From Classification to Ranking: Enhancing LLM Reasoning Capabilities for MBTI Personality Detection

Yuan Cao[1], Feixiang Liu[2], Xinyue Wang[1], Yihan Zhu[3], Hui Xu[1], Zheng Wang[1], Qiang Qiu[1]

[1]Institute of Computing Technology, Chinese Academy of Sciences
[2]University of Chinese Academy of Sciences
[3]Beihang University
caoyuan, wangxinyue, xuhui, wangzheng, qiuqiang@ict.ac.cn, liufeixiang@mails.ucas.ac.cn, zhuyihan@buaa.edu.cn

Abstract

Personality detection aims to measure an individual's corresponding personality traits through their social media posts. The advancements in Large Language Models (LLMs) offer novel perspectives for personality detection tasks. Existing approaches enhance personality trait analysis by leveraging LLMs to extract semantic information from textual posts as prompts, followed by training classifiers for categorization. However, accurately classifying personality traits remains challenging due to the inherent complexity of human personality and subtle inter-trait distinctions. Moreover, prompt-based methods often exhibit excessive dependency on expert-crafted knowledge without autonomous pattern-learning capacity. To address these limitations, we view personality detection as a ranking task rather than a classification and propose a corresponding reinforcement learning training paradigm. First, we employ supervised fine-tuning (SFT) to establish personality trait ranking capabilities while enforcing standardized output formats, creating a robust initialization. Subsequently, we introduce Group Relative Policy Optimization (GRPO) with a specialized ranking-based reward function. Unlike verification tasks with definitive solutions, personality assessment involves subjective interpretations and blurred boundaries between trait categories. Our reward function explicitly addresses this challenge by training LLMs to learn optimal answer rankings. Comprehensive experiments have demonstrated that our method achieves state-of-the-art performance across multiple personality detection benchmarks.

Introduction

Language-based personality detection is crucial for AI applications such as mental health screening and conversational agents, utilizing established psychological frameworks like Myers-Briggs Type Indicator (MBTI) (Myers 1962; Myers and McCaulley 1985). As illustrated in Figure 1, MBTI characterizes personality along four interdependent dimensions: Extraversion (E) vs. Introversion (I), Sensing (S) vs. iNtuition (N), Thinking (T) vs. Feeling (F), and Judging (J) vs. Perceiving (P). Crucially, these dimensions exhibit dynamic interactions governed by hierarchical cognitive functions that collectively shape behavioral expressions. For instance, in MBTI theory, an INTJ's dominant function is introverted intuition (Ni), while ENTJ's dominant function is extraverted thinking (Te), leading to different cognitive profiles even if only one letter differs. Prevailing approaches (Yang et al. 2021b; Zhu et al. 2024) fundamentally limit their efficacy by treating MBTI prediction as four independent binary classification tasks ("E/I", "S/N", "T/F", "J/P"), resulting in two critical shortcomings: dimensional decoupling that ignores psychometrically established interactions, static labeling that fails to capture complete personality expressions which neglects the position on the scale (Stajner and Yenikent 2021). While traditional models struggle with this interdependence, the recent success of Large Language Models (LLMs) offers new opportunities to approach personality modeling in a more holistic and reasoning-driven manner.

Currently, LLMs exhibit remarkable reasoning capabilities across diverse NLP tasks (Brown et al. 2020; Wei et al. 2022), but their application to personality assessment reveals fundamental limitations. Prompt-based methods such as PsyCoT (Yang et al. 2023b) mimic the way individuals complete psychological questionnaires through multi-turn dialogues. However, the limitations of prompt-based approaches are their over-reliance on expert-defined prompts and large-scale parameter models (Le et al. 2024; Islam et al. 2025), which makes them impractical for real-time applications due to high inference latency and computational cost. This gap necessitates novel frameworks capable of autonomously learning discriminative patterns from behavioral data.

To resolve these limitations, we pioneered a paradigm shift from categorical classification to preference ranking, a transformation aligned with psychometric realities where personality manifests itself through relative trait preferences rather than absolute categorical assignments (Heston, Thomas F and Gillette, Justin 2008; Hüllermeier et al. 2025). This ranking-centric approach directly addresses the core challenge of modeling dimensional interdependencies by conceptualizing personality assessment as a continuous spectrum rather than discrete categories.

Implementing this paradigm shift, we introduce PerDet-R1, a two-stage framework that eliminates categorical fallacies through ranking-based modeling. In the first stage, we perform supervised fine-tuning (SFT) by distilling comparative reasoning traces from advanced LLMs Qwen-plus

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

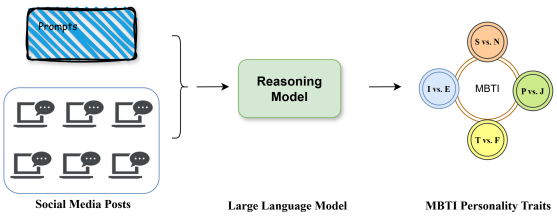

Figure 1: A framework for LLM-based MBTI personality detection. The framework leverages a single user's social media posts as input to corresponding personality traits along four binary dimensions of Introversion vs. Extroversion, Sensing vs. iNtuition, Think vs. Feeling, and Perception vs. Judging.

to generate ground-truth-aligned MBTI preference rankings from social media data, training a 7B-parameter LLM to output ordered type lists with explicit awareness of dimensional combinations and interactions. The second stage implements Group Relative Policy Optimization (GRPO), initializing the policy with our stage-one model and optimizing ranking positions via a Normalized Discounted Cumulative Gain (NDCG) reward function, thereby amplifying trait-cohort interactions through position-sensitive reinforcement while eliminating expert prompt dependency.

In summary, our core contributions are as follows:

- We propose the first list-wise ranking paradigm for personality detection, which eliminates the categorical fallacy inherent in existing methods.
- We introduce a trait-cohort advantage estimation for the reward model, which calculates the reward score by dimension similarity and NDCG to amplify subtle interactions, generating dense reward signals to effectively avoid the problem of reward hacking.
- Our method achieves state-of-the-art results on mainstream personality detection datasets, establishing a new standard for data-driven personality assessment.

Related Work

Personality Detection

Pre-trained language models (e.g., BERT, RoBERTa) dominated personality detection before the emergence of large language models (LLMs). While demonstrating performance gains through encoding concatenated user utterances, these approaches neglected temporal dynamics and inter-post relationships (Tandera et al. 2017; Xue et al. 2018; Lynn, Balasubramanian, and Schwartz 2020). To address the limitations, Yang et al. (2023a) address the unstructured post-integration challenge in personality detection through Deep Dynamic Graph Convolutional Network (D-DGCN). Departing from prior graph-based methods that require predefined adjacency structures. Concurrently with architectural innovations, LLM-based reasoning techniques emerged, and Chain-of-Thought (CoT) prompting has been introduced to steer LLMs in generating explicit reasoning pathways. Building upon standard CoT, PsyCoT (Yang et al. 2023b) innovatively adapts psychological questionnaires into a structured CoT framework. This approach simulates human self-assessment procedures via multi-turn dialogues, directing LLMs to focus on specific test items while leveraging historical ratings for more consistent and accurate reasoning. Hu et al. (2024) address the data scarcity challenge in personality detection through LLM-based text augmentation. Their method distills LLM capabilities (comprehension, summarization, sentiment analysis) to generate multi-perspective text analyses, semantic, affective, and linguistic. Bi et al. (2025) propose the LLM-Enhanced Text Mapping Model (ETM), a hybrid architecture addressing core limitations in traditional personality detection systems. ETM demonstrates how LLMs can dynamically construct psychologically-grounded representations. Although existing methods achieve notable success, they universally adopt classification-centric frameworks, either selecting among 16 discrete MBTI types or aggregating binary predictions across four dimensions. These approaches fundamentally diverge from our work, which harnesses LLMs' reasoning capabilities to reframe personality detection as comparative ranking. By amplifying inter-type distinctions through relative intensity assessments, our paradigm captures nuanced trait interactions obscured by classification methods, thereby advancing detection performance.

Fine-Tuning Language Models

Beyond architectural innovations, training strategies also play a crucial role in enhancing personality detection capabilities (Wang and Sun 2024; Serapio-García et al. 2025). Fine-tuning has emerged as a pivotal strategy for adapting language models to specialized tasks (Shen et al. 2025). Recent innovations integrate explicit reasoning traces during fine-tuning, where step-by-step rationales guide models through complex problem-solving (e.g., mathematics or coding) (Wei et al. 2022). These traces mitigate hallucination risks and enhance logical coherence, which proves essential for reasoning scenarios. Reinforcement Learning (RL) further augments reasoning capabilities. Algorithms like Group Relative Policy Optimization (GRPO) (Guo et al. 2024) optimize policies via reward signals. Many works use this training method to fine-tune the domain-specific LLMs. In the financial domain, Liu et al. (2025) introduced Fin-R1, which employs a two-stage training framework of SFT followed by RL, significantly enhancing its ability to perform complex financial reasoning tasks. Similarly, Papicchio et al. (2025) proposed Think2SQL, which also utilizes SFT and RL to improve the model's performance in converting natural language questions into SQL queries. These studies collectively demonstrate that the integration of SFT and RL can effectively improve the reasoning capabilities of language models in specific domains, also providing valuable insights for personality detection tasks. Building on these insights for personality detection, our work implements an SFT+GRPO framework with a novel Normalized Discounted Cumulative Gain (NDCG) reward function incorporating MBTI dimensional similarity metrics. This design intrinsically avoids reward hacking through position-aware discounting, enables automatic pattern dis-

covery from behavioral data, resolves expert dependency, and captures cross-dimensional trait interactions, establishing a new paradigm for completed personality-centric reasoning enhancement.

Rule-based Reward Modeling

Rule-based reward modeling has emerged as a critical paradigm for aligning large language models (LLMs), establishing structured computational frameworks that complement or supplant human preference data to enhance training efficiency and policy controllability (Xiong et al. 2023; Tong et al. 2024 ; Dong et al. 2024). This approach addresses fundamental limitations identified in the empirical analysis by Gao et al. (Gao, Schulman, and Hilton 2023) of the overoptimization of the reward model within direct alignment algorithms. Building on these insights, the Logic-RL framework (Xie et al., 2025) pioneers a dual-component reward architecture comprising Format Reward, which enforces structural compliance through formal grammar and reasoning pattern; Validation-Answer Reward, ensuring semantic fidelity via knowledge-constrained verification mechanisms. This integrated system synergistically guides models toward logically structured reasoning pathways while intrinsically mitigating reward hacking through cross-component dependency checks, where conflicting signals trigger differential reward attenuation. By leveraging formally specified rules and hierarchically structured rewards, the paradigm establishes robust behavioral guardrails against optimization shortcuts while providing domain-adaptable alignment frameworks and computationally tractable supervision alternatives, demonstrating particular efficacy in domains requiring simultaneous procedural correctness and substantive accuracy.

Method

We formulate personality detection as a multiclass ranking task, where the goal is to identify and rank the most relevant MBTI types for a user based on their social media posts. Formally, given a set $P = \{p_1, p_2, ..., p_n\}$ of social media posts by a user, the goal is to select the most relevant Top-K MBTI personality option $Y = \{y_1, y_2, ..., y_n\}$ for this user based on P. As the overall architecture shows in Figure 2, we adopt a two-stage LLM post-training process, namely SFT and Reasoning RL. Both need to transform the data to a specific format for model fine-tuning. The first stage focuses on developing LLM's reasoning abilities and also distills domain knowledge from the teacher model into the student model to achieve the effect of the cold start. The second stage aims to push LLM to compare the MBTI personality type for the candidate and select the most relevant answers. Followed by the presentation of some exciting results, and hope this provides the community with valuable insights.

Training Data Construction

The dataset construction involves a rigorous two-stage process. The purpose of each stage of fine-tuning is different, so the form and content of data are also different, which requires the reconstruction and expansion of the dataset to adapt to the fine-tuning requirements of LLM in different stages.

The original dataset V_{origin} is the public dataset that was collected from social media platforms. Each sample v in the original dataset contains two components: the posts and the personality label of this user. This format does not apply to fine-tuning generative language models. To build the dataset V for SFT, we utilize the Qwen-plus model with a specialized prompt, as shown in Figure 3, to generate the reasoning trace and the most relevant answers.

During the first stage, each sample v in the SFT training datasets V comprises three components, that is, where P denotes the content of the posts formatted as plain text, R represents the reasoning trace, and is formatted as <think>...</think>, Y corresponds to the Top-K answer list, which is formatted as <answer>[option 1, option 2, ..., option N]<answer>. Since not all personality types can be accurately inferred from social media posts alone, we apply rejection sampling to filter out samples where the ground truth does not appear in the model's Top-K predictions. By evaluating the model output, the sampling method can filter out low-quality inference chains, retaining only those that meet the standards. This ensures that the model's inference chains are not only correct but also highly readable. Specifically, we filter out the items for which the ground-truth label does not appear in the answer list. Finally, 1,000 samples were carefully selected as high-quality SFT training samples. We use them as initial data for cold start, which helps the model get started without a large amount of supervised data. By using carefully designed small amounts of inference data, the model can produce reasonable inference outputs in the early stages of training.

During the second stage, we use the original datasets that exclude corresponding SFT training samples as the RL training dataset since the cold start data should not appear in this stage. In conclusion, Stage 1 aims to instill basic reasoning patterns from a strong teacher model; Stage 2 uses reinforcement learning to optimize trait inference through reward feedback.

Supervised Fine-tuning

To enhance performance on specific tasks, researchers employ instruction tuning (Chung et al. 2024), also known as Supervised Fine-Tuning (SFT) (Ouyang et al. 2022); During this stage, the domain knowledge can be injected to the model (Hsieh et al. 2023) and and enables the model to distinguish different personality traits from social media posts. The objective at this stage is to instill foundational reasoning patterns in the model without overly emphasizing immediate reasoning performance. Additionally, by leveraging the knowledge from large-scale models, computational costs can be reduced.

We initially performed SFT on the instruct model `Qwen2.5-7B-Instruct`, and the training data was sampled from the reasoning model `Qwen-plus`. This method aims to fine-tune the instruct model to replicate the reasoning behavior of the reasoning model, that is, to compare and rank the candidate answers first, and then output the Top-K answers after standardized format processing.

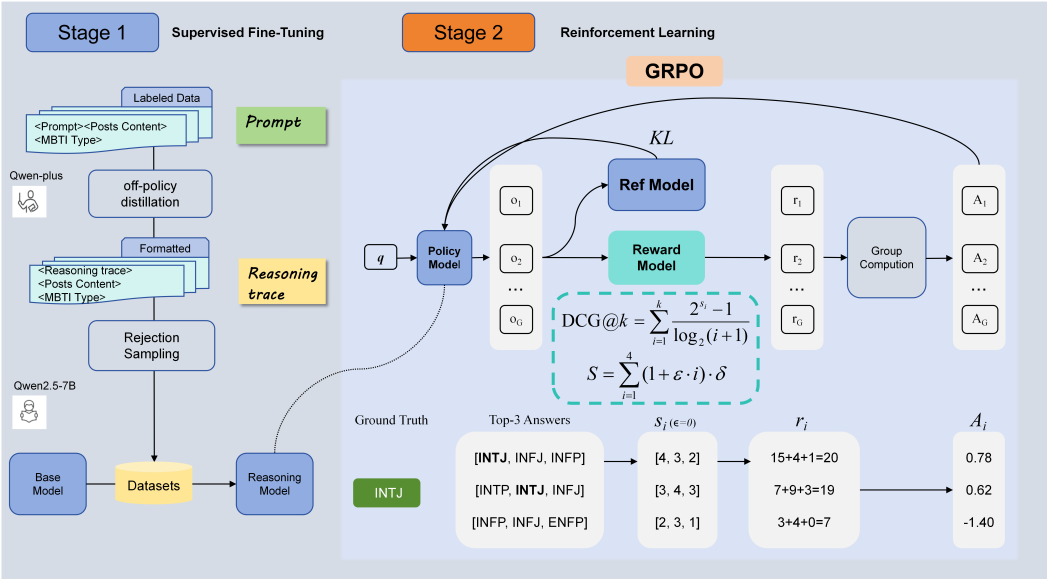

Figure 2: An overview of our PerDet R1, the two-stage LLM post-training workflow. In the first stage, we adopt Qwen-plus as the teacher model to generate a reasoning trace for comparing different MBTI types. Subsequently, the base model learning formatted data via SFT to get a reasoning model. In the second Stage, the reasoning model optimizes via Group Relative Policy Optimization (GRPO) with NDCG and dimension similarity reward function to enhance ranking capability.

Reasoning Reinforcement Learning

Reinforcement learning (RL) now plays a pivotal role in training LLMs. It is used not only to align responses with human preferences through Reinforcement Learning from Human (RLHF) (Ziegler et al. 2019; Ouyang et al. 2022; Lee et al. 2024), but also to enhance models' reasoning abilities, as recently demonstrated by DeepSeek-R1 (Guo et al. 2025). These applications underscore RL's vast potential to drive further advancements in LLMs.

We employ the Group Relative Policy Optimization (GRPO) algorithm, which offers a value-free alternative by computing normalized group-level advantages based directly on realized rewards. Let $\boldsymbol{x} \sim \mathcal{X}$ represent the prompt extracted from the conditional input distribution, and $\{\boldsymbol{y}_i\}_{i=1}^{G} \sim \pi_{\theta_{\text{old}}}(\cdot \mid \boldsymbol{x})$ denote the G response sequences generated by the frozen reference strategy $\pi_{\theta_{\text{old}}}$. For each response $\boldsymbol{y}_i = (y_{i,1}, \ldots, y_{i,T_i})$, the reward model assigns a scalar reward $R_i \in \mathbb{R}$. The relative advantage A_i of the $i-th$ response group is defined by normalizing the distribution of group rewards, that is to say, each response receives a reward r_i, from which they compute the group-relative advantage A_i:

$$A_i = \frac{r_i - \mu_r}{\sigma_r} \tag{1}$$

where μ_r and σ_r denote the mean and standard deviation of reward values within the group.

For each token position t in response y_i , define the state as $s_{i,t} = x || y_{i,<t}$ and the token-level probability ratio as:

$$p_{i,t}(\theta) = \frac{\pi_\theta(y_{i,t} \mid \mathbf{s}_{i,t})}{\pi_{\theta_{\text{old}}}(y_{i,t} \mid \mathbf{s}_{i,t})} \tag{2}$$

The policy update now maximizes the following objective function:

$$L_{\text{GRPO}}(\theta) = E\Big[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{T_i}\sum_{t=1}^{T_i}\min(p_{i,t}(\theta)A_i, \\ clip\left(p_{i,t}(\theta), 1-\epsilon, 1+\epsilon\right)A_i) - \beta KL\big[\pi_\theta \,\|\pi_{\theta_{\text{ref}}}\big]\Big] \tag{3}$$

Where ϵ is the hyper-parameter for clipping parameter and β controls the Kullback-Leibler divergence (KL) regularization by penalizing the deviation from the reference policy model.

Reward Function Design

To reduce LLM training costs, GRPO (Shao et al. 2024) eliminates the value model and relies solely on rule-based reward functions for optimization, making the reward signal critical for guiding exploration. In MBTI personality detection, a naive approach directly outputting personality types would permit straightforward reward assignment but fundamentally conflicts with the need of the GRPO method for differentiated group-wise rewards. Classification-based rewards inherently produce sparse and undifferentiated signals, causing minimal score variance within groups and not reflecting relative advantages. To avoid sparse and undifferentiated reward signals common in classification-based settings, we design a ranking-based reward function using dimension similarity and Normalized Discounted Cumulative Gain (NDCG) that captures both correctness and relevance of the output position.

The NDCG metric is widely adopted in recommendation systems and information retrieval as our reward func-

Prompt

system prompt:
Respond in the following format:
<think>⋯</think>/n<answer>⋯</answer>
user prompt:
Based on the following social medial posts and the official interpretations of 16 MBTI personality types, conduct a comprehensive analysis of top 3 most likely MBTI types, along with a brief explanation of why these types were chosen and their relative ranking. Finally output the top 3 most likely MBTI types in the format [Type1, Type2, Type3].

Reasoning Trace

<think>*The poster mentions feeling uncomfortable and anxious in crowds and prefers observing to taking the lead. Here's the MBTI type* ***ranking from most to least likely:***
ISTP : Introverted, practical, flexible, enjoys solitude, prone to social anxiety, and thinks rationally,most fitting for the posts.
*INTP: Also matches traits like rationality, introversion, and strong observation skills, but focuses more on abstract thinking than ISTP⋯****</think>***

<answer>*[ISTP, INTP, ISFP]****</answer>***

Figure 3: An example of prompts and generated reasoning trace with think and answer label for top-k MBTI types ranking.

tion. NDCG prioritizes relevance and ranking position, assigning a greater weight to items ranked higher in a list. This effectively measures whether the output list contains the ground-truth and the ordinal position of the correct answer in the output list. During the aforementioned Supervised Fine-Tuning (SFT) phase, the model has learned to generate Top-K MBTI types after reasoning analysis. Subsequently, the MBTI dimension similarity scoring is applied to compute the relevance values, which are then leveraged to calculate the final NDCG-based reward score. The NDCG at $k-th$ position $(k \leq K)$ is defined as:

$$\text{NDCG@}k = \frac{\text{DCG@}k}{\text{IDCG@}k} \tag{4}$$

DCG measures the cumulative benefits of actual ranking and discounts the results that are lower in the ranking:

$$\text{DCG@}k = \sum_{i=1}^{k} \frac{2^{s_i} - 1}{\log_2(i+1)} \tag{5}$$

IDCG is the DCG value arranged in descending order of correlation, which represents the theoretical optimal result:

$$\text{IDCG@}k = \sum_{i=1}^{k} \frac{2^{s^{\text{sorted}}} - 1}{\log_2(i+1)} \tag{6}$$

where s_i is the similarity score for ground-truth and the i-th predicted type in the Top K list. Since MBTI consists of four dimensions, each of dimension contains two opposing endpoints, resulting in sixteen different personality types. For these characteristics, we calculate similarity scores using the following formula:

$$\mathrm{s}\left(\hat{\mathrm{t}}, \mathrm{t}^*\right) = \sum_{i=1}^{4} (1 + \epsilon \cdot i) \cdot \delta\left(\hat{t}^{(i)}, t^{*(i)}\right) \tag{7}$$

Where $\epsilon \in \mathbb{R}^+$ denotes the dimension weight coefficient, a hyper-parameter controlling the contribution strength of each dimension's character match, let $T = \{\hat{t}_1, \hat{t}_2, \ldots, \hat{t}_k\}$ denotes the list of predicted MBTI types containing k elements, where each $\hat{t}_j$ represents a complete personality type prediction with four-letter format. And t^* represents the ground-truth from datasets. δ is the character matching function, defined as:

$$\delta(a, b) = \begin{cases} 1 & , a = b \\ 0 & , a \neq b \end{cases} \tag{8}$$

where δ measures the degree of match between the predicted and the ground truth. To avoid the problem of reward hacking in rule-based reward model, We also use the $\mathrm{s}\left(\hat{\mathrm{t}_1}, \mathrm{t}^*\right)$ as our Dimension Similarity Reward (DS Reward) which denotes dimension similarity of ground-truth and the first answer from top k predict list to enhance the effect of the first answer.We take the sum of NDCG and DS as the total reward function.

Methods	**Kaggle**		**PANDORA**	
	Macro-F1 (Binary)	F1-score (Multi)	Macro-F1 (Binary)	F1-score (Multi)
SVM	60.21	22.39	53.15	18.55
BERTconcat	60.61	23.77	53.91	10.03
BERTmean	66.04	26.38	56.52	17.56
D-DGCN	71.35	30.32	61.40	29.69
ChatGPT	63.89	34.55	60.08	29.87
Qwen-plus	76.43	38.63	61.34	25.08
PsyCoT	65.22	31.22	60.81	29.81
TAE	72.07	32.09	63.05	30.22
ETM	77.79	32.55	65.77	30.09
PerDet-R1 (our)	**80.57**	**41.34**	**66.10**	**35.08**

Table 1: Performance comparison on Kaggle and Pandora datasets. The average Macro-F1 for the 4-dimensional binary classification task and F1-Score for 16-types multiclass classification task.

Experiments

Datasets

Following previous studies (Yang et al. 2023a; Hu et al. 2024; Bi et al. 2025), we conducted experiments on the most commonly used personality detection datasets, including Kaggle[1] and PANDORA[2]. These datasets cover mainstream social platforms and forums. The Kaggle dataset was collected through the PersonalityCafe[3] forum, which is focused on personality type discussion. It comprises a total of 8675 users' data, each consisting of the 50 most recent posts. PANDORA was collected from Reddit[4], with dozens to hundreds of social media posts for each of 9067 users. As in previous work (Yang et al. 2023a), we remove words that match any personality type from the posts to avoid label leakage.

[1]https://www.kaggle.com/datasets/datasnaek/mbti-type
[2]https://psy.takelab.fer.hr/datasets/all/
[3]http://personalitycafe.com/forum
[4]https://www.reddit.com/

Performance is evaluated using the average Macro-F1 metric, which is the same as in previous work. We take the first answer in the top 3 MBTI answer list as the predicted value to calculate the average Macro-F1 for the 4-dimensional binary classification task, F1-Score for the 16 MBTI types multiclass classification task, and NDCG@k metric for the ranking task. We opted for Macro-F1 to ensure fairer comparisons with prior research findings. Furthermore, since our approach treats the four MBTI dimensions as a unified label, we employ F1-Score to comprehensively evaluate performance in multi-class classification tasks. Meanwhile, applying NDCG not only effectively reflects the training outcomes of stage 2 but also demonstrates whether the ranking task successfully ranks answers closest to the ground truth in the top position.

Baselines

To evaluate the proposed models intensively, we employ the following mainstream models as our baseline.

SVM (Cui and Qi 2017) and XGBoost (Tadesse et al. 2018): These methods concatenate all the posts of a user into a document first, and then employ SVM or XGBoost for classification based on features extracted using bag-of-words word models. They represent the advanced performance of traditional machine learning algorithms in personality detection tasks.

BERT-concat (Jiang, Zhang, and Choi 2019): This method concatenates a user's posts into an extended document and then utilizes BERT to encode this composite text for user representation.

BERT-mean (Keh and Cheng 2019): It uses BERT to encode each post individually, applies average pooling to get user feature embedding, and then maps the features to personality labels.

D-DGCN (Yang et al. 2023a): D-DGCN employs a dynamic graph to encode each post via domain-adapted BERT, then employs DGCN to encode the graph and obtains a user representation. And it's the SOTA method (Kaggle dataset) in a BERT-based model.

PsyCoT (Yang et al. 2023b): This method prompts the LLM to engage in reasonable personality reasoning. Introducing items from the questionnaire as a set of rigorous CoT. It guides the LLM to evaluate each item based on the author's text and then infer the final personality trait.

ChatGPT: We applied the 'gpt-3.5-turbo-0301' ver-sion of ChatGPT, and set the temperature to 0, making the outputs mostly deterministic for the identical inputs.

Qwen-plus (Yang et al. 2025): We applied the Qwen3 flagship model, and set the temperature to 1, making the outputs mostly deterministic for the identical inputs.

TAE (Hu et al. 2024): This method distills the useful knowledge from the LLM to address the data scarcity issue faced by small models in personality detection. It utilizes LLM to generate posts analysis from semantic, sentiment, and linguistic aspects, enhancing data and personality label representations.

ETM (Bi et al. 2025): This method leverages the text embedding and text generation capabilities of LLM to address the issues of poor user vector representation and the insufficient relationship between user vectors and personality labels in small model-based personality detection.

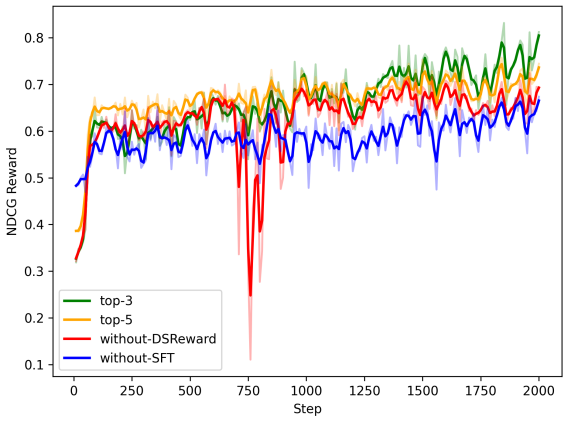

Figure 4: NDCG Reward (smoothed with 0.8 sigma) comparison during RL training stage on Kaggle Dataset.

Implementation Details

For the traditional deep learning method, we use PyTorch and use Adam Optimizer (Kingma and Ba 2015), the same as in the previous study. For LLM, we set up a batch size of 256 in the SFT stage and 128 in the RL stage with several generations of 16. The temperature parameter (τ) is maintained at 1. The fine-tuning process was conducted under the LoRA framework (Hu et al. 2021). We selected a rank of 32 and conducted the SFT throughout 3 epochs and RL over 2000 steps. The top k parameter is set to 3, and the dimension weight coefficient parameter ϵ is set to 0.1. Each post is limited to 128 tokens. We set the maximum number of posts per user to 50. We replace words that match any personality label with the <MASK> to avoid information leaks (Yang et al. 2023a). We use `Qwen-plus` for distilling reasoning trace and `Qwen2.5-7B-Instruct` as the base model that needs to be fine-tuned.

Model	Strategy	Kaggle		
		Binary	Multi-class	NDCG@3
Llama3-8B	w/o SFT, GRPO	35.29	8.80	51.15
	w/o GRPO	59.65	13.77	62.69
	SFT+GRPO	72.57	37.88	71.64
Qwen2.5-7B	w/o SFT, GRPO	35.67	6.21	49.46
	w/o SFT	68.11	17.21	67.58
	w/o GRPO	65.03	15.82	64.09
	classification	63.07	33.67	–
	w/o DS reward	70.06	28.64	69.38
	w/o NDCG reward	68.90	27.28	66.92
	Top-5	76.44	43.58	74.91
	SFT+GRPO	80.57	41.34	74.31

Table 2: Results of ablation study over different base models and training strategies, we mainly carry out ablation experiments on Qwen model and Kaggle dataset

Overall Results

Table 1 demonstrates that PerDet-R1 outperforms all baselines in binary Macro-F1 on benchmark datasets. Specifically, on the Kaggle and PANDORA datasets, compared with the currently best-performing ETM, our method achieved improvements of 2.78% and 0.83% in Macro-F1 scores, as well as 8.79% and 4.99% in F1-score scores. Although our primary objective is to rank complete MBTI types in the top position, this design simultaneously enhances the classification capability. It is worth noting that when evaluating personality trait detection as a multi-classification task, PerDet-R1 achieves a significantly higher F1-score compared to other methods. This advantage stems from modeling all four dimensions of MBTI as an integrated element during both two-stage training stages, along with full comparisons of differences between distinct categories. Overall, our method achieves outstanding performance due to two key enhancements. Firstly, during the SFT stage, it memorizes the strategy and thought process of analysis and comparison through rejection sampling and off-policy distillation, thereby acquiring reasoning capabilities. In the RL stage, it autonomously explores effective paths via the GRPO method, enhancing the model's generalization ability. Secondly, the NDCG-based reward function ingeniously integrates the two stages, modeling personality detection as a ranking task, thereby overcoming the fuzzy boundary issue caused by existing methods that simplistically separate the four dimensions into multiple binary classification tasks.

Through comparative analysis of two datasets, we observed that existing methods generally demonstrate lower accuracy on the Pandora dataset compared to the Kaggle dataset. This discrepancy likely stems from the fact that the Kaggle dataset originates from the MBTI forum, where users frequently employ abbreviations like "NS", "ExFJ", and "xNTP" during discussions. While these terms are not exhaustive answers, they effectively provide strong contextual cues that enhance the data-driven method.

Ablation Study

Table 2 presents the ablation results on the Kaggle dataset to facilitate a comprehensive investigation. PerDet-R1 integrates a two-stage training process (SFT followed by RL), a specially adapted reward function to enhance LLM reasoning capabilities, and a ranking-based inference strategy that replaces classification. PerDet-R1 achieves comprehensive performance improvement through integrated strategies, where removing any component causes overall performance degradation.

Training Strategy. For training stages, after removing GRPO, the NDCG score decreased by 8.95% for Llama3-8B and 10.22% for Qwen2.5-7B. When removing both SFT and GRPO, the NDCG score dropped substantially by 20.49% and 24.85%, respectively, on both models. As shown by the blue line in Figure 4, without the first-stage SFT, the performance improvement during later RL training becomes weak. This demonstrates the necessity of our two-stage training. Meanwhile, we conducted experiments using classification with SFT+GRPO instead of ranking on Qwen2.5-7B, which resulted in decreases of 17.50% in binary F1 and 7.67% in multi-class Macro-F1, confirming the significance of the ranking method.

Reward model. For the reward model in GRPO, we found that removing the dimension similarity reward let to NDCG reward decreased by 4.93% and 7.39%, respectively. Furthermore, as observed from the red line in Figure 4, the NDCG reward collapses abruptly during training without the dimension similarity reward. Similarly, the dimension similarity reward collapses when the NDCG reward is removed. This phenomenon confirms that both reward functions are essential for stable performance improvement.

Top-k. We used a variant number top k with k = 5 to verify whether a larger answer space facilitates exploration. The results show that while binary F1 decreased by 4.13%, multi-class Macro-F1 and NDCG increased by 2.24% and 0.60%, respectively.

These results demonstrate that our method leverages two-stage training and specially adapted reward modeling to enhance LLMs' reasoning capabilities in personality detection tasks significantly.

Conclusion

In this paper, we propose a list-wise ranking fine-tuning framework for personality detection that establishes a paradigm shift from classification on dimensions to personality trait ranking. Our method significantly enhanced LLM reasoning capabilities, particularly valuable when models inherently struggle with the task, by fundamentally reconceptualizing personality detection as a comparative ranking problem rather than disjoint binary classifications. This paradigm shift overcomes the critical limitation of existing approaches, which artificially fragment MBTI dimensions and disrupt their intrinsic interrelationships. Through a two-stage architecture featuring: 1) SFT with off-policy distillation for reasoning trace generation and rejection sampling for quality control, followed by 2) GRPO optimization with an NDCG-based reward function for generalization enhancement, our approach comprehensively models personality trait inter-class differences. Experimental results across benchmark datasets demonstrate state-of-the-art performance, validating the framework's efficacy. In future work, we will explore how to achieve higher-performance personality trait detection by utilizing multi-modal data such as profile avatars and facial expressions.

Acknowledgments

Author Y. Zhu gratefully acknowledge the financial support from the National Key R&D Program of China (Grant No. 2023YFC2506600, 2023YFC2506601).

References

Bi, W.; Kou, F.; Shi, L.; Li, Y.; Li, H.; Chen, J.; and Xu, M. 2025. Leveraging the Dual Capabilities of LLM: LLM-Enhanced Text Mapping Model for Personality Detection. In *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence*, volume 39, 23487–23495. Washington, DC: AAAI Press.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; and et al. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, volume 33, 1877–1901.

Chung, H. W.; Hou, L.; Longpre, S.; Zoph, B.; Tay, Y.; Fedus, W.; Li, Y.; Wang, X.; Dehghani, M.; Brahma, S.; Webson, A.; Gu, S. S.; Dai, Z.; Suzgun, M.; Chen, X.; Chowdhery, A.; Castro-Ros, A.; Pellat, M.; Robinson, K.; Valter, D.; Narang, S.; Mishra, G.; Yu, A.; Zhao, V.; Huang, Y.; Dai, A.; Yu, H.; Petrov, S.; Chi, E. H.; Dean, J.; Devlin, J.; Roberts, A.; Zhou, D.; Le, Q. V.; and Wei, J. 2022. Scaling Instruction-Finetuned Language Models. *arXiv preprint*, arXiv:2210.11416v5.

Cui, B.; and Qi, C. 2017. Survey Analysis of Machine Learning Methods for Natural Language Processing for MBTI Personality Type Prediction. Technical report cs229, Stanford University, Stanford, CA.

Dong, H.; Xiong, W.; Pang, B.; Wang, H.; Zhao, H.; Zhou, Y.; Jiang, N.; Sahoo, D.; Xiong, C.; and Zhang, T. 2024. RLHF Workflow: From Reward Modeling to Online RLHF. arXiv:2405.07863.

Gao, L.; Schulman, J.; and Hilton, J. 2023. Scaling Laws for Reward Model Overoptimization. In *Proceedings of the 40th International Conference on Machine Learning*, 10835–10866. Honolulu, HI: PMLR.

Guo, D.; Yang, D.; Zhang, H.; Song, J.; Zhang, R.; Xu, R.; Zhu, Q.; Ma, S.; Wang, P.; and Bi, X. 2025. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. *arXiv preprint*, arXiv:2501.12948v1.

Heston, T. F.; and Gillette, J. 2025. Do Large Language Models Have a Personality? A Psychometric Evaluation with Implications for Clinical Medicine and Mental Health AI. *medRxiv*.

Hsieh, C.-Y.; Li, C.-L.; Yeh, C.-K.; Nakhost, H.; Fujii, Y.; Ratner, A.; Krishna, R.; Lee, C.-Y.; and Pfister, T. 2023. Distilling Step-by-Step! Outperforming Larger Language Models with Less Training Data and Smaller Model Sizes. In *Findings of the Association for Computational Linguistics: ACL 2023*, 8003–8017. Toronto, Canada: Association for Computational Linguistics.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*.

Hu, L.; He, H.; Wang, D.; Zhao, Z.; Shao, Y.; and Nie, L. 2024. LLM vs Small Model? Large Language Model Based Text Augmentation Enhanced Personality Detection Model. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence*, volume 38, 18234–18242. Washington, DC: AAAI Press.

Hüllermeier, E.; Fürnkranz, J.; Cheng, W.; and Brinker, K. 2008. Label ranking by learning pairwise preferences. *Artificial Intelligence*, 172(16): 1897–1916.

Islam, M. A.; and Faruk, A. S. 2025. Prompt-Based LLMs for Position Bias-Aware Reranking in Personalized Recommendations. arXiv:2505.04948.

Jiang, H.; Zhang, X.; and Choi, J. D. 2020. Automatic Text-Based Personality Recognition on Monologues and Multiparty Dialogues Using Attentive Networks and Contextual Embeddings (Student Abstract). In *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence*, volume 34, 13821–13822. Palo Alto, CA: AAAI Press.

Keh, S. S.; and Cheng, I.-T. 2019. Myers-Briggs Personality Classification and Personality-Specific Language Generation Using Pre-trained Language Models. *arXiv preprint*, arXiv:1907.06333v1.

Kingma, D. P.; and Ba, J. 2015. Adam: A Method for Stochastic Optimization. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR 2015)*. San Diego, CA. ArXiv preprint arXiv:1412.6980.

Le, M.; Luu, T. N.; The, A. N.; Le, T. T.; Nguyen, T.; Nguyen, T. T.; Van, L. N.; and Nguyen, T. H. 2024. Adaptive Prompting for Continual Relation Extraction: A Within-Task Variance Perspective.

Lee, H.; Phatale, S.; Mansoor, H.; Mesnard, T.; Ferret, J.; Lu, K. R.; Bishop, C.; Hall, E.; Carbune, V.; Rastogi, A.; and Prakash, S. 2024. RLAIF vs. RLHF: Scaling Reinforcement Learning from Human Feedback with AI Feedback. In *Proceedings of the 41st International Conference on Machine Learning*, 26874–26901. Vienna, Austria: Proceedings of Machine Learning Research.

Liu, Z.; Guo, X.; Lou, F.; Zeng, L.; Niu, J.; Wang, Z.; Xu, J.; Cai, W.; Yang, Z.; Zhao, X.; Li, C.; Xu, S.; Chen, D.; Chen, Y.; Bai, Z.; and Zhang, L. 2025. Fin-R1: A Large Language Model for Financial Reasoning through Reinforcement Learning. *arXiv preprint*, arXiv:2503.16252v2.

Lynn, V.; Balasubramanian, N.; and Schwartz, H. A. 2020. Hierarchical Modeling for User Personality Prediction: The Role of Message-Level Attention. In Jurafsky, D.; Chai, J.; Schluter, N.; and Tetreault, J., eds., *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5306–5316. Online: Association for Computational Linguistics.

Mu, T.; Helyar, A.; Heidecke, J.; Achiam, J.; Vallone, A.; Kivlichan, I.; Lin, M.; Beutel, A.; Schulman, J.; and Weng, L. 2024. Rule Based Rewards for Language Model Safety. In Globerson, A.; Mackey, L.; Belgrave, D.; Fan, A.; Paquet, U.; Tomczak, J.; and Zhang, C., eds., *Advances in Neural Information Processing Systems*, volume 37, 108877–108901. Curran Associates, Inc.

Myers, I. B.; et al. 1962. *The myers-briggs type indicator*, volume 34. Consulting Psychologists Press Palo Alto, CA.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; Schulman, J.; Hilton, J.; Kelton, F.; Miller, L.; Simens, M.; Askell, A.; Welinder, P.; Christiano, P. F.; Leike, J.; and Lowe, R. 2022. Training Language Models to Follow Instructions with Human Feedback. In Koyejo, S.; Mohamed, S.; Agarwal, A.; Belgrave, D.; Cho, K.; and Oh, A., eds., *Advances in Neural Information Processing Systems*, volume 35, 27730–27744.

Papicchio, S.; Rossi, S.; Cagliero, L.; and Papotti, P. 2025.

Think2SQL: Reinforce LLM Reasoning Capabilities for Text2SQL. *arXiv preprint*, arXiv:2504.15077v2.

Rushton, S.; Mariano, J. M.; and Wallace, T. L. 2012. Program selection among pre-service teachers: MBTI profiles within a college of education. *Creative Education*, 3(1): 16–23.

Serapio-García, G.; Safdari, M.; Crepy, C.; Sun, L.; Fitz, S.; Romero, P.; Abdulhai, M.; Faust, A.; and Matarić, M. 2025. Personality Traits in Large Language Models. arXiv:2307.00184.

Shao, Z.; Wang, P.; Zhu, Q.; Xu, R.; Song, J.; Bi, X.; Zhang, H.; Zhang, M.; Li, Y. K.; Wu, Y.; and Guo, D. 2024. DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models. *arXiv preprint*, arXiv:2402.03300v3.

Shen, L.; Long, Y.; Cai, X.; Chen, G.; Razzak, I.; and Jameel, S. 2025. Less but Better: Parameter-Efficient Fine-Tuning of Large Language Models for Personality Detection. arXiv:2504.05411.

Stajner, S.; and Yenikent, S. 2021. 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*, 3580–3589. Online: Association for Computational Linguistics.

Tadesse, M. M.; Lin, H.; Xu, B.; and Yang, L. 2018. Personality predictions based on user behavior on the Facebook social media platform. *IEEE Access*, 6: 61959–61969.

Tandera, T.; Hendro; Suhartono, D.; Wongso, R.; and Prasetio, Y. L. 2017. Personality Prediction System from Facebook Users. *Procedia Computer Science*, 116: 604–611.

Wang, R.; and Sun, K. 2024. Continuous Output Personality Detection Models via Mixed Strategy Training. arXiv:2406.16223.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q. V.; Zhou, D.; et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35: 24824–24837.

Xie, T.; Gao, Z.; Ren, Q.; Luo, H.; Hong, Y.; Dai, B.; Zhou, J.; Qiu, K.; Wu, Z.; and Luo, C. 2025. Logic-RL: Unleashing LLM Reasoning with Rule-Based Reinforcement Learning. *arXiv preprint*, arXiv:2502.14768v1.

Xiong, W.; Dong, H.; Ye, C.; Wang, Z.; Zhong, H.; Ji, H.; Jiang, N.; and Zhang, T. ???? Iterative preference learning from human feedback: Bridging theory and practice for RLHF under KL-constraint. In *ICLR 2024 Workshop on Mathematical and Empirical Understanding of Foundation Models*.

Xue, D.; Wu, L.; Hong, Z.; Guo, S.; Gao, L.; Wu, Z.; Zhong, X.; and Sun, J. 2018. Deep learning-based personality recognition from text posts of online social networks. *Applied Intelligence*.

Yang, A.; Li, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Gao, C.; Huang, C.; Lv, C.; Zheng, C.; Liu, D.; Zhou, F.; Huang, F.; and Hu, F. 2025. Qwen3 Technical Report. *arXiv preprint*, arXiv:2505.09388v1.

Yang, F.; Yang, T.; Quan, X.; and Su, Q. 2021. Learning to Answer Psychological Questionnaire for Personality Detection. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, 1131–1142. Punta Cana, Dominican Republic: Association for Computational Linguistics.

Yang, T.; Deng, J.; Quan, X.; and Wang, Q. 2023a. Orders Are Unwanted: Dynamic Deep Graph Convolutional Network for Personality Detection. In *Proceedings of the 37th AAAI Conference on Artificial Intelligence*, volume 37, 13896–13904. Washington, DC: AAAI Press.

Yang, T.; Shi, T.; Wan, F.; Quan, X.; Wang, Q.; Wu, B.; and Wu, J. 2023b. PsyCoT: Psychological Questionnaire as Powerful Chain-of-Thought for Personality Detection. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 3305–3320. Singapore: Association for Computational Linguistics.

Zhu, Y.; Xia, Y.; Li, M.; Zhang, T.; and Wu, B. 2024. Data Augmented Graph Neural Networks for Personality Detection. In *Proceedings of the 38th AAAI Conference on Artificial Intelligence*, volume 38, 664–672. Washington, DC: AAAI Press.

Ziegler, D. M.; Stiennon, N.; Wu, J.; Brown, T. B.; Radford, A.; Amodei, D.; Christiano, P.; and Irving, G. 2020. Fine-Tuning Language Models from Human Preferences. *arXiv preprint*, arXiv:1909.08593v2.