeddings in downstream tasks with bag input types. Personality detection can be formulated as a multi-instance multi-label classification problem (Zhou et al., 2012). Mathematically, given a set of n utterances $\boldsymbol{U}_x = \{u_1, u_2, ..., u_n\}$ from an individual x, where $u_i = [\omega_{i1}, \omega_{i2}, ..., \omega_{im}]$ is i-th utterance with m tokens. The goal of the problem is to predict the T-dimensional personality traits $Y = [y_1, y_2, ..y_T]$ for given $\boldsymbol{U}_x$. For MBTI taxonomy, T = 4 and Y is a binary vector. In our method, we firstly train an encoder for a measurable personality representation $\{z_1, z_2, ..., z_n\} = \mathrm{encoder}(\boldsymbol{U})$ and then learn embedding pooling architecture for classification task $\mathcal{F}' : \mathrm{Pool}(z_1, z_2, ..., z_n) \rightarrow Y$. There is a difference from the traditional approach which learns a mapping function $\mathcal{F} : \boldsymbol{U} = \{u_1, u_2, ..., u_n\} \rightarrow Y$.

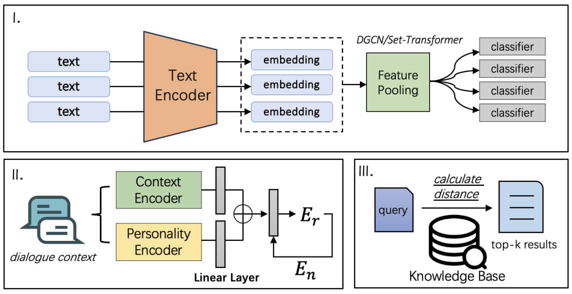


Figure 3: Illustration diagrams of different applications, including personality detection, emotion translation prediction, and retrieval.

Permutation-Invariant Pooling (PIP). Personality detection is a multi-instance multi-label learning task, challenged by the unstructured and variable nature of user utterances. Effective pooling of single-utterance embeddings requires: 1. Permutation invariance to avoid learning unintended sequential dependencies. 2. The ability to handle variable-sized inputs.

Inspired by the Set Transformer (Lee et al., 2019), we construct a permutation invariant decoder using multi-head attention blocks (MAB)[4] as the pooling architecture. Considering the MBTI taxonomy has four dimensions, we set up four corresponding seed vectors $\mathbf{S} = [s_1, s_2, s_3, s_4]$ combined with multi-head attention to aggregate features to four pooled outputs. We further model these pooled outputs using MAB to ensure permutation invariance and preserve information regarding interactions between outputs. The specific process is as follows:

$$\mathrm{Pool}(\mathbf{Z}) = \mathrm{MAB}(\mathrm{MAB}(\mathbf{S}, \mathbf{Z}), \mathrm{MAB}(\mathbf{S}, \mathbf{Z})) \quad (2)$$

where $\mathbf{Z}$ is set of utterance representations. This pooling architecture is notable for having only two learnable layers, which allows for significant performance improvements just through few-shot finetuning, as demonstrated in Table 3.

Multi-Label Classification. Using the MBTI taxonomy as an example, it consists of four independent binary classifications. We set up four binary classification heads corresponding to each MBTI dimension, each receiving one of the four pooled outputs. The probability outputs from these four classification heads are concatenated to calcu-

[4] Details of this module can refer to Appendix A.4

late binary cross-entropy loss:

$$\mathcal{L}_{bce} = -\frac{1}{N}\sum_{i=0}^{N}(y_i \text{log} p_i + (1 - y_i) log(1 - p_i)) \tag{3}$$

where p_i is a four-dimensional vector obtained by concatenating the prediction results from four binary classification heads after pooling the features of individual x_i.

5.2 Personality-Aware Retrieval

Personality-aware retrieval is a direct application of the well-measured properties of personality embeddings. Given a knowledge base $D = u_0, u_1, ..., u_n$ contains n utterances, we obtain the corresponding personality embeddings $Z = \{z_i | z_i = \text{encoder}(u_i)$ for $u_i \in D\}$. For query u_q, we get its corresponding embedding z_q and calculate the distance between z_q and each sample in the knowledge base D that distance $= \{d_i \mid d_i = ||z_q - z_i||_2$ for $z_i \in Z\}$. So that we can obtain the top-k samples that are most similar to the personality expressed by the query with top-k small distance. This retrieval process can be applied to decision-making in multi-agent systems and the field of NPC generation. We have also conducted related research, which is presented in the appendix.

5.3 Emotion Translation Predict

To provide consistent emotional interaction with users, dialog systems (Liu et al., 2022) should be capable to automatically select appropriate emotions for responses like humans. Current study (Wen et al., 2021) has already demonstrated that individual differences in emotion expression are closely related to personality traits. However, most dialog systems rely on user surveys or profiles for personality information, which isn't ideal for cold-start scenarios or adaptive interactions. We address this by integrating personality embeddings with an emotion translation model, using real-time personality information to predict emotions that react to the dialog context. Formally, dialogue context between user and dialogue system $C = \{u_1, u_2, ..., u_n\}$ contains n utterances from both user and dialogue system. e_i is the emotion expressed in u_i and we use e_r to represent the response emotion we want to predict. We obtain pretrained personality embedding $z_i = M_p(u_i)$ and incorporate it into the learning process of emotion transition model F_{et}: $e_r = F_{\text{et}}(e_n \mid C, z_n)$.

6 Experiment

In this section, we demonstrate the measurability of learning personality representations from the constructed PTCD dataset, as well as their practicality in various applications.

6.1 Implementation Details

We use the pre-trained *bert-base-uncased* BERT model from (Wolf et al., 2019) as our backbone. We set the maximum length of a post to 100 for both datasets. Due to our permutation-invariant pooling architecture, we do not have to set the maximum number of utterances. About measurable embedding learning, we employ AdamW optimizer with a warm-up schedule and 0.01 weight decay. The learning rate is set to $5e^{-5}$ for all benchmark datasets. Regarding the dataset processing, for the Kaggle dataset, we followed prior works by removing certain instances of information leakage and noise while maintaining the same partitioning. For the Essays dataset, which consists of numerous semantically uncorrelated sentences concatenated into long texts, we restored these long texts into collections of individual sentences based on punctuation for representation learning. Details of datasets can refer to appendix A.1.

6.2 Main Results

6.2.1 App. ①: Personality Detection

We primarily validate the feasibility of measurable embeddings through personality classification.

Baselines. For the personality detection task, we conducted a comprehensive comparison of various approaches. The traditional machine learning methods include: LIWC+SVM (Tighe et al., 2016), W2V+CNN (Rahman et al., 2019). The deep learning methods include: AttRCNN (Xue et al., 2018), DDGCN (Yang et al., 2023a). The LLM-based (Large Language Model-based) methods include: TAE (Hu et al., 2024), PsyCoT (Yang et al., 2023b). For more details about methods, please refer to the appendix A.3.

Direct Personality Detection. Firstly, we directly tested the effectiveness of our work using the personality classification corresponding to the PTCD dataset. Specifically, we fixed the parameters of the encoder trained with triplet loss and performed fine-tuning only on the pooling layer. Table 3 shows that our method combining measurable embedding learning with permutation-invariant pool-

Method	I/E		S/N		T/F		J/P		Average	
	ACC	**F1**	**ACC**	**F1**	**ACC**	**F1**	**ACC**	**F1**	**ACC**	**F1**
LIWC+SVM	60.12	56.53	58.44	57.13	61.27	57.92	61.33	57.98	60.29	57.39
BERT	78.55	75.10	82.77	76.19	84.33	79.25	76.19	74.44	80.46	76.26
RoBERTa	77.59	73.44	80.15	78.80	80.22	78.12	81.02	77.92	79.75	77.04
DDGCN	79.82	79.27	79.36	75.50	83.26	73.79	84.71	81.64	81.78	77.55
PsyCOT	57.12	52.33	76.81	43.44	57.97	42.25	56.52	56.29	62.10	48.58
GPT-4o	86.95	91.73	55.07	43.63	82.60	89.83	76.81	46.66	75.36	74.93
MEL+PIP	**95.65**	**94.26**	**94.20**	**91.48**	**95.65**	**90.84**	**94.20**	**93.15**	**94.94**	**92.93**

Table 3: Performance on testing sets of PTCD. Average results over 3 runs are reported. For each metric, the best results are marked in bold.

Method	I/E		S/N		T/F		J/P		Average	
	ACC	**F1**	**ACC**	**F1**	**ACC**	**F1**	**ACC**	**F1**	**ACC**	**F1**
LIWC+SVM	69.55	50.01	77.52	55.67	75.35	59.82	66.32	57.20	72.18	55.68
BERT	77.30	62.50	84.90	54.04	78.30	77.93	69.50	68.80	77.50	65.82
RoBERTa	77.10	61.89	86.50	57.59	79.60	78.69	70.60	70.07	78.45	67.06
AttRCNN	-	59.74	-	64.08	-	78.77	-	66.44	-	67.25
DDGCN	78.10	70.26	84.40	60.66	79.30	78.91	73.30	71.73	78.78	70.39
PsyCOT	79.00	66.56	85.00	61.70	75.00	74.80	57.00	57.83	74.00	65.22
TAE	-	**70.90**	-	66.21	-	**81.17**	-	70.20	-	72.07
GPT-4o	73.27	60.93	85.97	**72.84**	76.65	72.23	66.91	45.25	75.68	62.82
MEL+PIP	76.98	65.44	78.02	66.20	72.10	60.33	68.97	64.72	74.01	64.17
MEL+DGCN	**80.25**	70.22	**86.74**	68.59	**81.02**	80.25	**78.89**	**75.00**	**81.72**	**73.52**

Table 4: Performance on Kaggle dataset which is labeled by MBTI taxonomy. For each metric, the best results are marked in bold.

ing outperforms all competing methods by a considerable margin on all metrics: with the encoder frozen and only the pooling layer fine-tuned, we achieved an overall performance improvement of 12% over other mainstream methods.

Transfer Personality Detection. Secondly, we tested the transfer ability of pretrained personality embeddings on other publicly datasets. As shown in Table 4 and 5, we outperformed domain-specific training methods in domain adaptation scenarios across various open-source datasets with different classification rules. This demonstrates the strong transferability of measurable embeddings and suggests that training on a dataset within the same domain is likely to yield even more exceptional results. More details of experiment can refer to appendix A.5.

6.2.2 App. ②: Emotion Translation Predict

Current study (Wen et al., 2021) has released a dialog-emotion dataset PELD which includes 6,510 dialogue triples of daily conversations with emotion labels. We use RoBERTa as our base model in line with mainstream methods. Keeping the training methodology unchanged, we only use our pre-trained personality embeddings to integrate personality information, assessing the effectiveness of the learned personality embeddings in emotion translation prediction. Experiment result shown in Table 7 demonstrates that integrating personality information into emotion translation prediction in the form of single-sentence personality embeddings can significantly enhance task performance.

6.2.3 App. ③: Similarity-Based Personality Retrieval

Retrieval is an important application for embeddings, and we evaluate the performance of measurable embeddings in the similarity-based personality retrieval task. Specifically, we split the test set of the PTCD dataset into individual sentences to serve as the knowledge base. For a given query, we use the trained encoder to obtain the corresponding embedding and calculate the cosine similarity with the embeddings in the knowledge base, selecting the top N most similar utterances. For comparison, we also tested a BERT model trained using MLM and MLL. Due to the lack of relevant datasets in the field of personality retrieval, we employ two approaches to validate the performance of the retrieval task: (i) **Objective metric**. We use utterances from classic roles in the knowledge base as inputs and test the proportion of N retrieved similar utterances that belong to the same or similar personalities. This objective metric demonstrates the embeddings' utility in style transfer and NPC generation. (ii) **Subjective metric**. We invited 40 annotators through a well-established crowdsourcing platform to evaluate which of the two

Method	AGR		CON		EXT		NEU		OPN		Average	
	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1
LIWC+SVM	51.78	47.50	51.99	52.00	51.22	49.20	51.09	50.90	54.05	52.40	52.03	50.40
W2V+CNN	-	46.16	-	52.11	-	39.40	-	58.14	-	59.80	-	51.12
BERT	56.84	54.72	57.57	56.41	58.54	58.42	56.60	56.36	60.00	59.76	57.91	57.13
RoBERTa	59.03	57.62	57.81	56.72	57.98	57.20	56.93	56.80	60.16	59.88	58.38	57.64
PsyCOT	61.13	61.13	59.92	57.41	59.76	59.74	56.68	56.68	60.73	57.30	59.64	58.43
GPT-4o	63.15	62.55	60.82	60.02	49.25	47.88	55.98	54.39	61.06	60.56	58.05	57.08
MEL+PIP	60.97	58.82	61.34	57.52	60.65	57.43	55.30	55.21	60.82	58.49	59.82	57.49
MEL+DGCN	**63.33**	**61.80**	**62.50**	**61.08**	**62.76**	**60.24**	**61.68**	**61.45**	**62.05**	**61.32**	**62.42**	**61.18**

Table 5: Performance on Essays which is labeled by Big 5 taxonomy. For each metric, best results are marked in bold.

Method	ACC	F1	Excat ACC
bert-uncased-base	54.72	52.17	5.8
MLM	75.29	71.02	32.13
MLM+MLL	80.34	76.88	44.93
E5-Mistral-7B	82.69	79.82	46.75
Ours	**94.94**	**92.93**	**77.20**

Table 6: Different embedding performance in personality detection task.

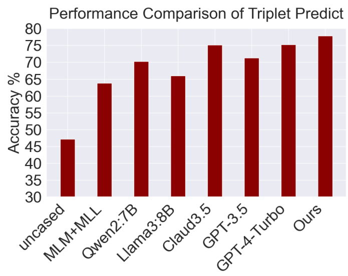


Figure 4: Comparison of Prediction Accuracy for Different Types of Triplets.

methods' retrieved content is more similar to the query. In cases where the similarity was difficult to determine, annotators could select the option "uncertain." As shown in Figure 5, our method demonstrates a clear advantage across various objective and subjective evaluation metrics.

6.3 Discussion

Compare Different Embeddings. It is worth noting that there is no off-the-shelf personality embedding method available yet, so we compare performance on downstream tasks with other common deep learning methods and text-embedding LLM. As shown in Table 6, the triplet learning method demonstrates a significant performance advantage.

Triplet Prediction. To evaluate the metric properties of personality representations, we directly compare the performance of different models on the triplet selection task. Specifically, we use the untrained BERT-base-Chinese model, a BERT model pre-trained with the Musk Language Model and multi-label learning techniques, open-source large language models (Touvron et al., 2023; Chu et al., 2024), and closed-source models (Anthropic, 2024; Achiam et al., 2023) for comparison. The results are shown in Figure 4, benefiting from strong generalization capabilities, large language models perform better on the triplet selection task than traditional deep learning methods, but our methods outperforms all mainstream closed-source large models, demonstrating the effect of triplet learning in personality perception.

Method	Negative	Netural	Positive	M-avg	W-avg
RoBERTa	0.415	0.430	0.323	0.389	0.390
RoBERTa-P	0.401	0.505	0.176	0.361	0.430
PET-CLS	0.492	0.474	0.327	0.431	0.445
Ours	**0.552**	**0.514**	**0.527**	**0.531**	**0.536**

Table 7: Results for sentiment prediction for dialogue emotion translation about F1-score. M-avg and W-avg indicate macro-averaged and weighted-averaged F1.

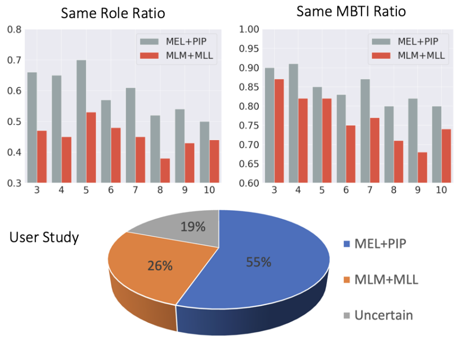


Figure 5: Comparison of Personality Retrieval. The bar chart illustrates the differences in retrieval results between the two embedding learning methods, specifically in terms of the proportion of same-source and same-personality roles. The pie chart represents the subjective results from our user study.

7 Conclusion

In this work, we propose a new task paradigm in text-based personality learning. We construct a triplet personality dataset to address the trade-off between data scale and annotation quality in traditional personality-text datasets. Based on this dataset, we obtain pretrained personality representation with desirable metric properties through triplet learning. We then integrate personality embeddings into various downstream applications and experimentally demonstrate the transferability of the triplet-learned embeddings. Although this work may seem anti-trend in the era of large models, we hope to spark interest among NLP researchers in the practical applications of personality analysis.

Limitations

To inspire future research, we summarize the limitations of our methods. Firstly, due to regional and platform restrictions, the constructed triple dataset contains only a single language. In the future, this approach can be extended to explore cross-linguistic personality representation learning. Secondly, the design of triplet learning is relatively straightforward, and there remains room for exploration in utilizing triplet data to learn better personality representations. Thirdly, the application of personality representation in the recommendation domain has yet to be explored. The coupling effects of personality representation in friend recommendation and music recommendation warrant further investigation.

Ethics Statement

We are following the ethics code for psychological research by which researchers may dispense with informed consent of each participant for archival research, for which disclosure of responses would not place participants at risk of criminal or civil liability, or damage their financial standing, employability, or reputation, and if confidentiality is protected. The PTCD dataset adheres to ethical standards through its compliance with open-source licensing agreements (CC-BY 4.0) for all data sources (as announced in Section 3), ensuring lawful reuse. Annotation was conducted via the YouLing platform, where annotators signed legally binding labor agreements and provided informed consent[5], guaranteeing fair compensation and regulatory compliance.

Acknowledgments

This paper is supported by the Fundamental Research Funds for the Zhejiang Provincial Universities (No. 226-2025-00004) and CCF-NetEase ThunderFire Innovation Research Funding (NO.202516).

[5] For details of the agreement, please refer to https://corp.163.com/gb/legal.html.

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*.

Mohammad Hossein Amirhosseini and Hassan Kazemian. 2020. Machine learning approach to personality type prediction based on the myers–briggs type indicator®. *Multimodal Technologies and Interaction*, page 9.

Na Min An, Sania Waheed, and James Thorne. 2024. Capturing the relationship between sentence triplets for llm and human-generated texts to enhance sentence embeddings. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 624–638.

Anthropic. 2024. Claude 3.5 sonnet. https://www.anthropic.com/news/claude-3-5-sonnet.

Shlomo Argamon, Sushant Dhawle, Moshe Koppel, and James W Pennebaker. Lexical predictors of personality type.

Nuo Chen, Yan Wang, Haiyun Jiang, Deng Cai, Yuhan Li, Ziyang Chen, Longyue Wang, and Jia Li. 2023. Large language models meet harry potter: A bilingual dataset for aligning dialogue agents with characters. *Preprint*, arXiv:2211.06869.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, et al. 2024. Qwen2-audio technical report. *arXiv preprint arXiv:2407.10759*.

Alexis Conneau and Douwe Kiela. 2018. Senteval: An evaluation toolkit for universal sentence representations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Philip J Corr and Gerald Matthews. 2020. *The Cambridge handbook of personality psychology*. Cambridge University Press.

Brandon Cui and Calvin Qi. 2017. Survey analysis of machine learning methods for natural language processing for mbti personality type prediction. *Final Report Stanford University*.

Alexander Philip Dawid and Allan M Skene. 1979. Maximum likelihood estimation of observer error-rates using the em algorithm. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 28(1):20–28.

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

Luciano Floridi and Massimo Chiriatti. 2020. Gpt-3: Its nature, scope, limits, and consequences. *Minds and Machines*, 30:681–694.

James Foulds and Eibe Frank. 2010. A review of multi-instance learning assumptions. *The knowledge engineering review*, 25(1):1–25.

ME Francis and Roger J Booth. 1993. Linguistic inquiry and word count. *Southern Methodist University: Dallas, TX, USA*.

Adrian Furnham. 1996. The big five versus the big four: the relationship between the myers-briggs type indicator (mbti) and neo-pi five factor model of personality. *Personality and individual differences*, 21(2):303–307.

Zhe Gan, Yunchen Pu, Ricardo Henao, Chunyuan Li, Xiaodong He, and Lawrence Carin. 2017. Learning generic sentence representations using convolutional neural networks. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 2390–2400. Association for Computational Linguistics.

Jingsheng Gao, Yixin Lian, Ziyi Zhou, Yuzhuo Fu, and Baoyuan Wang. 2023. LiveChat: A large-scale personalized dialogue dataset automatically constructed from live streaming. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15387–15405, Toronto, Canada. Association for Computational Linguistics.

T Gao, X Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. In *EMNLP 2021-2021 Conference on Empirical Methods in Natural Language Processing, Proceedings*.

Tao Ge, Xin Chan, Xiaoyang Wang, Dian Yu, Haitao Mi, and Dong Yu. 2024. Scaling synthetic data creation with 1,000,000,000 personas. *arXiv preprint arXiv:2406.20094*.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

Felix Hill, Kyunghyun Cho, and Anna Korhonen. 2016. Learning distributed representations of sentences from unlabelled data. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1367–1377.

Robert Hogan. 2017. *Personality and the fate of organizations*. Psychology Press.

Linmei Hu, Hongyu He, Duokang Wang, Ziwang Zhao, Yingxia Shao, and Liqiang Nie. 2024. Llm vs small model? large language model based text augmentation enhanced personality detection model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18234–18242.

Rong Hu and Pearl Pu. 2011. Enhancing collaborative filtering systems with personality information. In *Proceedings of the fifth ACM conference on Recommender systems*, pages 197–204.

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.

Ryan Kiros, Yukun Zhu, Ruslan Salakhutdinov, Richard S. Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Skip-thought vectors. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 3294–3302.

Mariusz Kleć, Alicja Wieczorkowska, Krzysztof Szklanny, and Włodzimierz Strus. 2023. Beyond the big five personality traits for music recommendation systems. *EURASIP Journal on Audio, Speech, and Music Processing*, 2023(1):4.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.

Juho Lee, Yoonho Lee, Jungtaek Kim, Adam Kosiorek, Seungjin Choi, and Yee Whye Teh. 2019. Set transformer: A framework for attention-based permutation-invariant neural networks. In *International conference on machine learning*, pages 3744–3753. PMLR.

Cheng Li, Ziang Leng, Chenxi Yan, Junyi Shen, Hao Wang, Weishi MI, Yaying Fei, Xiaoyang Feng, Song Yan, HaoSheng Wang, Linkang Zhan, Yaokai Jia, Pingyu Wu, and Haozhen Sun. 2023. Chatharuhi: Reviving anime character in reality via large language model. *Preprint*, arXiv:2308.09597.

Zheng Li, Dawei Zhu, Qilong Ma, Weimin Xiong, and Sujian Li. 2024. Eerpd: Leveraging emotion and emotion regulation for improving personality detection. *arXiv preprint arXiv:2406.16079*.

Meng Liu, Hongyang Gao, and Shuiwang Ji. 2020. Towards deeper graph neural networks. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 338–348.

Yifan Liu, Wei Wei, Jiayi Liu, Xianling Mao, Rui Fang, and Dangyang Chen. 2022. Improving personality consistency in conversation by persona extending. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pages 1350–1359.

Veronica Lynn, Niranjan Balasubramanian, and H Andrew Schwartz. 2020. Hierarchical modeling for user personality prediction: The role of message-level attention. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 5306–5316.

Robert R McCrae and Paul T Costa Jr. 1989. Reinterpreting the myers-briggs type indicator from the perspective of the five-factor model of personality. *Journal of personality*, 57(1):17–40.

Yash Mehta, Navonil Majumder, Alexander Gelbukh, and Erik Cambria. 2020. Recent trends in deep learning based personality detection. *Artificial Intelligence Review*, 53(4):2313–2339.

Jianmo Ni, Gustavo Hernandez Abrego, Noah Constant, Ji Ma, Keith Hall, Daniel Cer, and Yinfei Yang. 2022. Sentence-t5: Scalable sentence encoders from pretrained text-to-text models. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1864–1874.

James W Pennebaker. 2001. Linguistic inquiry and word count: Liwc 2001.

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.

Konstantinos V Petrides and Stella Mavroveli. 2018. Theory and applications of trait emotional intelligence. *Psychology: The Journal of the Hellenic Psychological Society*, 23(1):24–36.

Md Abdur Rahman, Asif Al Faisal, Tayeba Khanam, Mahfida Amjad, and Md Saeed Siddik. 2019. Personality detection from text using convolutional neural network. In *2019 1st international conference on advances in science, engineering and robotics technology (ICASERT)*, pages 1–6. IEEE.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 3980–3990. Association for Computational Linguistics.

Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander Alemi. 2017. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31.

Tommy Tandera, Derwin Suhartono, Rini Wongso, Yen Lina Prasetio, et al. 2017. Personality prediction system from facebook users. *Procedia computer science*, 116:604–611.

Edward P. Tighe, Jennifer C. Ureta, Bernard Andrei L. Pollo, Charibeth K. Cheng, and Remedios de Dios Bulos. 2016. Personality trait classification of essays with the application of feature reduction. In *Proceedings of the 4th Workshop on Sentiment Analysis where AI meets Psychology (SAAIP 2016) co-located with 25th International Joint Conference on Artificial Intelligence (IJCAI 2016), New York City, USA, July 10, 2016*, volume 1619 of *CEUR Workshop Proceedings*, pages 22–28. CEUR-WS.org.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Quan Tu, Shilong Fan, Zihang Tian, Tianhao Shen, Shuo Shang, Xin Gao, and Rui Yan. 2024. Charactereval: A chinese benchmark for role-playing conversational agent evaluation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 11836–11850. Association for Computational Linguistics.

Xiujuan Wang, Yi Sui, Kangfeng Zheng, Yutong Shi, and Siwei Cao. 2021. Personality classification of social users based on feature fusion. *Sensors*, 21(20):6758.

Zhiyuan Wen, Jiannong Cao, Ruosong Yang, Shuaiqi Liu, and Jiaxing Shen. 2021. Automatically select emotion for response via personality-affected emotion transition. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 5010–5020.

Ross B Wilkinson and Wendy A Walford. 2001. Attachment and personality in the psychological health of adolescents. *Personality and Individual Differences*, 31(4):473–484.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

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

Feifan Yang, Xiaojun Quan, Yunyi Yang, and Jianxing Yu. 2021a. Multi-document transformer for personality detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 14221–14229.

Shenzhi Yang, Jun Xia, Jingbo Zhou, Xingkai Yao, and Xiaofang Zhang. 2025. Nodereg: Mitigating the imbalance and distribution shift effects in semi-supervised node classification via norm consistency. *arXiv preprint arXiv:2503.03211*.

Tao Yang, Jinghao Deng, Xiaojun Quan, and Qifan Wang. 2023a. Orders are unwanted: dynamic deep graph convolutional network for personality detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 13896–13904.

Tao Yang, Tianyuan Shi, Fanqi Wan, Xiaojun Quan, Qifan Wang, Bingzhe Wu, and Jiaxiang Wu. 2023b. Psycot: Psychological questionnaire as powerful chain-of-thought for personality detection. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 3305–3320. Association for Computational Linguistics.

Tao Yang, Feifan Yang, Haolan Ouyang, and Xiaojun Quan. 2021b. Psycholinguistic tripartite graph network for personality detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4229–4239.

Minghua Zhang, Yunfang Wu, Weigang Li, and Wei Li. 2018. Learning universal sentence representations with mean-max attention autoencoder. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4514–4523.

Tong Zhao, Yozen Liu, Leonardo Neves, Oliver Woodford, Meng Jiang, and Neil Shah. 2021. Data augmentation for graph neural networks. In *Proceedings of the aaai conference on artificial intelligence*, volume 35, pages 11015–11023.

Yinhe Zheng, Guanyi Chen, Minlie Huang, Song Liu, and Xuan Zhu. 2019. Personalized dialogue generation with diversified traits. *arXiv preprint arXiv:1901.09672*.

Yinhe Zheng, Rongsheng Zhang, Minlie Huang, and Xiaoxi Mao. 2020. A pre-training based personalized dialogue generation model with persona-sparse data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9693–9700.

Zhi-Hua Zhou, Min-Ling Zhang, Sheng-Jun Huang, and Yu-Feng Li. 2012. Multi-instance multi-label learning. *Artificial Intelligence*, 176(1):2291–2320.

Guangcheng Zhu, Ruixuan Xiao, Haobo Wang, Zhen Zhu, Gengyu Lyu, and Junbo Zhao. 2025. Large margin representation learning for robust cross-lingual named entity recognition. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4270–4291.

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

A More Details of Implement

A.1 Details of Dataset

In personality detection task, we primarily utilized our own constructed PTCD dataset, along with two publicly available personality datasets: Kaggle[6] and Essays (Pennebaker and King, 1999).

We collect high-quality corpora from different domains to construct the PTCD dataset. In the begining, we collect 6,145,276 utterances for Dialogue domain, 230,795 dialogue instances of 1,388 fictional characters for Literature Domain and 237,562 utterances for Description domain. After preprocessing, the distribution of data in different domains can be referred to Table 8. During the triplet proposal phase, we generated a total of 40,000 triplets. After annotation completion, we applied rigorous filtering based on annotation consistency (using the Dawid-Skene algorithm (Dawid and Skene, 1979) with a 0.2 confidence threshold), resulting in 36,294 high-quality annotated triplets.

The Kaggle dataset is sourced from PersonalityCafe[7] and contains 45 to 50 social media posts for each of the 8,675 users along with their corresponding MBTI personality type.

The Essays dataset is a comprehensive collection of text data tailored for personality recognition tasks, specifically emphasizing the Big 5 personality traits. Under specific guidelines, volunteers were asked to write freely to convey their thoughts within a set time limit. This dataset includes 2,468 long texts, each paired with the corresponding author's Big Five personality traits.

Have to note that the Pandora dataset[8] is also widely used in related work; however, access to this dataset requires permission from the author. Unfortunately, we did not receive response from the author before the completion of this paper.

A.2 Details of Data Processing in Essays

In the Essays dataset, the text from each volunteer is concatenated into a long passage. In our learning paradigm, this long text needs to be segmented into individual sentences. Text segmentation is not the focus of this study, so we adopt a straightforward punctuation-based approach: sentences are divided based on periods, question marks, and exclamation marks.

Domain	Dialogue	Literature	Description	All
Num	49,996	55,459	57,050	162,505

Table 8: The distribution of different domains of corpus in PTCD.

A.3 Details of Baselines

The details of baseline methods used in the personality detection task are as follows: LIWC+SVM (Tighe et al., 2016) uses LIWC (Pennebaker, 2001) to extract psycholinguistic features and applies SVM as the classifier. W2V+CNN (Rahman et al., 2019) uses non-pretrained CNN network (LeCun et al., 1998)with word2vec algorithm to learn text embedding. AttRCNN (Xue et al., 2018) utilizes a hierarchical structure that incorporates a variant of Inception (Szegedy et al., 2017) to encode each post. DDGCN (Yang et al., 2023a) employs a domain-adapted BERT to encode each post, along with a dynamic deep graph network to aggregate posts in a non-sequential manner. TAE (Hu et al., 2024) enhances the performance of smaller models in personality detection by utilizing text augmentations from LLM and employing contrastive learning techniques. PsyCoT (Yang et al., 2023b) employs psychological questionnaires as chain-of-thought (CoT) process, utilizing LLM to conduct multi-turn dialogue evaluations.

A.4 Details of Multi-head Attention Blocks

Regarding the specific structure of multi-head attention blocks (MAB) is as follows:

$$\begin{aligned}
&\mathrm{MAB}(Q,K) = \mathrm{LayerNorm}(O+\mathrm{FNN}(O)) \\
&\textbf{where } O = \mathrm{LayerNorm}(Q+M) \\
&\textbf{and } M = \mathrm{Multihead}(Q,K,K;\omega) \\
&\textbf{and } \mathrm{FNN}(O) = \mathrm{ReLU}(\mathrm{FC}(O))
\end{aligned} \tag{4}$$

where Q, K are both set inputs, ω is a hyperparameter that controls the number of attention heads.

A.5 Details of Transfer Personality Detection

Given the significant differences in language and domain between our triplet dataset and these open-source datasets, we opted to translate the PTCD dataset into English for training and then evaluated its transfer performance on the open-source datasets.

Additionally, pooling architecture mentioned in section 5.1 only has two layers to learn, which is

[6] https://www.kaggle.com/datasnaek/mbti-type

[7] http://personalitycafe.com/forum

[8] https://psy.takelab.fer.hr/datasets/all

Method	PIP		DGCN	
	PTCD	Kaggle	PTCD	Kaggle
Epoch Time	6.71s	4.02s	70.10s	58.42s

Table 9: Comparison of epoch time of different datasets.

Method	ACC	F1	Excat ACC
MLL encoder	80.43	76.88	44.93
Embedding Fusion	86.34	84.19	60.12
Single Classifier	89.49	87.50	66.67
Ours	**94.94**	**94.94**	**77.20**

Table 10: Ablation study on PTCD dataset. ACC refers to the accuracy of a single dimension under the MBTI taxonomy, while Exact ACC denotes the probability of having all four dimensions accurately classified.

too simple to maintain excellent performance in domain adaptation scenarios. To better capture the information between corpora, we follow previous work (Yang et al., 2023a). Based on the trained personality representation, we dynamically compute the adjacency graph and use DGCN (Liu et al., 2020) to complete the aggregation of features. In this manner, we outperformed domain-specific training methods in domain adaptation scenarios across different open-source datasets.

B Additional Experimental Results

B.1 Training Complexity

To demonstrate the advantage of our PIP architecture in terms of time complexity, we compared the time required for training convergence using PIP and DGCN. As shown in Table 9, PIP has a significant advantage over DGCN in terms of time complexity.

B.2 Ablation Study of Personality Detection

To demonstrate the effectiveness of the two components separately, we also tested the settings of training on PTCD dataset. To be specific, we test three ablation settings:

- MLL-encoder: In this setup, we stop using triplet loss to train the encoder. Instead, we split the PTCD dataset into individual sentences for multi-label learning to assess the importance of measurable embedding learning.
- Embedding Fusion: This setting replaces the Permutation-Invariant Pooling architecture by averaging the set input embeddings and conducting Multi-Label Classification to validate the effectiveness of our pooling structure.
- Single Classifier: In this scenario, we limited the pooling output to one channel and conducted a 16-class classification using MBTI rules, confirming the effectiveness of transforming the task into a multi-label learning format.

Method	5	10	15	20
Zero-Shot	31.64	33.47	32.02	33.48
Few-Shot (MLM + MLL)	29.62	30.76	30.14	29.76
Few-Shot (Ours)	**38.73**	**35.76**	**37.84**	**36.75**

Table 11: The ratio of $rank = 1$ cases to total cases across three methods under different numbers of examples, specifically for $N = 5, 10, 15,$ and 20.

Method	5	10	15	20
Few-Shot (MLM + MLL)	56.24	54.69	56.70	57.44
Few-Shot (Ours)	**58.58**	**58.97**	**60.01**	**59.30**

Table 12: The win rate relative to *Zero-Shot* under different numbers of examples, specifically for $N = 5, 10, 15,$ and 20.

As shown in Table 10, all ablation settings resulted in varying degrees of performance degradation, demonstrating the effectiveness of each component of our framework. Additionally, we can observe the importance ranking of measurable embedding learning, pooling architecture, and classifier.

B.3 Retrieval-based enhancements for role-playing agent

In game development, ensuring the dialogues from NPC align with their predefined personalities is crucial, particularly for user-customized companion NPCs. To evaluate the effectiveness of embedding models for role-playing agents, we investigate the changes of agents' responses before and after introducing retrieval examples, assessing the alignment of these responses with the character's personality. To ensure equitable evaluation, we construct a dataset based on CharacterEval (Tu et al., 2024) that retains test data and character descriptions containing personality information. Additionally, it includes a dialogue library derived from dialogue snippets from CharacterEval.

We compare three methods: 1) *Zero-Shot*, using the standard prompt from CharacterEval, to generate responses; 2) *Few-Shot (MLM + MLL)*,

Table 13: Prompt of Retrieval-Based Role-Playing Agent

# Prompt of Retrieval-Based Role-Playing Agent

Now, please act as a role-play expert. Based on the information provided below, engage in a conversation by impersonating the role of role and strictly adhere to the character's traits.

{role_information}

Below are text samples with similar characteristics for your reference:

{examples through retrieval}

Dialogue history as follows:

{dialogue context}

employing a BERT model trained using MLM and MLL to select dialogues that exhibit similar personalities as examples, guiding the agent to generate responses; and 3) *Few-Shot (Ours)*, applying our method to retrieve similar dialogues. In the latter two methods, we utilize embedding models trained on our PTCD datasets to encode personality descriptions and select the N most similar dialogues from the library based on cosine similarity, thereby enhancing the consistency of the agent's responses with the target personality. To measure the similarity between the responses and the personality descriptions, we use a judge model to rank the outputs generated by the three methods. Specifically, we utilize GPT-3.5 as the agent and GPT-4 for the judge model.

As illustrated in Table 11 and Table 12, our method consistently outperforms competing approaches across both evaluation metrics: the proportion of $rank = 1$ samples and the win rate. This performance advantage is particularly pronounced when the number of examples $N = 15$. The empirical results indicate that the samples retrieved by our method demonstrate higher relevance to character personality traits, thereby providing robust evidence for the superior performance of our method.

C Details of Prompt

We provide the LLM prompts used in the paper below.

D Analysis about MBTI and Big 5

Firstly, both the MBTI and Big5 frameworks have been extensively studied in existing works (Hu et al., 2024; Yang et al., 2023b,a). Psychological studies (Furnham, 1996; McCrae and Costa Jr, 1989) have also demonstrated correlations between the dimensions of these two frameworks. Furthermore, our experiments also demonstrate that our method achieves SOTA performance on both MBTI and Big5 classification tasks (as shown in Table 2 and Table 3), which fully proves the generalization of our work. While future research may increasingly focus on Big5-based studies, we believe that this novel learning paradigm is more critical for personality representation research.

# *Personality Conception*

Assume the role of a certified personality psychology expert analyzing fictional character profiles. Your task is to generate original personality descriptions using the MBTI framework, strictly avoiding verbatim content from the input text. Follow this exact structure:
1. Extraversion (E) vs. Introversion (I)
State preference: Analyze energy source and social interaction patterns with 2 specific behavioral examples
2. Sensing (S) vs. Intuition (N)
State preference: Describe information processing style with 1 concrete decision-making example
3. Thinking (T) vs. Feeling (F)
State preference: Explain decision-making approach with 1 conflict resolution example
4. Judging (J) vs. Perceiving (P)
State preference: Characterize lifestyle orientation with 1 time-management/crisis-handling example
Additional Requirements:
Use plain language avoiding technical terms while maintaining psychological accuracy
For ambiguous traits: specify "Insufficient evidence to determine" with rationale
Maintain third-person perspective throughout
Keep total length under 150 words
Character Profiles: { Character Profiles }

Table 14: Prompt of Triplet Mining.

# *Sentence Filter*

Please determine whether the following text contains information that reflects the personality of the speaker. Please base your judgment on the following criteria:
Important criteria: Opinion and attitude - Express a clear opinion or attitude.
Auxiliary criteria: Language style - a particular way of speaking is used.
Please answer "yes" or "no" according to the above criteria.
Examples:
1."What's the use of being smart when you're all alone in the end?" -Yes
2. "Nice weather." - No
Text under test: utterance
Note: Only reply "yes" or "no" and give a reason of no more than 50 words.

# *Obtain Personality Tendcy*

Based on the following sentence, determine this person's MBTI personality type.
Sentence:
{corpus}
MBTI personality types include the following options:
–E (Extraversion)/I (Introversion)
–S (Sensing)/N (Intuition)
–T (Thinking)/F (Feeling)
–J (Judging)/P (Perceiving)
Please analyze the corpus and determine the most suitable MBTI personality type combination.
Note: Output only the personality type combination without any additional content.

Table 15: Prompt of Personality Detection.

# *Prompt of Personality Detection*

Judge the person’s MBTI personality label based on the following text paragraphs. These texts are descriptions of a person’s actions, thoughts, and feelings.
Text paragraph:
{corpus}
The MBTI Personality label has the following options:
1. E (Extroversion) /I (introversion)
2. S (Sense of feeling) /N (intuition)
3. T (thinking) /F (feeling)
4. J (Judgment) /P (perception)
Please analyze the most appropriate MBTI personality label combination based on the corpus.
Note: Output personality label combination directly, do not output other content.

# *Examples*

Example: The posts of this user are:’Wow, thank you for this thread! Physical vs. metaphysical is a great topic! I find that I am very much the same way your are. How can I put it ... I have my days. :) The more I develop my xSxJ, the ...||| my room. I like to be in my bed, next to my books, with my fan on and laptop nearby.||| I wouldn’t say that I can read souls-but I can see potential. I can sense sadness, happiness, uneasiness! etc. I can tell when someone is not happy where they are and with what they are doing...||| thank you for being so polite! :) ||| I find eye contact is key. I acknowledge their existence and importance by maintaining eye contact with them throughout the conversation. Not by staring into their eyes in a creeper way, but by making ...||| As an INFJ male I can somewhat relate to your post. A very close lady friend of mine and I were like this for years! I had always liked her and could read her fairly well. I knew when she needed ...’
Result: INFJ