Can Large Language Models Understand You Better? An MBTI Personality Detection Dataset Aligned with Population Traits

Bohan Li[1], Jiannan Guan[1], Longxu Dou[2], Yunlong Feng[1], Dingzirui Wang[1], Yang Xu[1], Enbo Wang[1], Qiguang Chen[1], Bichen Wang[1], Xiao Xu[1], Yimeng Zhang[2], Libo Qin[3], Yanyan Zhao[1], Qingfu Zhu[1], Wanxiang Che[1]*

[1] Harbin Institute of Technology [2] Individual Researcher [3] Central South University

{bhli,jnguan}@ir.hit.edu.cn, car@ir.hit.edu.cn

Abstract

The Myers-Briggs Type Indicator (MBTI) is one of the most influential personality theories reflecting individual differences in thinking, feeling, and behaving. MBTI personality detection has garnered considerable research interest and has evolved significantly over the years. However, this task tends to be overly optimistic, as it currently does not align well with the natural distribution of population personality traits. Specifically, (1) the self-reported labels in existing datasets result in incorrect labeling issues, and (2) the hard labels fail to capture the full range of population personality distributions. In this paper, we optimize the task by constructing MBTIBENCH, the first manually annotated high-quality MBTI personality detection dataset with soft labels, under the guidance of psychologists. As for the first challenge, MBTIBENCH effectively solves the incorrect labeling issues, which account for 29.58% of the data. As for the second challenge, we estimate soft labels by deriving the polarity tendency of samples. The obtained soft labels confirm that there are more people with non-extreme personality traits. Experimental results not only highlight the polarized predictions and biases in LLMs as key directions for future research, but also confirm that soft labels can provide more benefits to other psychological tasks than hard labels.[1]

1 Introduction

Personality, a key psychological concept, refers to individual differences in thinking, feeling, and behavior (Corr and Matthews, 2009). Among personality models, the Myers-Briggs Type Indicator (MBTI) is one of the most recognized non-clinical frameworks, with broad applications in areas like stance detection (Hosseinia et al., 2021). Recently, automatically detecting a person's MBTI type from their written content (e.g., tweets and blogs) (Plank and Hovy, 2015; Gjurković et al., 2021) has become a growing area of research with both academic significance (Stajner and Yenikent, 2020; Khan et al., 2005) and practical applications (Bagby et al., 2016) .

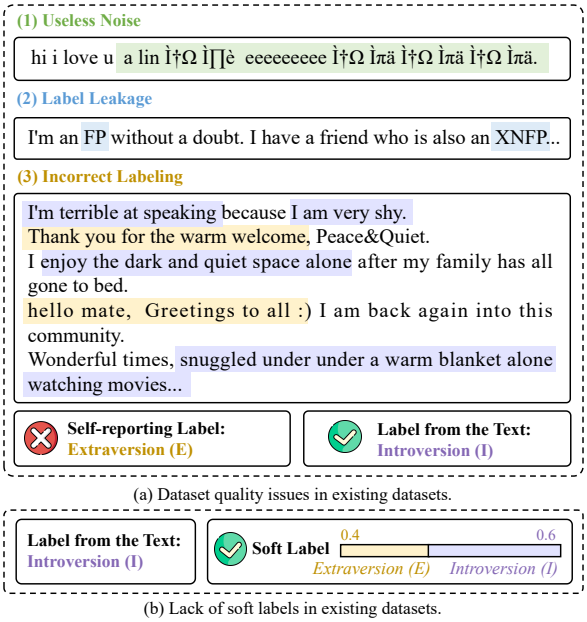


Figure 1: Our MBTIBENCH focuses on the above two limitations in existing MBTI personality detection datasets: data quality issues and the lack of soft labels.

However, MBTI personality detection tends to be overly optimistic, as it currently does not align well with the natural distribution of personality traits within the population. (1) Given the widespread use of MBTI, existing datasets are often derived from social media posts, with labels typically sourced from users' self-reported MBTI types (Gjurković et al., 2021). Assessments conducted by non-professional psychological institutions, along with inaccurate self-perception, can lead to **a mismatch between the self-reported personality traits and the actual linguistic patterns in the text** (McDonald, 2008; Müller and

*Corresponding author.

[1]The code and data are available at https://github.com/Personality-NLP/MbtiBench.

Proceedings of the 31st International Conference on Computational Linguistics, pages 5071–5081
January 19–24, 2025.

Attitudes	Perceiving Functions	Judging Functions	Lifestyle Preferences
Interact with the world	*Gather information*	*Make decisions*	*Organize their lives*
Extraversion (E)	**Sensing (S)**	**Thinking (T)**	**Judgment (J)**
Energized by interaction	*Practical and realistic*	*Logical and analytical*	*Organized and structured*
Social and talkative	*Detail-oriented*	*Fact-based*	*Planning-oriented*
Introversion (I)	**iNtuition (N)**	**Feeling (F)**	**Perception (P)**
Energized by solitude	*Imaginative and innovative*	*Compassionate and empathetic*	*Flexible and spontaneous*
Reflective and thoughtful	*Future-focused*	*Values-driven*	*Adaptable*

Figure 2: Definitions of the four dimensions in MBTI personality theory.

Moshagen, 2019; Paulhus and Vazire, 2007). For example, a user who self-identifies as an *Extraversion* type might exhibit more *Introversion* traits in their posts (Figure 1 (a)). (2) From a psychological standpoint, **personality is not binary but rather complex, nuanced, and highly individualized** (Wu et al., 2022). Most people don't display extreme personality traits; instead, they tend to fall somewhere in the middle (Tzeng et al., 1989). However, existing datasets use only binary MBTI labels, missing the full spectrum of personality traits in most people (Harvey and Murry, 1994; Wu et al., 2022) (Figure 1 (b)).

In this paper, we make a step toward solving the challenge. We optimize the task by constructing MBTIBENCH, the first MBTI personality detection dataset aligned with population traits. (1) To solve the data quality issues related to self-reported labels, we propose the first data filtering guidelines for MBTI personality detection and apply them on existing datasets to ensoure data quality. We manually re-annotate the cleaned datasets under the guidance of psychological experts, aligning each post with correct labels that best descrtibe the personality polarity. (2) To capture the full range of population personality traits, we estimate soft labels for MBTI personality detection by deriving the polarity tendencies of samples. The obtained soft labels confirm the above opinion that there are more people with non-extreme personality traits.

We analyze our MBTIBENCH in detail from multiple perspectives to explore the influence of soft labels compared to hard labels. By integrating personality traits into stress identification, we demonstrate that the use of soft labels significantly enhances performance on personality-related tasks. We further evaluate six large language models (LLMs) and four prompting methods on MBTIBENCH, and highlight the polarized predictions and biases as key directions.

Our contributions are as follows:

- We are the first to challenge the over-optimism in MBTI personality detection from the nature of population personality traits.
- We optimize MBTI personality detection by creating MBTIBENCH, the first soft-labeled MBTI dataset, manually annotated under the guidance of psychologists.
- We highlight the polarized predictions and biases in LLMs as future research directions. Our experimental results also confirm that soft labels can provide more benefits to other psychological tasks than hard labels.

2 MBTI Personality Detection

2.1 Existing Datasets Overview

There are three widely used datasets in personality detection based on MBTI, including Twitter (Plank and Hovy, 2015), Kaggle[2],PANDORA (Gjurković et al., 2021). These three datasets are collected from social media posts. Each sample is a set of posts from the corresponding user, and the labels are obtained through users self-reporting their posts (see Appendix C.1). However, as shown in Figure 1, there are issues of label leakage and irrelevant noise that significantly compromise the quality and factual accuracy of the data itself. Moreover, users' lack of clarity in self-reporting often leads to mismatches between self-reported labels and their posts (McDonald, 2008). To address the above issues, we clean and re-annotate existing datasets to construct our MBTIBENCH.

2.2 Problem Overivew

We introduce the definitions of the four dimensions in MBTI personality theory in Figure 2.

The task of MBTI personality detection, which involves automatically inferring a person's MBTI type from their textual content, has attracted significant interest from researchers due to its broad range of potential applications (Khan et al., 2005;

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

Bagby et al., 2016). Given the growing interest in understanding personality through text, MBTI personality detection, that automatically inferring an individual's MBTI type from their written content, has become a prominent area of research, with a broad range of practical applications (Khan et al., 2005; Bagby et al., 2016). As for hard labels, existing MBTI datasets lay out a binary classification based on four distinct dimensions independently (*E/I*, *S/N*, *T/F*, *J/P*), and draw the typology of the person according to the combination of those four values (e.g. *ESTJ*).

In this paper, we construct a new dataset called MBTIEVAL with soft labels. Soft labels are continual representations of polarity tendency mapped in $[0, 1]$, and we define the four dimensions of soft labels as the degree of *E*, *S*, *T*, and *J* polarity, respectively. For example, 40% *Extraverslon* simultaneously represents 60% *Introversion*.

3 Dataset Construction

In this paper, we re-annotate three existing datasets to solve the dataset quality issues. We introduce the construction details below.

3.1 Design Principles

To ensure our annotation quality, we respectively establish data filtering guidelines for useless noise and label leakage issues, and data annotation guidelines for incorrect labeling issues.[3]

3.1.1 Data Filtering Guidelines

We are the first to summarize data filtering guidelines for label leakage and useless noise issues in MBTI personality detection.

We divide the label leakage errors into three categories: (1) Direct Personality Leakage involves direct references to personality types through specific letters ("ENFP") or complete personality type words ("Introverted"). (2) Personality Trait Leakage involves specific MBTI trait descriptors providing enough context to infer certain personality types. For example, *Te* indicates *extroverted Thinking*. (3) Cross-Theory Trait Leakage involves traits from other personality theories like the Big Five (Furnham, 1996), confirming MBTI personality characteristics.

We divide the useless noise issues into three categories: (1) Information-insufficient Samples indicate samples that are too short (less than 100 words) to contain valuable information for inferring personality types. (2) Garbled Text indicates blocks of text that appears as random, unintelligible characters and symbols. (3) Link and Media References indicates hyperlinks or media file names, which are useless for personality detection.

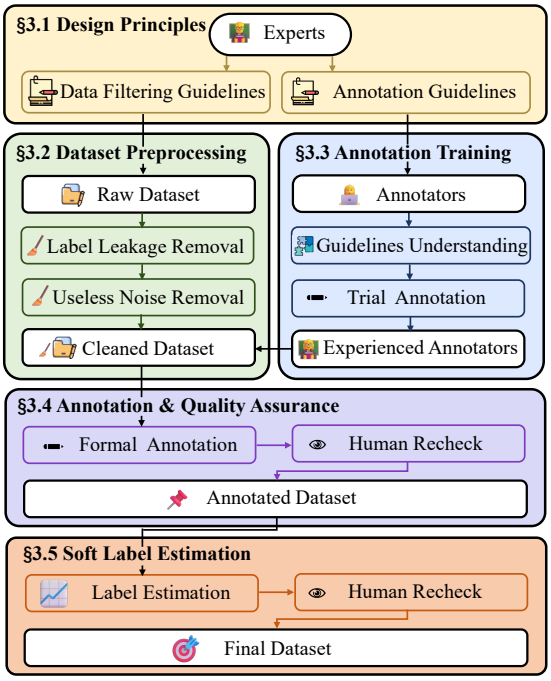


Figure 3: MBTIBENCH construction worklow.

3.1.2 Data Annotation Guidelines

To address incorrect labeling issues and accurately label personality types,we refer to the methodology outlined in Stajner and Yenikent (2021) to construct our annotation guidelines. Psychology PhD students participate in the formulation of these guidelines. We discuss the personality traits for each dimension based on the dataset, analyzing and adjusting our guidelines through trial annotations. Finally, we annotate the trial samples, which serve as expert guidelines.

3.2 Dataset Preprocessing

We reconstruct the test sets of the three most commonly used personality detection datasets, including Twitter, Kaggle, and PANDORA. Following Stajner and Yenikent (2021), we select six samples for each type, totaling 286 samples across the three datasets. [4] To ensure that the samples provide useful data for personality detection, we filter the dataset. Specifically, we manually remove sen-

[3]We provide guideline examples in Appendix D.1

[4]There are only four samples for the ESFJ type in the test set of PANDORA.

tences involving label leakage and useless noise from the samples, using the annotation guidelines from Section 3.1, resulting in a cleaned dataset awaiting annotation.

3.3 Annotation Training

3.3.1 Annotation Guidelines Understanding

We employ three experienced annotators who hold either a Bachelor's or Master's degree in English and can use English fluently with extensive annotation experience. We conduct training sessions for them under the guidance of psychology experts, using expert-annotated examples to explain the definition and judgment signals for each MBTI dimension. Annotators read through the entire instance and make independent judgments for each dimension. We refer to the 4-point Likert scale (Likert, 1932) and ask the annotators to assign, for each instance and each MBTI dimension separately, two polar intensity labels. For example, in the *E/I* dimension, they could assign *E*-, *E*+, *I*-, or *I*+ labels to reflect the degree of classification signals or their confidence level. In each pilot round, the dataset they annotate consists of 16 instances, which are not used in the final round to ensure the integrity of the final data.

3.3.2 Trial Annotation

We distribute trial annotation samples to the annotators, requiring them to annotate the same samples independently. This is to assess their consistency and understanding of the guidelines, as well as the effectiveness of the guidelines themselves. The trial annotation samples obtained during the pilot rounds show how the proposed guidelines are used in practice and highlight the most challenging aspects of the annotation process (Shi et al., 2023; Chen et al., 2023).

After the trial annotation, we determine the final result for each label by voting and measuring annotation quality using annotation accuracy and Fleiss' Kappa (Fleiss, 1971). If the accuracy is below 0.8 or the Fleiss' Kappa does not exceed 0.45, we repeat the "Annotation Training" process, starting again from "Annotation Guidelines Understanding" to ensure the effectiveness of the trial annotation.[5] We conduct three rounds of trial annotation, during which the accuracy and Kappa for the four dimensions steadily increase.

3.4 Annotation and Quality Assurance

3.4.1 Formal Annotation

Once the trial annotation is completed, the annotators become fully acquainted with the annotation guidelines and procedures. The annotators are provided with the cleaned dataset and are expected to annotate independently. We pay them at a rate that is no less than the local average, and they are entitled to take adequate breaks to mitigate the effects of fatigue. Annotators are provided with the cleaned dataset and annotate independently. They are compensated at a rate no less than the local average and are given adequate breaks to reduce fatigue.

3.4.2 Human Recheck

After the formal annotation, we conduct a human recheck of each annotated instance to ensure the annotation quality. If an instance is found to be unreasonably annotated, we summarize the issues and communicate them with the respective annotator. These problematic samples are then mixed with other normal samples and re-annotated by the annotators to prevent direct suggestions or interference. The re-annotated instances account for less than 15% of the total. The final Fleiss' Kappa scores[6] are not only satisfactory for personality detection (Jiang et al., 2019) but also suitable for soft label estimation, which demonstrates our annotation quality.

3.5 Soft Label Generation

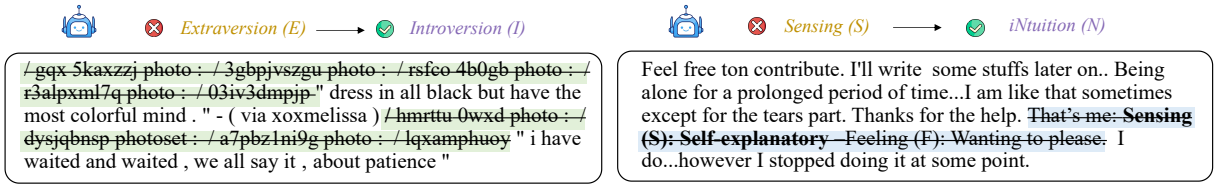
3.5.1 Label Estimation

Inspired by the Dawid-Skene model (Dawid and Skene, 1979), we adopt the EM algorithm (Dempster et al., 1977) to derive the polarity tendency of samples. The algorithm effectively fits annotators of varying quality and comprehensively produces realistic rankings for each sample (Passonneau and Carpenter, 2013). We estimate a more objective soft label by processing the hard labels provided by the three annotators, thereby accurately reflecting the comprehensive inclination of the data across various dimensions. In this paper, we use the *E/I* dimension as an example to introduce the algorithm for estimating soft labels (Algorithm 1).

1.Initial True Label Calculation: First, we take the annotated data as the initial input and calculate

[5] A Fleiss' Kappa value above 0.4 is satisfactory for subjective tasks (Jiang et al., 2019).

[6] 0.4779 for *E/I*, 0.4686 for *S/N*, 0.5517 for *T/F*, 0.4622 for *J/P*.



(a) After removing useless noises, the model can classify Introversion (I) traits more accurately.

(b) After removing label leakage, the model struggles to distinguish Sensing (S) traits.

Figure 4: After solving the data quality problem, the performance of the model is affected.

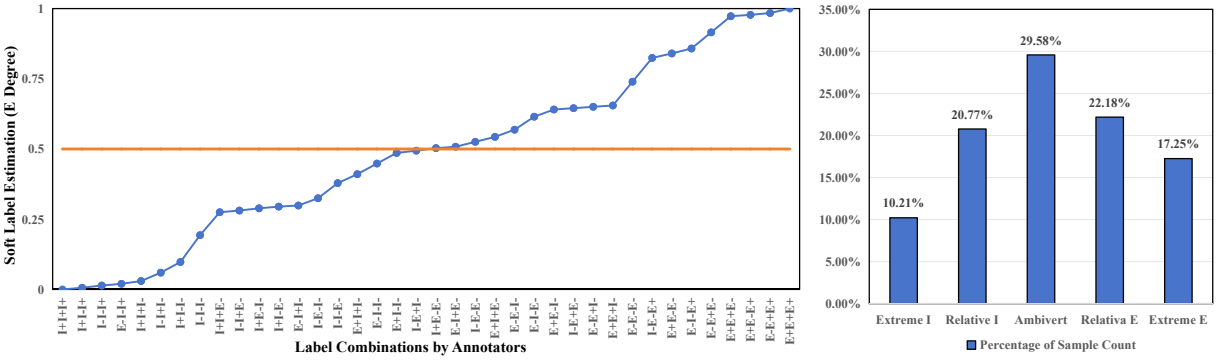


Figure 5: The mapping from hard labels to soft labels of the *E/I* dimension in MBTIBENCH. In (a), we use an EM algorithm to convert annotator label combinations (x-axis) into soft labels ranging from 0 to 1 (y-axis). We then define Extreme I, Relative I, Ambivert, Relative E, and Extreme E based on these soft label values, as shown on the x-axis of (b). The y-axis in (b) shows the sample count for each category.

the median of the three annotators' labels as the initial true label.

2.Co-occurrence Matrix Calculation: Next, we compare each annotator's label with the initial true label to obtain a matrix $\mathbf{M} \in \mathcal{R}^{3\times4\times4}$ where m_{ijk} represents the joint frequency of category j and category k in i-th annotator labels.

3.Transition Matrix Formation: Using Bayes' theorem to merge E+ and E- into E, and store the result in a new matrix T:

$$T_{i0k} = \frac{(N_{i0k} \times P_{E+}) + (N_{i1k} \times P_{E-})}{P_E}.$$

P_{E+}, P_{E-}represent the proportions of labels classified as E+,E-,and P_E is the sum of them.

4.Iterative Calculation: Calculate the initial posterior probability of E from the transition matrix, then update the matrix and recompute until changes are within a predefined tolerance:

$$P(E \mid r_1, r_2, r_3) = \frac{\prod_{i=1}^{3} P_{r_i 0} \times P_E}{\sum_{j=0}^{1} \prod_{i=1}^{3} P_{r_i j} \times P_E},$$

where r_i represents the label labeled by the i-th annotator.

5.Tendency Calculation: Since the midpoint where posterior probability is 0.5, we accumulate and average values on both sides.

6. Normalization: Finally, we normalize the cumulative frequencies on both sides of the midpoint. Each combination is assigned a final ratio between 0 and 1, reflecting the label's tendency. We define this ratio as the "soft label" of the type *Extraversion*.

4 Dataset Quality Analysis

We analyze the quality of our datasets from three perspectives by addressing the limitations found in existing datasets (Liu et al., 2020).

4.1 Human Recheck

Under the guidance of psychology experts, we review the estimated soft labels to ensure they align with the degree of personality polarity reflected in texts. We find that the use of the EM algorithm in estimating soft labels effectively integrates the annotators' varying accuracy and labeling habits, producing accurate soft labels (Raykar et al., 2010a) (Table 3 in Appendix D.4).

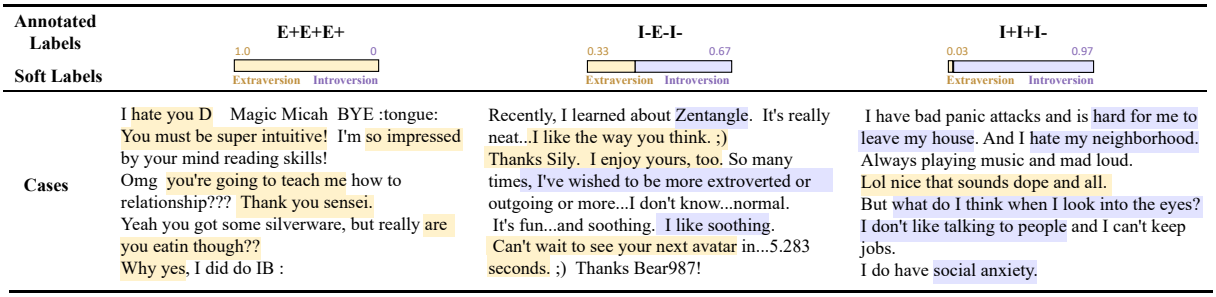


Figure 6: Three cases and their corresponding annotated labels and soft labels.

Dimensions	Kaggle	Twitter	PANDORA
Useless Noise	15.92%	43.13%	2.65%
Label Leakage	31.21%	0.62%	16.56%
Incorrect Labeling	29.17%	29.17%	29.79%

Table 1: Issues solved in three datasets.

4.2 Self-reported Labels Influence Model Performance

We analyze the useless noise issues, self-reported label leakage, and incorrect self-reported labels in Table 1 and Figure 4. The datasets exhibit varying degrees of quality issues before refinement that impact model performance. More dataset statics are in Appendix Table 4.

4.3 Soft Labels Align with Population Traits

We illustrate the estimated soft labels and the corresponding sample count distribution of *E/I*, *S/N*, *T/F*, and *J/P* in Figure 5 and Appendix Figure 11-13. Soft labels exhibit a smooth distribution from 0 to 1, and there are more non-extreme samples than extreme ones. This is consistent with the population distribution and demonstrates that our soft labels align with population traits better compared to hard labels (Harvey and Murry, 1994).

5 Experiment Setup

To evaluate model performance on MBTIBENCH and explore future directions, we conduct experiments across six backbone models and four prompting methods on MBTIBENCH (Liu et al., 2022). In this section, we provide an overview of the experiments.

5.1 Experimental Details

We employ several widely used and powerful LLMs as our backbones, including closed-source LLMs: GPT-4 (Achiam et al., 2023) (`gpt-4o-mini` and `gpt-4o`), and open-source LLMs: Qwen2 series (Yang et al., 2024) (`Qwen2-7B-Instruct` and `Qwen2-72B-Instruct`), Llama3.1 series (Dubey et al., 2024) (`Meta-Llama-3.1-8B-Instruct` and `Meta-Llama-3.1-70B-Instruct`). Qwen2 series have a context window of 32K tokens, while Llama 3.1 series and GPT-4o series both have context windows of 128K tokens, which are sufficient to process a large volume of user posts and perform Chain-of-Thought (CoT) (Wei et al., 2022) reasoning. Moreover, we employ the average soft labels of each dimension as the *baseline*, to evaluate model performance.

Following the methodologies outlined in Yang et al. (2023), we adopt four prompting approaches: (1) Zero-shot involves directly presenting the task description and requiring the model to directly complete the task. (2) Step-by-step (Kojima et al., 2022) employs the additional phrase *Let's think step by step* based on zero-shot in the inference prompt.[7] (3) Few-shot employs two additional examples in the inference prompt based on zero-shot, with each example containing a post and the corresponding complete personality label.[8] (4) PsyCoT (Yang et al., 2023) requires the model to answer MBTI scale questions and refer to the answers for final personality judgments. For detailed information on inference parameters and post length, please refer to Appendix F.

Following Raykar et al. (2010b) and Gjurković et al. (2021), we directly predict soft labels to evaluate the model's ability to make more detailed estimations of personality traits. Specifically, $t(\cdot)$ instructs the model to assess the degree of each personality trait and generate a score within the range of 1 to 9, for example, *[[7.25]]*.

We employ the Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) to measure the

[7] *Zero-shot* in Yang et al. (2023).
[8] We manually choose the two examples.

Backbones	Methods	E/I		S/N		T/F		J/P		Rank
		S-RMSE	S-MAE	S-RMSE	S-MAE	S-RMSE	S-MAE	S-RMSE	S-MAE	
Baseline	-	2.66	2.29	2.70	2.31	2.82	2.43	3.04	2.59	-
gpt-4o-mini	Zero-shot	**2.78**	**2.17**	**2.77**	**2.24**	**2.37**	**1.99**	**3.11**	**2.60**	**1.00**
	Step-by-step	2.82	2.23	3.02	2.42	2.67	2.13	3.52	2.89	2.63
	Few-shot	2.96	2.43	2.84	2.41	2.65	2.15	3.15	**2.60**	2.38
	PsyCoT	3.02	2.34	4.65	3.90	3.61	2.87	3.96	3.12	3.88
gpt-4o	Zero-shot	2.91	2.28	**2.90**	**2.23**	**2.62**	**2.06**	3.80	3.15	**1.75**
	Step-by-step	3.07	2.38	2.96	2.33	2.73	2.19	3.51	2.88	2.38
	Few-shot	**2.82**	**2.13**	3.34	2.74	2.76	2.16	**3.44**	**2.81**	1.88
	PsyCoT	3.36	2.67	5.02	4.25	3.23	2.49	5.20	4.27	4.00
Qwen2-7B	Zero-shot	3.11±0.00	2.59±0.00	**2.68**±0.01	**2.29**±0.01	2.90±0.00	2.48±0.00	3.16±0.02	2.65±0.01	1.75
	Step-by-step	3.26±0.08	2.70±0.06	2.69±0.00	2.30±0.00	2.92±0.00	2.50±0.00	3.16±0.00	2.66±0.00	2.88
	Few-shot	3.51±0.01	2.80±0.01	3.29±0.01	2.71±0.01	4.03±0.04	3.23±0.04	3.20±0.02	2.68±0.02	4.00
	PsyCoT	**2.68**±0.00	**2.30**±0.00	2.70±0.00	2.30±0.01	**2.82**±0.00	**2.43**±0.00	**3.10**±0.01	**2.62**±0.01	**1.38**
Qwen2-72B	Zero-shot	2.58±0.00	2.11±0.00	**2.70**±0.01	**2.28**±0.01	3.00±0.02	2.46±0.02	3.10±0.01	2.61±0.01	**1.75**
	Step-by-step	**2.58**±0.01	**2.10**±0.01	**2.70**±0.01	2.30±0.01	**2.92**±0.01	**2.41**±0.01	3.12±0.02	2.62±0.01	**1.75**
	Few-shot	2.76±0.02	2.26±0.01	3.09±0.02	2.56±0.02	3.29±0.03	2.70±0.02	**3.09**±0.01	**2.60**±0.01	2.75
	PsyCoT	2.64±0.01	2.15±0.01	3.99±0.03	3.23±0.03	4.59±0.03	3.75±0.03	4.66±0.04	3.75±0.04	3.75
Llama3.1-8B	Zero-shot	**2.77**±0.02	**2.39**±0.01	**2.83**±0.01	**2.37**±0.00	**2.86**±0.02	**2.44**±0.01	**3.10**±0.01	**2.64**±0.01	**1.00**
	Step-by-step	3.76±0.05	3.05±0.03	3.82±0.07	3.07±0.08	3.93±0.12	3.21±0.13	3.73±0.02	2.99±0.03	3.88
	Few-shot	3.67±0.00	3.00±0.01	3.42±0.00	2.77±0.00	3.66±0.00	3.01±0.00	3.44±0.00	2.87±0.00	2.50
	PsyCoT	3.31±0.02	2.80±0.02	3.55±0.00	2.93±0.01	3.54±0.00	2.96±0.00	3.70±0.01	3.12±0.00	2.63
Llama3.1-70B	Zero-shot	**2.68**±0.01	**2.22**±0.00	**3.15**±0.01	2.58±0.01	2.89±0.01	2.41±0.01	**3.24**±0.02	**2.72**±0.01	**1.50**
	Step-by-step	2.88±0.01	2.39±0.01	3.20±0.01	2.63±0.01	2.92±0.02	2.38±0.02	3.37±0.02	2.84±0.01	2.75
	Few-shot	2.93±0.02	2.47±0.02	3.17±0.02	**2.55**±0.02	**2.75**±0.01	**2.24**±0.02	3.25±0.02	2.78±0.01	2.00
	PsyCoT	2.94±0.02	2.34±0.01	3.96±0.01	3.17±0.01	3.65±0.03	2.93±0.03	4.00±0.02	3.19±0.03	3.75

Table 2: We conduct experiments using soft labels and use S-RMSE and S-MAE as evaluation metrics, where lower values indicate better performance. The best and second-best results across the four methods for each backbone are **bolded** and underlined.

differences between the predicted soft labels by LLMs and the estimated golden ones. Considering the limited precision of our estimated soft labels, we introduce Segmented MAE (S-MAE) and Segmented RMSE (S-RMSE). Instead of computing the error based solely on the raw predictions, we discretize the continuous 0-1 distribution into ten intervals and evaluate the corresponding RMSE and MAE within these segments. By framing the problem in this way, the error becomes less sensitive to specific small deviations in continuous values and more focused on whether the prediction falls within an appropriate range. Further details are introduced in Appendix F.4.

6 Results and Analysis

6.1 Model Evaluation

In this paper, we conduct experiments on MBTIBENCH across four inference prompt methods and six backbones introduced in Section 5.1. The main results of our experiments are shown in Table 2 and Figure 7. We address the following research questions (RQs):

RQ 1: How do different backbone models exhibit biases in soft label prediction? Different backbone models exhibit varying biases in their soft label predictions (Figure 7). GPT-4o shows less bias compared to Llama 3.1, while Qwen2 more frequently assigns a score of 5. We hypothesize that these discrepancies stem from differences in their training corpora, leading to varied score distributions.

RQ 2: How do different methods perform on MBTIBENCH? The Zero-shot method demonstrates superior performance across all six backbone models. As shown in Table 2, it achieves the lowest average RMSE and MAE. In contrast, other methods may amplify the inherent biases of the models. For example, the PsyCoT method tends to predict 0 in certain dimensions, while few-shot methods are more influenced by the provided examples.

RQ 3: How does model performance vary across different dimensions? Different backbone models exhibit varying biases in their soft label pre-

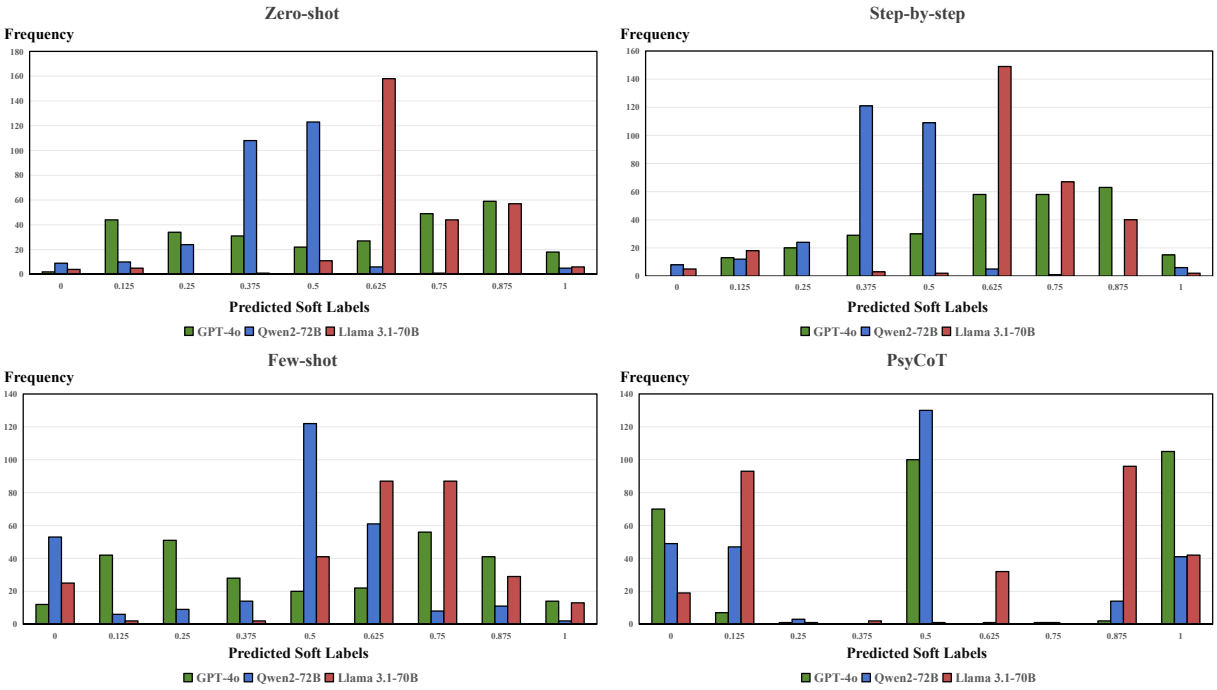


Figure 7: The score distribution of LLMs on MBTIBENCH for the *T/F* dimension.

dictions (Figure 7 and Appendix Figure 20- 22).[9] GPT-4o shows less bias compared to Llama 3.1, while Qwen2 more frequently assigns a score of 5. We hypothesize that these discrepancies stem from differences in their training corpora, leading to varied score distributions.

6.2 Can Soft Labels Better Align with Population Traits?

To validate the effectiveness of soft labels, we follow Turcan and McKeown (2019) and incorporat personality traits into the stress identification task. This task is both challenging and representative, with substantial real-world significance. We use the Dreaddit stance detection dataset (Turcan and McKeown, 2019), where the LLM is provided with a *post* from one poster and tasked with determining whether the poster is likely to suffer from very severe stress or not (*w/o MBTI*). We predict the user's MBTI personality types using both hard-label (*w/ Hard*) and soft-label (*w/ Soft*) approaches with LLM, based on the *post*. These personality types are then used as auxiliary inputs for stress identification. To evaluate the robustness of our results, we repeat the stress identification experiments 10 times and perform a t-test to assess the statistical significance of the outcomes. The results are shown in Figure 8.

[9]We show the results of each first round on three backbones.

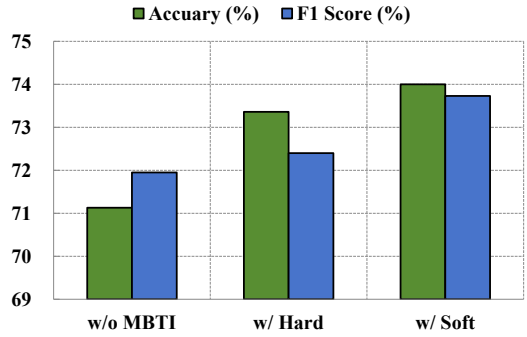


Figure 8: We integrate personality traits into stress identification, demonstrating that the use of soft labels (*w/ Soft*) significantly enhances performance on personality-related tasks.

We analyze the experimental results conclude that: (1) The inferior performance of *w/o MBTI* (Acc: 72.13, F1: 71.95) indicates that MBTI personality traits contribute to stress identification. (2) *w/ Soft* (Acc: 74.00, F1: 73.73) outperforms *w/ Hard* (Acc: 73.36, F1: 72.40), indicating that soft labels more accurately align with population personality traits than hard labels.

7 Related Works

The Big Five and MBTI are two widely used personality frameworks in the fields of computational linguistics and natural language processing (Yang et al., 2021). In this paper, we focus on MBTI, one of the most widely used non-clinical psychometric

assessments because it translates well into the behavioral context (Stajner and Yenikent, 2021). It is widely adopted in diverse real-world applications like understanding team building processes in work environments (Kuipers et al., 2009), for career suggestions (Garden and Sloan, 2011), and in marketing and consumer behavior (Gountas and Gountas, 2000). MBTI distinguishes itself in applied psychological settings by offering an approach that simplifies interpersonal and organizational dynamics, making it particularly valuable for enhancing team functionality and personal development initiatives (Myers et al., 1980).

The quality issues of datasets will affect the accurate evaluation of LLM performance and the iteration of methods in personality detection tasks (Liu et al., 2023b). There have been some early explorations of dataset quality in previous work (Stajner and Yenikent, 2021), but they only annotated one commonly used Twitter dataset. These studies do not consider the impact of data noise and do not annotate complete samples. Additionally, these works do not validate their conclusions through model experiments. These limitations indicate that there is still room for improvement in the discussion of dataset quality.

8 Conclusion

MBTI is one of the most popular personality theories and has garnered considerable research interest over the years. In this paper, we point out the over-optimism in MBTI personality detection from the nature of population personality traits. Specifically, (1) the self-reported labels in existing datasets result in data quality issues and (2) the hard labels fail to capture the full range of population personality distributions. We construct MBTIBENCH, the first high-quality manually annotated MBTI personality detection dataset with soft labels, under the guidance of psychologists. We filter and annotate the data, and estimate soft labels by deriving the polarity tendency of samples. Experiment results highlight the polarized predictions and bias as future directions. We integrate personality traits into stance detection, demonstrating the effectiveness of soft labels on personality-related tasks.

Acknowledgments

We gratefully acknowledge the support of the National Natural Science Foundation of China (NSFC) via grant 62236004, 62206078, 62441603 and 62476073.

We are immensely grateful to Professor Mengyue Wu for her insightful guidance and encouragement, which continue to be instrumental in shaping the core contributions of our paper. We also extend our sincere thanks to Zeming Liu and Jun Xu for their generous contributions of time and expertise in offering suggestions on manuscript writing. Their input significantly enhances the quality and depth of our paper. We are deeply grateful to Xin Hao for the insightful suggestions regarding the annotation guidelines, which are invaluable to this paper. We would like to express our profound appreciation to Lina Zhang for her meticulous work on the figures, which have become a highlight of our paper.

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*.

Shlomo Engelson Argamon, Sushant Dhawle, Moshe Koppel, and James W. Pennebaker. 2005. Lexical predictors ofpersonality type.

R Michael Bagby, Tara M Gralnick, Nadia Al-Dajani, and Amanda A Uliaszek. 2016. The role of the five-factor model in personality assessment and treatment planning. *Clinical Psychology: Science and Practice*, 23(4):365.

Fabio Celli and B. Lepri. 2018. Is big five better than mbti? a personality computing challenge using twitter data. In *Italian Conference on Computational Linguistics*.

Bao Chen, Yuanjie Wang, Zeming Liu, and Yuhang Guo. 2023. Automatic evaluate dialogue appropriateness by using dialogue act. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7361–7372, Singapore. Association for Computational Linguistics.

Mark Chen and Jerry Tworek et.al. 2021. Evaluating large language models trained on code. *ArXiv*, abs/2107.03374.

Cindy Chung and James Pennebaker. 2011. The psychological functions of function words. In *Social communication*, pages 343–359. Psychology Press.

Philip J Corr and Gerald Ed Matthews. 2009. *The Cambridge handbook of personality psychology*. Cambridge University Press.

A. Philip Dawid and Allan Skene. 1979. Maximum likelihood estimation of observer error-rates using

the em algorithm. *Journal of The Royal Statistical Society Series C-applied Statistics*, 28:20–28.

Arthur P. Dempster, Nan M. Laird, and Donald B. Rubin. 1977. Maximum likelihood from incomplete data via the em - algorithm plus discussions on the paper.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76:378–382.

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*, 21:303–307.

Annamaria Garden and Sloan. 2011. Relationships between mbti profiles , motivation profiles , and career paths.

Alastair J. Gill and Jon Oberlander. 2019. Taking care of the linguistic features of extraversion. *Proceedings of the Twenty-Fourth Annual Conference of the Cognitive Science Society*.

Matej Gjurković, Mladen Karan, Iva Vukojević, Mihaela Bošnjak, and Jan Snajder. 2021. PANDORA talks: Personality and demographics on Reddit. In *Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media*, pages 138–152.

John Gountas and Sandra Gountas. 2000. A new psychographic segmentation method using jungian mbti variables in the tourism industry. *Tourism Analysis*, 5:151–156.

Robert James Harvey and William D. Murry. 1994. Scoring the myers-briggs type indicator: Empirical comparison of preference score versus latent-trait methods. *Journal of Personality Assessment*, 62:116–129.

Marjan Hosseinia, Eduard Constantin Dragut, Dainis Boumber, and Arjun Mukherjee. 2021. On the usefulness of personality traits in opinion-oriented tasks. In *Recent Advances in Natural Language Processing*.

Hang Jiang, Xianzhe Zhang, and Jinho D. Choi. 2019. Automatic text-based personality recognition on monologues and multiparty dialogues using attentive networks and contextual embeddings. *ArXiv*, abs/1911.09304.

Amir A Khan, Kristen C Jacobson, Charles O Gardner, Carol A Prescott, and Kenneth S Kendler. 2005. Personality and comorbidity of common psychiatric disorders. *The British Journal of Psychiatry*, 186(3):190–196.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *ICML 2022 Workshop on Knowledge Retrieval and Language Models*.

Michal Kosinski, David Stillwell, and Thore Graepel. 2013. Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences*, 110:5802 – 5805.

Ben S Kuipers, Malcolm J Higgs, Natalia V Tolkacheva, and Marco C de Witte. 2009. The influence of myers-briggs type indicator profiles on team development processes: An empirical study in the manufacturing industry. *Small Group Research*, 40(4):436–464.

Rensis Likert. 1932. A technique for the measurement of attitude scales.

William F. Line. 1948. Description and measurement of personality. *American Journal of Psychiatry*, 104:591–592.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023a. Visual instruction tuning. *ArXiv*, abs/2304.08485.

Zeming Liu, Ping Nie, Jie Cai, Haifeng Wang, Zheng-Yu Niu, Peng Zhang, Mrinmaya Sachan, and Kaiping Peng. 2023b. XDailyDialog: A multilingual parallel dialogue corpus. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12240–12253, Toronto, Canada. Association for Computational Linguistics.

Zeming Liu, Haifeng Wang, Zheng-Yu Niu, Hua Wu, Wanxiang Che, and Ting Liu. 2020. Towards conversational recommendation over multi-type dialogs. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1036–1049, Online. Association for Computational Linguistics.

Zeming Liu, Jun Xu, Zeyang Lei, Haifeng Wang, Zheng-Yu Niu, and Hua Wu. 2022. Where to go for the holidays: Towards mixed-type dialogs for clarification of user goals. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1024–1034, Dublin, Ireland. Association for Computational Linguistics.

François Mairesse, Marilyn A. Walker, Matthias R. Mehl, and Roger K. Moore. 2007. Using linguistic cues for the automatic recognition of personality in conversation and text. *J. Artif. Intell. Res.*, 30:457–500.

Jennifer McDonald. 2008. Measuring personality constructs: The advantages and disadvantages of self-reports, informant reports and behavioural assessments.

Sascha Müller and Morten Moshagen. 2019. Controlling for response bias in self-ratings of personality: A comparison of impression management scales and the overclaiming technique. *Journal of Personality Assessment*, 101:229 – 236.

Isabel Briggs Myers et al. 1980. Gifts differing: Understanding personality type.

R. Passonneau and Bob Carpenter. 2013. The benefits of a model of annotation. *Transactions of the Association for Computational Linguistics*, 2:573–573.

Delroy L. Paulhus and Simine Vazire. 2007. The self-report method.

James W. Pennebaker and Laura A. King. 1999. Linguistic styles: language use as an individual difference. *Journal of personality and social psychology*, 77 6:1296–312.

Barbara Plank and Dirk Hovy. 2015. Personality traits on twitter—or—how to get 1,500 personality tests in a week. In *WASSA@EMNLP*.

Vikas C Raykar, Shipeng Yu, Linda H Zhao, Gerardo Hermosillo Valadez, Charles Florin, Luca Bogoni, and Linda Moy. 2010a. Learning from crowds. *Journal of machine learning research*, 11(4).

Vikas Chandrakant Raykar, Shipeng Yu, Linda H. Zhao, Gerardo Hermosillo, Charles Florin, Luca Bogoni, and Linda Moy. 2010b. Learning from crowds. *J. Mach. Learn. Res.*, 11:1297–1322.

Klaus R. Scherer. 2003. Vocal communication of emotion: A review of research paradigms. *Speech Commun.*, 40:227–256.

Xiaoming Shi, Zeming Liu, Chuan Wang, Haitao Leng, Kui Xue, Xiaofan Zhang, and Shaoting Zhang. 2023. MidMed: Towards mixed-type dialogues for medical consultation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8145–8157, Toronto, Canada. Association for Computational Linguistics.

Sanja Stajner and Seren Yenikent. 2020. A survey of automatic personality detection from texts. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6284–6295, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Sanja Stajner and Seren Yenikent. 2021. Why is mbti personality detection from texts a difficult task? In *Conference of the European Chapter of the Association for Computational Linguistics*.

Elsbeth Turcan and Kathy McKeown. 2019. Dreaddit: A Reddit dataset for stress analysis in social media. In *Proceedings of the Tenth International Workshop on Health Text Mining and Information Analysis (LOUHI 2019)*, pages 97–107, Hong Kong. Association for Computational Linguistics.

Oliver C. S. Tzeng, Roger Ware, and Jeaw mei Chen. 1989. Measurement and utility of continuous unipolar ratings for the myers-briggs type indicator. *Journal of Personality Assessment*, 53:727–738.

Ben Verhoeven, Walter Daelemans, and Barbara Plank. 2016. Twisty: A multilingual twitter stylometry corpus for gender and personality profiling. In *International Conference on Language Resources and Evaluation*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*.

Cort J. Willmott and Kenji Matsuura. 2005. Advantages of the mean absolute error (mae) over the root mean square error (rmse) in assessing average model performance. *Climate Research*, 30:79–82.

Wen Wu, C. Zhang, Xixin Wu, and Philip C. Woodland. 2022. Estimating the uncertainty in emotion class labels with utterance-specific dirichlet priors. *IEEE Transactions on Affective Computing*, 14:2810–2822.

Kosuke Yamada, Ryohei Sasano, and Koichi Takeda. 2019. Incorporating textual information on user behavior for personality prediction. In *Annual Meeting of the Association for Computational Linguistics*.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Feifan Yang, Tao Yang, Xiaojun Quan, and Qinliang Su. 2021. Learning to answer psychological questionnaire for personality detection. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1131–1142.

Tao Yang, Tianyuan Shi, Fanqi Wan, Xiaojun Quan, Qifan Wang, Bingzhe Wu, and Jiaxiang Wu. 2023. Psycot: Psychological questionnaire as powerful chain-of-thought for personality detection. *ArXiv*, abs/2310.20256.