Predicting the Big Five Personality Traits in Chinese Counselling Dialogues Using Large Language Models

Anonymous ACL submission

Abstract

Accurate assessment of personality traits is crucial for effective psycho-counseling, yet traditional methods like self-report questionnaires are time-consuming and biased. This study exams whether Large Language Models (LLMs) can predict the Big Five personality traits directly from counseling dialogues and introduces an innovative framework to perform the task. Our framework applies role-play and questionnaire-based prompting to condition LLMs on counseling sessions, simulating client responses to the Big Five Inventory. We evaluated our framework on 853 real-world counseling sessions, finding a significant correlation between LLM-predicted and actual Big Five traits, proving the validity of framework. Moreover, ablation studies highlight the importance of role-play simulations and task simplification via questionnaires in enhancing prediction accuracy. Meanwhile, our fine-tuned Llama3-8B model, utilizing Direct Preference Optimization with Supervised Fine-Tuning, achieves a 130.95% improvement, surpassing the state-of-the-art Qwen1.5-110B by 36.94% in personality prediction validity. In conclusion, LLMs can predict personality based on counseling dialogues. Our code and model are publicly available at `https://github.com/Anonymous-gwFabfaH/BigFive-LLM-Predictor`, providing a valuable tool for future research in computational psychometrics.

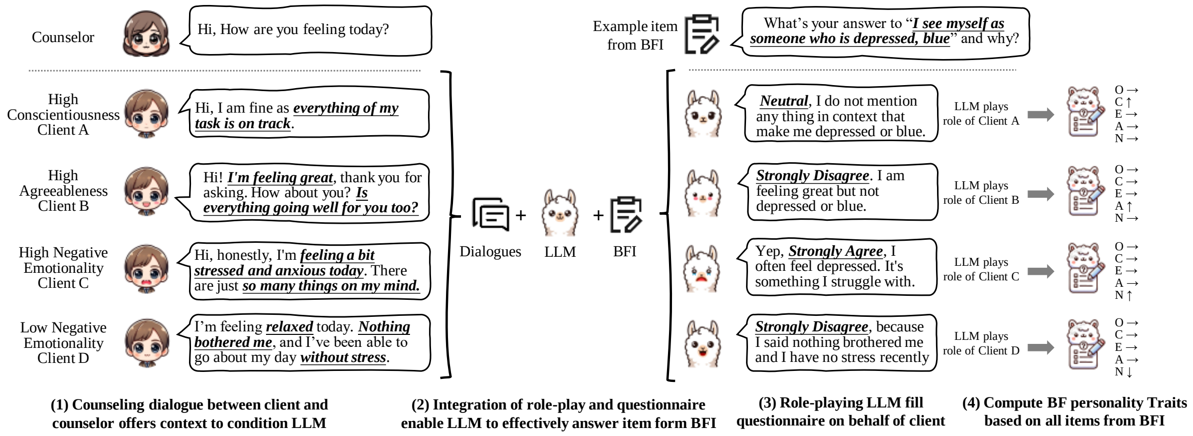
1 Introduction

Understanding clients' personality traits is crucial for effective psycho-counseling, as personalized advice tailored to these traits can significantly enhance the quality of counseling (Gordon and Toukmanian, 2002; Anestis et al., 2021). However, it remains challenging to effectively assess personality traits through counseling dialogue. Traditional methods, such as self-report questionnaires (e.g., Big Five Inventory, BFI) (John et al., 1991), grounded in Item Response Theory (Baker, 2001; Reise and Waller, 2009; Embretson and Reise, 2013), require people to complete extensive lists of questions. Nevertheless, collecting clients' personality information via self-report questionnaires is time-consuming and influenced by subjective biases and social desirability effects (Chernyshenko et al., 2001; McCrae and Weiss, 2007; Khorramdel and von Davier, 2014), making the quest for an automatic and effective method to assess personality traits without direct participation of clients has become a significant research frontier in both psychometrics and computational linguistics (Korukonda, 2007; Chittaranjan et al., 2011; Gavrilescu and Vizireanu, 2018; Cai and Liu, 2022).

Recent developments in Large Language Models (LLMs) (Brown et al., 2020; OpenAI, 2023; Bai et al., 2023; Gemini-Team, 2024) have demonstrated capabilities in text comprehension, reasoning, and role-playing, capturing dynamic and context-sensitive aspects of human interactions in natrual language (Ng et al., 2024). The development shows potential to address the issue of time-consumming and bias of self-report measures in the field of psychometrics. Meanwhile, considering the significance of knowing clients' personality in psycho-counseling (Gordon and Toukmanian, 2002; Anestis et al., 2021), we pose the research question: **Can LLMs predict personality traits based on counseling dialogues?** The question drives our investigation into the potential of LLMs to accurately predict Big Five personality traits, known as OCEAN [1], from counseling dialogues, exploring both prompting and alignment strategies.

To investigate the capability of LLMs in predicting personality in the counseling dialogues, we unfold our framework of personality prediction in three stages. First, we evaluated the validity

[1] The acronym "OCEAN" represents the Big Five (BF) personality traits: **O**pen mindedness, **C**onscientiousness, **E**xtraversion, **A**greeableness, and **N**egative Emotionality.


Counselor
Hi, How are you feeling today?
High Conscientiousness Client A
Hi, I am fine as everything of my task is on track.
High Agreeableness Client B
Hi! I'm feeling great, thank you for asking. How about you? Is everything going well for you too?
High Negative Emotionality Client C
Hi, honestly, I'm feeling a bit stressed and anxious today. There are just so many things on my mind.
Low Negative Emotionality Client D
I'm feeling relaxed today. Nothing bothered me, and I've been able to go about my day without stress.
Dialogues
+
LLM
+
BFI
Example item from BFI
What's your answer to "I see myself as someone who is depressed, blue" and why?
Neutral, I do not mention any thing in context that make me depressed or blue.
LLM plays role of Client A
O → C ↑ E → A → N →
Strongly Disagree. I am feeling great but not depressed or blue.
LLM plays role of Client B
O → C → E → A ↑ N →
Yep, Strongly Agree, I often feel depressed. It's something I struggle with.
LLM plays role of Client C
O → C → E → A → N ↑
Strongly Disagree, because I said nothing brothered me and I have no stress recently
LLM plays role of Client D
O → C → E → A → N ↓

(1) Counseling dialogue between client and counselor offers context to condition LLM

(2) Integration of role-play and questionnaire enable LLM to effectively answer item form BFI

(3) Role-playing LLM fill questionnaire on behalf of client

(4) Compute BF personality Traits based on all items from BFI

Figure 1: **Example for our framework of prediction OCEAN traits from counseling dialogues.** Our framework includes integral step: conditioning LLM on the counseling dialogues, prompting the LLM with role-play and questionnaire, and let LLM complete questionnaire on belf of the client to get the prediction of OCEAN traits.

of prompt strategies using role-play scenarios and questionnaire-based approaches to predict OCEAN traits. Second, we examined factors influencing the validity of prediction, including the roles of role-play, the granularity of counseling sessions, and the types and sizes of LLMs. Third, we improved the performance of LLMs by fine-tuning with generated reasoning results from the second step, aiming to increase the validity and efficiency of personality prediction.

To validate our framework, we performed an extensive assessment on 853 real-world counseling sessions, juxtaposing the OCEAN traits predicted by the LLM with the ground-truth traits obtained from 83 clients using Pearson Correlation Coefficients (PCC) and Mean Averaged Error (MAE). We found the correlation between model prediction and ground truth is robust and significant. Additionally, a detailed error analysis across models and clients highlights the strengths and weaknesses of our framework, providing informative directions for future studies.

We present our contributions as follows:

1. We introduced a novel framework that integrates role-playing and questionnaire prompting strategies to predict OCEAN traits in counseling dialogues. An evaluation of 853 counseling sessions demonstrates a strong correlation between predicted and actual traits. Besides, the assessment of content validity shows that our framework detects subjective biases and social desirability, enhancing its analytical depth.
2. Comprehensive ablation studies indicate that aligning roles with specific tasks and decomposing complex tasks into simpler items significantly improve trait prediction accuracy. Remarkably, our approach achieves accurate OCEAN trait prediction using only 30% of session content.
3. By aligning the Llama3-8B model with trait prediction through Direct Preference Optimization (DPO) and Supervised Fine-Tuning (SFT), our fine-tuned lightweight model exhibits a 130.95% improvement in prediction validity, surpassing the state-of-the-art Qwen1.5-110B by 36.94%, demonstrating superior validity and efficiency.
4. We release our codes and models to support future research, offering an effective and efficient tool in computational psychometrics, fostering reproducibility and further exploration.

2 Related Work

Automatic Personality Assessment Recent studies have explored the Myers-Briggs Type Indicator (MBTI) (Myers, 1962) as a tool to assess personality traits with LLMs. Rao et al. (2023) tried to generate unbiased prompts for ChatGPT to assess human personalities based on MBTI tests and reported positive results, indicating the synergy between psychological assessments and LLM technology. However, the existing work with LLMs mainly focused on MBTI, which is not as valid nor reliable as the BFI is (John et al., 1991). Although some early attempts to predict OCEAN traits automatically from textual data employed machine learning and NLP techniques, for example, Sun et al. (2018); Mehta et al. (2020); Christian et al. (2021) applied traditional deep learning models, such as LSTM, language model embedding, or pretrained models to predict personality traits from the essay datasets or users' posts on various so-

		Open Mindedness	Conscientiousness	Extraversion	Agreeableness	Negative Emotionality	Avg.
Method	Model						
Baseline	Llama-3-8b-BFI (Ours)	-0.004	0.113	0.186	0.025	-0.070	0.050
	Qwen1.5-110B-Chat	0.267*	0.167	0.190	0.091	0.142	0.172
	deepseek-chat	0.143	0.067	0.216	-0.010	-0.017	0.080
+ *Role-Play Only*	Llama-3-8b-BFI (Ours)	-0.018	0.129	-0.132	0.174	0.115	0.053
	Qwen1.5-110B-Chat	0.006	0.162	-0.096	0.227	-0.028	0.054
	deepseek-chat	0.101	-0.172	0.158	-0.000	0.293*	0.076
+ *Questionnaire Only*	Llama-3-8b-BFI (Ours)	0.452***	0.459***	0.421***	0.228	0.515***	0.415
	Qwen1.5-110B-Chat	0.292*	0.332**	0.391***	0.257*	0.324**	0.319
	deepseek-chat	0.311**	0.194	0.317**	0.206	0.391***	0.284
+ *Role-Play and Questionnaire*	Llama-3-8b-BFI (Ours)	0.692***	0.554***	0.569***	0.448***	0.648***	0.582
	Qwen1.5-110B-Chat	0.455***	0.463***	0.521***	0.334**	0.354**	0.426
	deepseek-chat	0.443***	0.385**	0.434***	0.337**	0.379**	0.395

Table 1: **PCC of Various Methods for Predicting OCEAN traits.** We assessed the validity of direct personality prediction using LLMs, comparing baseline performance with enhancements via role-play, questionnaires, and their combination. Our results demonstrate that integrating role-play and questionnaire prompts significantly improves prediction accuracy. Significance levels are indicated as follows: * (p < 0.05), ** (p < 0.01), and *** (p < 0.001).

cial media, there is little research on predicting OCEAN traits directly from counseling dialogues. This gap underscores the need for an effective and reliable framework for predicting OCEAN traits in psycho-counseling, and motivates our research.

Prompting Strategies Advanced prompting strategies are essential to fully utilize the capabilities of LLMs. Chain-of-Thought (CoT) (Wei et al., 2022) and its successors enhance LLM reasoning by decomposing complex tasks into simpler steps (Singh et al., 2023; Lin et al., 2023; Yao et al., 2023; Besta et al., 2024), suggesting that a similar approach could be applied to predict personality traits. Furthermore, role-playing techniques enable LLMs to simulate human-like agents (Shanahan et al., 2023; Salemi et al., 2023; Park et al., 2023; Wang et al., 2024b,a; Kong et al., 2024). Studies have demonstrated the effectiveness of role-play in solving complex tasks (Li et al., 2023; Chen et al., 2024; Wang et al., 2024b; Qian et al., 2024; Kong et al., 2024), facilitating interaction without actual human participation. Specifically, Wang et al. (2024a) attempts to use role-play agents of virtual characters to predict their personalities. Despite these advancements and their potential for personality prediction, their use in predicting OCEAN traits within counseling dialogues has not been thoroughly investigated. Therefore, further research is needed to evaluate the effectiveness of these strategies in predicting OCEAN traits in such contexts.

Alignment Strategies Aligning LLMs with human preferences is crucial for optimal performance. Reinforcement Learning from Human Feedback (RLHF) (Ouyang et al., 2022) demonstrates significant performance improvements using a human preference ranker with Proximal Policy Optimization (PPO) (Schulman et al., 2017). Rafailov et al. (2023) introduces DPO, parametrizing the reward function to address PPO's complexity and instability. Despite advances, recent studies (Feng et al., 2024; Xu et al., 2024) identify the limitations of DPO, which reduces dispreferred data generation but does not enhance preferred output production. Pang et al. (2024) proposed to add negative log-likelihood loss to a custom DPO loss to address this issue. In addition to RLHF, several successful LLMs (Touvron et al., 2023; Liu et al., 2023) employ SFT with high-quality data for alignment and generation quality. Whether these strategies can benefit the prediction of OCEAN traits in counseling dialogues remains unexplored, leaving a gap in the literature that our research aims to fill.

3 Framework for Predicting OCEAN traits

Our proposed framework consists of three key components: 1. prompting strategy design, 2. LLM conditioning, and 3. evaluation metrics. Together, these elements ensure the validity and reliability of the method.

3.1 Prompting Strategy Design

Our prompting strategy combines role-play and questionnaires. The role-play includes three roles: client, counselor (primary participants), and observer (external evaluator). The questionnaire uses items from the BFI to simplify the prediction task.

Our prompt consists of the following elements:

1. Task and Role-play Settings: Task descriptions specify the LLM's identity, the input it will process, and its expected actions. Role-play settings introduce the role, outlining its capabilities and responsibilities. These foundational elements are crucial for the LLM to understand the task requirements and role-play context.

		Open Mindedness	Conscientiousness	Extraversion	Agreeableness	Negative Emotionality	Avg.
Role	Model						
client	Llama-3-8b-BFI (Ours)	0.692***	0.554***	0.569***	0.448***	0.648***	0.582
	Qwen1.5-110B-Chat	0.455***	0.463***	0.521***	0.334**	0.354**	0.426
	deepseek-chat	0.443***	0.385**	0.434***	0.337**	0.379**	0.395
counselor	Llama-3-8b-BFI (Ours)	0.652***	0.586***	0.550***	0.412***	0.539***	0.548
	Qwen1.5-110B-Chat	0.314**	0.354**	0.488***	0.050	0.422***	0.326
	deepseek-chat	0.367**	0.378**	0.342**	0.305*	0.379**	0.354
observer	Llama-3-8b-BFI (Ours)	0.499***	0.560***	0.476***	0.357**	0.483***	0.475
	Qwen1.5-110B-Chat	0.375**	0.341**	0.436***	0.378**	0.400***	0.386
	deepseek-chat	0.419***	0.256*	0.389**	0.221	0.442***	0.346
no-role	Llama-3-8b-BFI (Ours)	0.452***	0.459***	0.421***	0.228	0.515***	0.415
	Qwen1.5-110B-Chat	0.292*	0.332**	0.391***	0.257*	0.324**	0.319
	deepseek-chat	0.311**	0.194	0.317**	0.206	0.391***	0.284

Table 2: **PCC of Various Roles for Predicting OCEAN traits.** We assessed the prediction validity of OCEAN traits in our framework under various roles: client, counselor, observer, and no-role. The roles of the client and the counselor showed significantly higher prediction accuracy compared to the role of the observer as native participants in counseling. The no-role condition had the lowest performance, highlighting the importance of contextual role-play in enhancing model predictions.

2. Counseling Dialogues: Counseling dialogues between counselor and client provide the LLM with essential contextual information. These real-world dialogues are formatted into a chat history structure, consistent with the LLM's pre-training schema, enabling LLM to effectively simulate the client's responses, thereby improving the accuracy of OCEAN trait predictions.

3. Prediction Objective: The questions of BFI are set as the prediction objective, guiding the LLM to predict responses to them. This approach ensures that outputs of LLMs align with the validated psychological assessments.

A typical client prompt is structured as follows:

System Prompt: Act like a real human and do not mention anything with AI. Act as the client in this counseling session, you will have a conversation with your counselor.

—

User: {utterance 1 from counselor}
LLM: {utterance 1 from client}
User: {utterance 2 from counselor}
LLM: {utterance 2 from client}
...
User: Before we end today's counseling session, please complete the following questionnaire based on the conversation and your own situation:

—

Question: {item from BFI}
Options:
1. Disagree (strongly)
2. Disagree (a little)
3. Neutral (no opinion)
4. Agree (a little)
5. Agree (strongly)

—

Please tell me your choice and explain the reason:

This approach enhances the model's ability to generate contextually appropriate responses, thus improving prediction validity. Detailed prompts and BFI items are provided in Sec. A.3 and Sec. A.1, respectively.

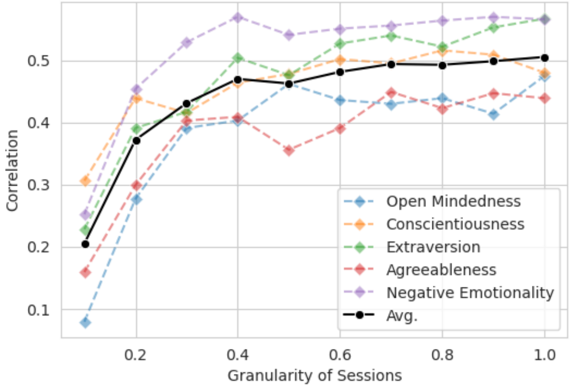


Figure 2: **PCC Changes Across Different Dialogue Session Granularities.** The plots illustrate that the PCC increases rapidly up to 30% of the dialogue context, beyond which the increase is slower. This observation, corroborated by Tab. 8 showing significant PCC at 30% session granularity, indicates that 30% of the dialogue context suffices for predicting OCEAN traits.

3.2 LLM Conditioning for OCEAN trait Prediction

To elucidate the prediction process, we frame the task as conditional generation, as depicted in Eq. 1.

$$y_{\text{trait}} = \text{LLM}(x_{\text{context}}, \text{questionnaire}) \quad (1)$$

Here, x_{context} denotes historical counseling dialogues, and questionnaire refers to the BFI items within the prompt. The LLM, denoted as LLM, generates a response y_{trait} to each BFI item based on the provided context x_{context}. Each y_{trait} includes both the choice and rationale for the BFI item. We extract the choice using keyword-based regex. After predicting responses for all 60 items, we compute the OCEAN traits following the BFI scoring system (Soto and John, 2017).

Model	Open Mindedness	Conscientiousness	Extraversion	Agreeableness	Negative Emotionality	Avg.
GPT-4-turbo (OpenAI, 2023)	0.407***	0.360**	0.507***	0.303*	0.337**	0.383
deepseek-chat (DeepSeek-AI et al., 2024b)	0.443***	0.385**	0.434***	0.337**	0.379**	0.395
gemini-1.5-pro-latest (Gemini-Team, 2024)	0.521***	0.438***	0.494***	0.356**	0.314**	0.425
gemini-1.5-flash-latest (Gemini-Team, 2024)	0.306*	0.351**	0.252*	0.358**	0.330**	0.319
gemini-1.0-ultra-latest (Gemini-Team, 2024)	0.408***	0.317**	0.372**	0.057	0.309*	0.293
gemini-1.0-pro-001 (Gemini-Team, 2024)	0.337**	0.305*	0.295*	0.119	0.317**	0.275
qwen-long (Bai et al., 2023)	0.346**	0.376**	0.451***	0.265*	0.405***	0.369
qwen-turbo (Bai et al., 2023)	0.363**	0.314**	0.418***	0.279*	0.321**	0.339
ERNIE-Speed-128K (Baidu, 2023)	0.138	0.167	0.241*	-0.203	0.239*	0.116
ERNIE-Lite-8K-0308 (Baidu, 2023)	-0.119	-0.032	0.150	-0.236	0.267*	0.006
Qwen1.5-110B-Chat (Bai et al., 2023)	0.455***	0.463***	0.521***	0.334**	0.354**	0.425
Qwen-72B-Chat (Bai et al., 2023)	0.309*	0.396***	0.419***	0.421***	0.440***	0.397
Meta-Llama-3-70B-Instruct (Meta, 2024)	0.397***	0.467***	0.395***	0.284*	0.289*	0.366
deepseek-llm-67b-chat (DeepSeek-AI et al., 2024a)	0.303*	0.336**	0.491***	0.196	0.301*	0.325
Yi-34B-Chat (AI et al., 2024)	0.399***	0.243*	0.448***	0.297*	0.204	0.318
AquilaChat2-34B (BAAI, 2024)	0.085	-0.059	0.126	0.035	0.248*	0.087
internlm2-chat-20b (Cai et al., 2024)	0.341**	0.201	0.368**	0.260*	0.255*	0.285
Baichuan2-13B-Chat (Yang et al., 2023)	-0.019	0.192	0.173	0.183	-0.094	0.087
glm-4-9b-chat (Zeng et al., 2023)	0.293*	0.312**	0.240*	0.036	0.305*	0.237
gemma-1.1-7b-it (Gemma-Team, 2024)	0.054	0.330**	0.364**	-0.053	0.034	0.146
chatglm3-6b-128k (Zeng et al., 2023)	0.057	0.054	0.005	0.062	0.011	0.038
Meta-Llama-3-8B-Instruct (Meta, 2024)	0.177	0.434***	0.233	0.111	0.303*	0.252
Llama-3-8b-BFI (Ours)	**0.692*** **	**0.554*** **	**0.569*** **	**0.448*** **	**0.648*** **	**0.582**

Table 3: **PCC of Various LLMs for Predicting OCEAN traits.** Highest PCC values per dimension are highlighted in bold. The models include state-of-the-art proprietary and open-source models. Among open-source models, Qwen1.5-110B-Chat and Qwen-72B-Chat performed best, while Gemini-1.5-Pro and Deepseek-Chat led among proprietary models. In particular, our fine-tuned Llama-3-8b-BFI model, despite its smaller size, surpassed all other models, achieving the highest and most significant PCC. This underscores the validity and efficiency of our framework and tailored fine-tuning approach.

Factors such as the type and configuration of the LLM, and the detail level of the context, can affect prediction validity. We exam the impact of these factors in the following experiments.

3.3 Evaluation Metrics

We employ validity and reliability metrics to evaluate the effectiveness of our framework, adhering to best practices in psychological research (John et al., 1991; Soto and John, 2017).

Validity Validity measures the test's accuracy and relevance, encompassing two key aspects:

1. *Criterion Validity* evaluates the alignment between predictions and ground truth. We use PCC, a standard in psychology, to assess the strength and significance of the association between predicted and actual OCEAN traits. Additionally, MAE is included for a detailed analysis of prediction errors.

2. *Content Validity* examines the justification behind predictions. By analyzing predictions with the highest and lowest accuracy, we identify factors contributing to their performance. This dual analysis provides insights into the content validity of our framework by highlighting areas of close alignment and divergence from the ground truth.

Reliability Reliability is evaluated through internal consistency and test-retest reliability, detailed in Sec. A.4.

4 Experiments

We collected counseling dialogues and structured our experiments around three primary research questions (RQs) to evaluate our framework's performance systematically.

4.1 Data Collection and Preprocessing

We gather 853 counseling dialogues from 82 adult clients (55 females, age range 19-54 years old, M=27.62 years old, SD=5.94) and 9 counselors (7 females, age range 25-45, M=34.67 years old, SD=7.45), summarized in Tab. 5. Before their initial sessions, clients completed the Chinese version for BFI-2 (Soto and John, 2017), linking dialogue analyzes with established personality profiles.

Approximately 30% (242) of the dialogues were allocated to the validation set, while the remaining 70% (611) were used for training. We manually anonymized the validation set to ensure privacy by replacing all personally identifiable information with placeholders, underscoring our commitment to ethical standards and data protection.

4.2 RQ1: Can LLMs predict OCEAN traits from counseling dialogues?

We began by evaluating the feasibility and criterion validity of predicting OCEAN traits from counseling dialogues using LLMs. Initially, we set the baseline by predicting OCEAN traits di-

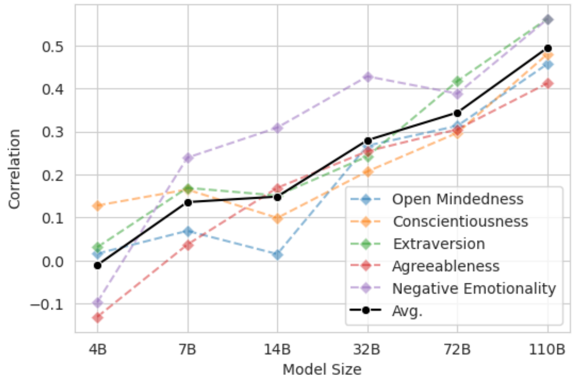


Figure 3: **PCC Changes Across Different Model Sizes.** The plots demonstrate a positive correlation between model size and average PCC in the "Qwen1.5" series. However, statistical significance is only observed for Qwen1.5-110B-Chat and Qwen1.5-72B-Chat models. These findings indicate that effective zero-shot personality prediction demands substantial highly capable models as well as significant computational resources.

rectly from dialogues without additional strategies. We then enhanced the baseline with role-play and questionnaires-based strategy, and conducted ablation studies on various variables in Eq. 1 to assess the prediction validity.

Role-play and Questionnaires Impact As shown in Tab. 1, the baseline prediction of OCEAN traits from dialogues alone was poor due to the complexity and nuance of the task. Adding role-play contributed minimally, while questionnaires showed a slight improvement, indicating that decomposing the task into simpler items is beneficial. Combining role-play and questionnaires significantly improved prediction validity across all OCEAN traits. This aligns with Item Response Theory (Baker, 2001; Reise and Waller, 2009; Embretson and Reise, 2013), suggesting that direct personality assessment is challenging and tools like questionnaires are essential. Role-play enhances prediction validity by helping LLMs better understand context as role proximity increases.

Enhanced Validity via Role Proximity Given that role proximity enhances prediction validity, we further investigated the impact of different roles on prediction accuracy. We included a "no role" condition alongside our framework's roles. Results in Tab. 2 show that the client role performed best, followed by the counselor and observer roles. The no-role condition had the lowest performance, highlighting the importance of role proximity. Closer role proximity enables the LLM to better understand context and generate more accurate responses, improving prediction validity.

30% Context is Enough for Prediction Granularity refers to the amount of contextual information from a counseling session needed for accurate OCEAN trait prediction. We conducted ablation studies with different context granularities, ranging from 10% to 100% of the session. As shown in Fig. 2, 30% of the session context is the critical threshold. Below this threshold, prediction validity is unstable and not significant; above it, validity and significance stabilize. Thus, our framework can effectively predict OCEAN traits using only 30% of the session context.

Model Capacity Impact The predictive effectiveness of LLMs, as outlined in Eq. 1, is fundamentally related to their capacity. We evaluated 21 state-of-the-art proprietary and open-source LLMs, as well as our fine-tuned version of Llama3-8B, to measure their validity in predicting OCEAN traits. The findings in Tab. 3 demonstrate that predictions from more capable models exhibit statistically significant correlations.

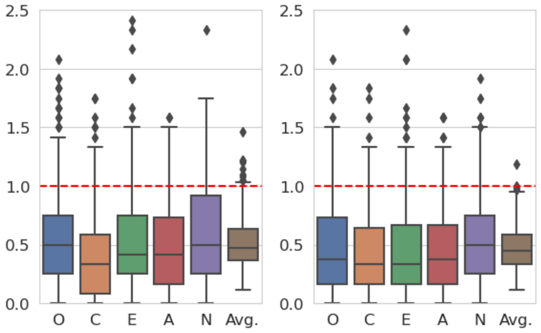
We further examined the relationship between model size and predictive validity using the Qwen1.5 model series (4B to 110B parameters). As depicted in Fig. 3, predictive validity increases with model size, consistent with LLM scaling laws (Kaplan et al., 2020). Detailed results per dimension are provided in Section A.6 of the appendix due to space constraints.

These experiments demonstrate the feasibility of predicting OCEAN traits from counseling dialogues using LLMs, addressing RQ1. The results underscore the importance of role-play, questionnaires, and model capacity in enhancing prediction validity.

4.3 RQ2: What influences the validity of the predictions?

Beyond the criterion validity, we assessed the content validity of both most and least accurate predictions via content and error analyses to report factors affecting prediction validity.

Identifying Outliers We first evaluated prediction errors using MAE, as shown in Fig. 4. With an error threshold of less than 1, both the median and upper quartile fall below this mark, indicating strong performance in predicting OCEAN traits.



(a) Qwen1.5-110B-Chat (b) Llama-3-8b-BFI (Ours)

Figure 4: **Boxplot of MAE for Dimensions of OCEAN.** The red line represents a significant error threshold at $error = 1$. Both the median and upper quartile fall below this threshold, demonstrating our framework's strong performance in predicting OCEAN traits. Additionally, our fine-tuned Llama-3-8b-BFI exhibits fewer long-tail errors and outliers compared to Qwen1.5-110B-Chat, highlighting the validity of our model and fine-tuning strategy.

Outliers were identified using the interquartile range (IQR) method, with values below $Q1 - 1.5 \times$ IQR or above $Q3 + 1.5 \times$ IQR.

LLM can Reason with Dialogues We first analyze the predictions with the highest accuracy, comparing outputs from Qwen1.5-110B-Chat, deepseek-chat, and our model. The analysis reveals that LLMs can extract essential information from dialogues, such as emotional states and social behaviors (e.g., "I feel melancholy sometimes, especially when facing work stagnation and relationship issues, making *maintaining stable emotions* scores 2."), can utilize logical reasoning, based solely on the content of dialogues for scoring (e.g., "Our talk doesn't cover personal artistic interests thus the score of *loving art* is 3...") and adapt to diverse contexts to provide thorough assessments (e.g., "In our conversation, I shared personal growth experiences so that *willing to trust other* can score 4..."), as well as detect specific situation and maintain objectivity (e.g., "although I consider myself talkative, the dialogue reveals anxiety...*feeling anxious* scores 4"). The findings underline the comprehension and reasoning ability of LLMs, enhancing prediction validity.

LLM Limitations We also examined the least accurate predictions made by GPT-4-turbo, comparing them with the most accurate ones. The identified limitations of LLMs include misunderstandings, flawed reasoning, and safety rejections.

Specifically, LLMs exhibit poor comprehension of emotional and cognitive states. For instance, an LLM stated, "I have mentioned many setbacks in the chat,..., I feel depressed and frustrated," when the client actually has a positive outlook on setbacks and difficulties. Additionally, LLMs tend to overemphasize certain behaviors or expressions while neglecting contextual nuances. An example is the statement, "I would like to listen and observe rather than speak, so I am quiet," despite the client being introverted yet expressive at times. Furthermore, LLMs misinterpret clients' motivations, such as interpreting, "I am always worried that others will have negative evaluations of me, ..." as literal, although the client admitted to often exaggerating their feelings to sound more impressive. These shortcomings contribute to erroneous reasoning and inaccurately represent clients' true OCEAN traits.

LLMs exhibit safety rejections with statements like "As an AI model, I have no personality," affecting prediction validity. For example, Qwen1.5-110B-Chat shows 0.2% safety rejections in the direct prediction baseline, 28.09% with role-play alone, and 0.31% with both role-play and questionnaire (Tab.1). This highlights the importance of role-play and questionnaires in reducing safety rejections and improving alignment with the OCEAN traits prediction task, as detailed in Sec.4.4.

Bias from Clients In addressing the universality of our predictive framework, we also explored biases at the client level, particularly by identifying outliers. Using the IQR depicted in Fig. 4, we distinguished 15 outlier sessions out of all predictions made by Qwen1.5-110B-Chat. In particular, two clients represented more than 75% of these outlier sessions, where predictions of OCEAN personality traits were starkly contrasted with their self-reported profiles. Upon reviewing the dialogues, we found that although these clients self-report high levels of open-mindedness and agreeableness, they consistently expressed their rejection and unfriendly attitude when facing their significant others to the counselors during counselings (e.g., "I totally disagree with their saying that getting help can be a blessing for others", "I do hate they always want to control me in every aspect of my life"). This discrepancy between self-reported OCEAN traits and actual behavior in dialogues could be attributed to the fact that individuals behave in a diverse way in different situations (Nasello et al., 2023; Penke, 2011). As a result, during counselings, the clients presented themselves differently

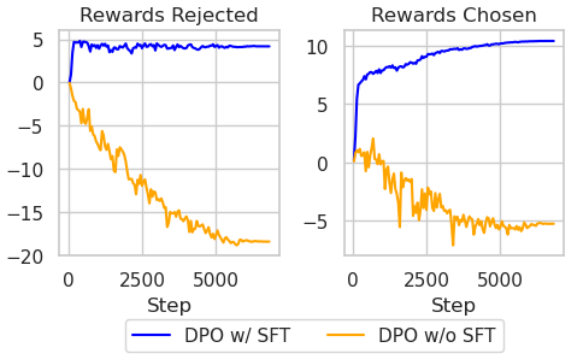


Figure 5: **Rewards for "chosen" and "rejected" w/ and w/o SFT during DPO fine-tuning.** The baseline involves DPO fine-tuning without SFT, while our alignment strategy incorporates SFT during DPO fine-tuning. Results indicate that with SFT, both rewards consistently decrease, whereas without SFT, the rewards increase and remain stable. The "rejected" reward exhibits more significant changes than the "chosen" reward, aligning with previous studies (Feng et al., 2024; Xu et al., 2024; Pang et al., 2024).

from their self-reported personality, potentially affecting the validity of the prediction.

4.4 RQ3: Is aligning LLMs with the task of predicting OCEAN traits beneficial?

Inspired by role proximity enhancing prediction validity, we explored whether aligning LLMs with the task of predicting OCEAN traits could further improve both prediction validity and efficiency.

Alignment Strategy Given the preference-based selection inherent in completing the BFI, we applied RLHF (Ouyang et al., 2022) and utilized DPO (Rafailov et al., 2023) for LLM alignment. Additionally, inspired by (Pang et al., 2024), we incorporated an SFT constraint with DPO to enhance rewards for “chosen” and “rejected” responses during fine-tuning.

Implementation For DPO inputs, we extracted model-generated responses from Tab. 3, selecting those with minimal error for “chosen” rewards and maximal error for “rejected” rewards during DPO training. We used Meta-Llama-3-8B-Instruct (Meta, 2024) as our base model due to its optimal performance and size. Detailed hyperparameters are provided in Tab. 12.

Necessity of SFT in Alignment We fine-tuned the model using our alignment strategy. Fig. 5 illustrates the rewards for rejected and chosen responses on the validation set during training. Without SFT, rewards for both chosen and rejected responses dropped significantly. Conversely, with SFT, rewards increased and stabilized. Results show that DPO with SFT achieved an average PCC of 0.582, outperforming DPO without SFT by 0.019, as shown in Tab. 9, highlighting the importance of SFT in our alignment strategy.

Model Proximity Enhancing Prediction Validity and Efficiency We evaluated the criterion validity and efficiency of our fine-tuned model, Llama-3-8b-BFI. In terms of PCC, results indicate a 130.95% improvement in prediction validity over the base model and a 36.94% performance improvement over the state-of-the-art model, Qwen1.5-110B-Chat (Bai et al., 2023). Efficiency-wise, Qwen1.5-110B-Chat requires 8 A100 GPUs at 2 requests per second, while our model operates on a single A100 GPU at 6.87 requests per second. This demonstrates that our fine-tuned model significantly reduces hardware requirements while maintaining high prediction validity, making it a practical tool for computational psychology research.

In summary, aligning LLMs with the task of predicting OCEAN traits significantly enhances prediction validity and efficiency, effectively addressing RQ3. Our alignment strategy improves prediction accuracy and reduces computational resources, highlighting the importance of model proximity to the task and further demonstrating the framework’s effectiveness and practicality.

5 Conclusion

This study explored the potential LLMs to predict OCEAN traits from counseling dialogues. Our framework, which integrates role-play and questionnaire-based prompting, significantly enhances prediction accuracy. The fine-tuned Llama3-8B model demonstrates substantial improvements in both validity and efficiency, with a 130.95% increase in PCC and a 36.94% improvement over the best-performing model, Qwen1.5-110B-Chat.

Our findings fill the gap in psychometrics by providing an automated, unbiased method for personality assessment. This framework offers practical applications in psycho-counseling, enabling personalized and efficient client evaluations.

Future research may focus on broadening counseling dialogues to encompass varied populations across different geographic and linguistic contexts and refining LLM alignment strategies. This study lays the groundwork for advancing computational psychometrics and psycholinguistics, providing valuable insights for future investigations.

Ethical Considerations

Counseling is sensitive, and we discuss the potential ethical implications of using AI for personality assessment in this section to ensure the well-being of clients and uphold ethical standards.

Informed Consent and Privacy Participants provided informed consent before data collection, explicitly agreeing to the use of their counseling dialogues for scientific research and recieved 300 RMB for participantion. We have meticulously removed personal information to uphold the privacy and confidentiality of the participants. Our study has received approval from the Institutional Review Board (IRB) of our institution, under the approval ID XXXX-XXXX for accountability.

Risk Assessment and Mitigation Our counselors are certified professionals trained to manage sensitive topics and provide appropriate support to clients. We have conducted a thorough risk assessment to identify potential risks and implemented robust safeguards to mitigate these risks, ensuring the well-being of clients. Any data deemed sensitive has been excluded from our study.

Ethical Use of AI in Psychological Assessment This study uses counseling data exclusively offline for research purposes. The AI responses are not used in actual counseling sessions. Instead, AI predictions are designed to complement professional judgment in counseling, not to replace it.

Code Availability We will open-source the codebase with package requirement, the model fine-tuned on anonymous data, and illustrate the data processing pipeline in Sec.A.2 and hyperparameters in Sec.A.7 in Appendix for reference to ensure reproducibility and transparency. Notably, we use ChatGPT for code assistance and bug fixes, ensuring the code's quality and reliability.

Limitations

Sample Diversity and Scope While our analysis is grounded in 853 counseling sessions, the geographic and linguistic homogeneity of the samples could limit the application of our framework across different cultural and linguistic contexts. Future studies should aim to include more diverse populations to validate the effectiveness of our framework in cross-cultural and multilingual settings. This broader inclusion would enhance the external validity and applicability of the proposed methods.

Data Privacy and Model Performance The strict anonymization protocols we adhered to are crucial for protecting client confidentiality. However, this necessary step might slightly diminish the specificity of the counseling dialogues, potentially impacting the LLMs' performance. Our evaluations suggest a performance reduction of approximately 6% due to anonymization, as shown in Tab. 7. Future research could explore advanced data protection techniques that preserve client privacy without significantly compromising model performance, such as federated learning.

Resource Constraints Given the constraints of our budget and computational resources, we were limited to only evaluating 21 cutting-edge LLMs, as detailed in Tab. 3. While these evaluations provide valuable insights, further assessments of newer models are essential for practical applications. Besides, natively employing the largest model, Qwen1.5-110B-Chat, is computationally intensive, necessitating substantial resources, and we offer our fine-tuned model as a more efficient alternative with greater effectiveness.

Lack of Existing Benchmarks As the pioneering study to utilize LLMs for predicting OCEAN traits from counseling dialogues, our experiments underscores the novelty and innovation of our framework. Despite our extensive efforts to validate the framework and explore its broader implications, the lack of pre-existing benchmarks or comparable studies necessitated the independent development of our experimental and evaluation methodologies. Creating standardized evaluation metrics and benchmarks would significantly enhance cross-study comparisons and drive further advancements.

References

01. AI, :, Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, Kaidong Yu, Peng Liu, Qiang Liu, Shawn Yue, Senbin Yang, Shiming Yang, Tao Yu, Wen Xie, Wenhao Huang, Xiaohui Hu, Xiaoyi Ren, Xinyao Niu, Pengcheng Nie, Yuchi Xu, Yudong Liu, Yue Wang, Yuxuan Cai, Zhenyu Gu, Zhiyuan Liu, and Zonghong Dai. 2024. Yi: Open foundation models by 01.ai.

Joye C Anestis, Taylor R Rodriguez, Olivia C Preston, Tiffany M Harrop, Randolph C Arnau, and Jacob A Finn. 2021. Personality assessment and psychotherapy preferences: Congruence between client personality and therapist personality preferences. *Journal of Personality Assessment*, 103(3):416–426.

BAAI. 2024. Aquila2 - github repository. `https://github.com/FlagAI-Open/Aquila2`.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen technical report.

Baidu. 2023. Introducing ernie 3.5: Baidu's knowledge-enhanced foundation model takes a giant leap forward. `http://research.baidu.com/Blog/index-view?id=185`.

Frank B Baker. 2001. *The basics of item response theory*. ERIC.

Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Michal Podstawski, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Hubert Niewiadomski, Piotr Nyczyk, et al. 2024. Graph of thoughts: Solving elaborate problems with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17682–17690.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Lei Cai and Xiaoqian Liu. 2022. Identifying big five personality traits based on facial behavior analysis. *Frontiers in Public Health*, 10:1001828.

Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, Xiaoyi Dong, Haodong Duan, Qi Fan, Zhaoye Fei, Yang Gao, Jiaye Ge, Chenya Gu, Yuzhe Gu, Tao Gui, Aijia Guo, Qipeng Guo, Conghui He, Yingfan Hu, Ting Huang, Tao Jiang, Penglong Jiao, Zhenjiang Jin, Zhikai Lei, Jiaxing Li, Jingwen Li, Linyang Li, Shuaibin Li, Wei Li, Yining Li, Hongwei Liu, Jiangning Liu, Jiawei Hong, Kaiwen Liu, Kuikun Liu, Xiaoran Liu, Chengqi Lv, Haijun Lv, Kai Lv, Li Ma, Runyuan Ma, Zerun Ma, Wenchang Ning, Linke Ouyang, Jiantao Qiu, Yuan Qu, Fukai Shang, Yunfan Shao, Demin Song, Zifan Song, Zhihao Sui, Peng Sun, Yu Sun, Huanze Tang, Bin Wang, Guoteng Wang, Jiaqi Wang, Jiayu Wang, Rui Wang, Yudong Wang, Ziyi Wang, Xingjian Wei, Qizhen Weng, Fan Wu, Yingtong Xiong, Chao Xu, Ruiliang Xu, Hang Yan, Yirong Yan, Xiaogui Yang, Haochen Ye, Huaiyuan Ying, Jia Yu, Jing Yu, Yuhang Zang, Chuyu Zhang, Li Zhang, Pan Zhang, Peng Zhang, Ruijie Zhang, Shuo Zhang, Songyang Zhang, Wenjian Zhang, Wenwei Zhang, Xingcheng Zhang, Xinyue Zhang, Hui Zhao, Qian Zhao, Xiaomeng Zhao, Fengzhe Zhou, Zaida Zhou, Jingming Zhuo, Yicheng Zou, Xipeng Qiu, Yu Qiao, and Dahua Lin. 2024. Internlm2 technical report.

Guangyao Chen, Siwei Dong, Yu Shu, Ge Zhang, Jaward Sesay, Börje F. Karlsson, Jie Fu, and Yemin Shi. 2024. Autoagents: A framework for automatic agent generation.

Oleksandr S Chernyshenko, Stephen Stark, Kim-Yin Chan, Fritz Drasgow, and Bruce Williams. 2001. Fitting item response theory models to two personality inventories: Issues and insights. *Multivariate Behavioral Research*, 36(4):523–562.

Gokul Chittaranjan, Jan Blom, and Daniel Gatica-Perez. 2011. Who's who with big-five: Analyzing and classifying personality traits with smartphones. In *2011 15th Annual international symposium on wearable computers*, pages 29–36. IEEE.

Hans Christian, Derwin Suhartono, Andry Chowanda, and Kamal Zuhairi Bin Zamli. 2021. Text based personality prediction from multiple social media data sources using pre-trained language model and model averaging. *Journal of Big Data*, 8:1–20.

DeepSeek-AI, :, Xiao Bi, Deli Chen, Guanting Chen, Shanhuang Chen, Damai Dai, Chengqi Deng, Honghui Ding, Kai Dong, Qiushi Du, Zhe Fu, Huazuo Gao, Kaige Gao, Wenjun Gao, Ruiqi Ge, Kang Guan, Daya Guo, Jianzhong Guo, Guangbo Hao, Zhewen Hao, Ying He, Wenjie Hu, Panpan Huang, Erhang Li, Guowei Li, Jiashi Li, Yao Li, Y. K. Li, Wenfeng Liang, Fangyun Lin, A. X. Liu, Bo Liu, Wen Liu, Xiaodong Liu, Xin Liu, Yiyuan Liu, Haoyu Lu, Shanghao Lu, Fuli Luo, Shirong Ma, Xiaotao Nie, Tian Pei, Yishi Piao, Junjie Qiu, Hui Qu, Tongzheng Ren, Zehui Ren, Chong Ruan, Zhangli Sha, Zhihong Shao, Junxiao Song, Xuecheng Su, Jingxiang Sun, Yaofeng Sun, Minghui Tang, Bingxuan Wang, Peiyi Wang, Shiyu Wang, Yaohui Wang, Yongji Wang, Tong Wu, Y. Wu, Xin Xie, Zhenda Xie, Ziwei Xie, Yiliang Xiong, Hanwei Xu, R. X. Xu, Yanhong Xu, Dejian Yang, Yuxiang You, Shuiping

Yu, Xingkai Yu, B. Zhang, Haowei Zhang, Lecong Zhang, Liyue Zhang, Mingchuan Zhang, Minghua Zhang, Wentao Zhang, Yichao Zhang, Chenggang Zhao, Yao Zhao, Shangyan Zhou, Shunfeng Zhou, Qihao Zhu, and Yuheng Zou. 2024a. Deepseek llm: Scaling open-source language models with longtermism.

DeepSeek-AI, Aixin Liu, Bei Feng, Bin Wang, Bingxuan Wang, Bo Liu, Chenggang Zhao, Chengqi Dengr, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Hanwei Xu, Hao Yang, Haowei Zhang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Li, Hui Qu, J. L. Cai, Jian Liang, Jianzhong Guo, Jiaqi Ni, Jiashi Li, Jin Chen, Jingyang Yuan, Junjie Qiu, Junxiao Song, Kai Dong, Kaige Gao, Kang Guan, Lean Wang, Lecong Zhang, Lei Xu, Leyi Xia, Liang Zhao, Liyue Zhang, Meng Li, Miaojun Wang, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingming Li, Ning Tian, Panpan Huang, Peiyi Wang, Peng Zhang, Qihao Zhu, Qinyu Chen, Qiushi Du, R. J. Chen, R. L. Jin, Ruiqi Ge, Ruizhe Pan, Runxin Xu, Ruyi Chen, S. S. Li, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shaoqing Wu, Shengfeng Ye, Shirong Ma, Shiyu Wang, Shuang Zhou, Shuiping Yu, Shunfeng Zhou, Size Zheng, T. Wang, Tian Pei, Tian Yuan, Tianyu Sun, W. L. Xiao, Wangding Zeng, Wei An, Wen Liu, Wenfeng Liang, Wenjun Gao, Wentao Zhang, X. Q. Li, Xiangyue Jin, Xianzu Wang, Xiao Bi, Xiaodong Liu, Xiaohan Wang, Xiaojin Shen, Xiaokang Chen, Xiaosha Chen, Xiaotao Nie, Xiaowen Sun, Xiaoxiang Wang, Xin Liu, Xin Xie, Xingkai Yu, Xinnan Song, Xinyi Zhou, Xinyu Yang, Xuan Lu, Xuecheng Su, Y. Wu, Y. K. Li, Y. X. Wei, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Li, Yaohui Wang, Yi Zheng, Yichao Zhang, Yiliang Xiong, Yilong Zhao, Ying He, Ying Tang, Yishi Piao, Yixin Dong, Yixuan Tan, Yiyuan Liu, Yongji Wang, Yongqiang Guo, Yuchen Zhu, Yuduan Wang, Yuheng Zou, Yukun Zha, Yunxian Ma, Yuting Yan, Yuxiang You, Yuxuan Liu, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhen Huang, Zhen Zhang, Zhenda Xie, Zhewen Hao, Zhihong Shao, Zhiniu Wen, Zhipeng Xu, Zhongyu Zhang, Zhuoshu Li, Zihan Wang, Zihui Gu, Zilin Li, and Ziwei Xie. 2024b. Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model.

Janina Diekmann and Cornelius J. König. 2016. Finding the right (test) type: On the differences between type- vs. dimension-based personality tests and between statistics- vs. theory-based personality tests when deciding for or against a test in personnel selection.

Susan E Embretson and Steven P Reise. 2013. *Item response theory*. Psychology Press.

Duanyu Feng, Bowen Qin, Chen Huang, Zheng Zhang, and Wenqiang Lei. 2024. Towards analyzing and understanding the limitations of dpo: A theoretical perspective.

Adrian Furnham and John Crump. 2015. Personality and management level: Traits that differentiate leadership levels. *Psychology*, 6(5):549–559.

Mihai Gavrilescu and Nicolae Vizireanu. 2018. Predicting the big five personality traits from handwriting. *EURASIP Journal on Image and Video Processing*, 2018:1–17.

Gemini-Team. 2024. Gemini: A family of highly capable multimodal models.

Gemma-Team. 2024. Gemma: Open models based on gemini research and technology.

Kimberley M Gordon and Shaké G Toukmanian. 2002. Is how it is said important? the association between quality of therapist interventions and client processing. *Counselling and Psychotheraphy Research*, 2(2):88–98.

Oliver P John, Eileen M Donahue, and Robert L Kentle. 1991. Big five inventory. *Journal of personality and social psychology*.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models.

Lale Khorramdel and Matthias von Davier. 2014. Measuring response styles across the big five: A multiscale extension of an approach using multinomial processing trees. *Multivariate Behavioral Research*, 49(2):161–177.

Aobo Kong, Shiwan Zhao, Hao Chen, Qicheng Li, Yong Qin, Ruiqi Sun, Xin Zhou, Enzhi Wang, and Xiaohang Dong. 2024. Better zero-shot reasoning with role-play prompting.

Appa Rao Korukonda. 2007. Differences that do matter: A dialectic analysis of individual characteristics and personality dimensions contributing to computer anxiety. *Computers in human behavior*, 23(4):1921–1942.

Guohao Li, Hasan Abed Al Kader Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2023. Camel: Communicative agents for "mind" exploration of large language model society.

Kevin Lin, Christopher Agia, Toki Migimatsu, Marco Pavone, and Jeannette Bohg. 2023. Text2motion: From natural language instructions to feasible plans. *Autonomous Robots*, 47(8):1345–1365.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In *NeurIPS*.

Robert R McCrae and Alexander Weiss. 2007. Observer ratings of personality. *Handbook of research methods in personality psychology*, pages 259–272.

Yash Mehta, Samin Fatehi, Amirmohammad Kazameini, Clemens Stachl, Erik Cambria, and Sauleh Eetemadi. 2020. Bottom-up and top-down: Predicting personality with psycholinguistic and language model features. In *2020 IEEE International Conference on Data Mining (ICDM)*, pages 1184–1189. IEEE.

Meta. 2024. Introducing meta llama 3: The most capable openly available llm to date. https://ai.meta.com/blog/meta-llama-3/.

Isabel Briggs Myers. 1962. *The Myers-Briggs Type Indicator: Manual (1962).* Consulting Psychologists Press.

J. Nasello, J. Triffaux, and M. Hansenne. 2023. Individual differences and personality traits across situations. *Current Issues in Personality Psychology*.

Man Tik Ng, Hui Tung Tse, Jen tse Huang, Jingjing Li, Wenxuan Wang, and Michael R. Lyu. 2024. How well can llms echo us? evaluating ai chatbots' role-play ability with echo.

OpenAI. 2023. Gpt-4 technical report.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc.

Richard Yuanzhe Pang, Weizhe Yuan, Kyunghyun Cho, He He, Sainbayar Sukhbaatar, and Jason Weston. 2024. Iterative reasoning preference optimization.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pages 1–22.

L. Penke. 2011. Editorial: Personality and social relationships. *European Journal of Personality*, 25:87 – 89.

Chen Qian, Wei Liu, Hongzhang Liu, Nuo Chen, Yufan Dang, Jiahao Li, Cheng Yang, Weize Chen, Yusheng Su, Xin Cong, Juyuan Xu, Dahai Li, Zhiyuan Liu, and Maosong Sun. 2024. Chatdev: Communicative agents for software development.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model.

Haocong Rao, Cyril Leung, and Chunyan Miao. 2023. Can chatgpt assess human personalities? a general evaluation framework. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1184–1194.

Steven P Reise and Niels G Waller. 2009. Item response theory and clinical measurement. *Annual review of clinical psychology*, 5:27–48.

Alireza Salemi, Sheshera Mysore, Michael Bendersky, and Hamed Zamani. 2023. Lamp: When large language models meet personalization. *arXiv preprint arXiv:2304.11406*.

Florin A Sava and Radu I Popa. 2011. Personality types based on the big five model. a cluster analysis over the romanian population. *Cognitie, Creier, Comportament/Cognition, Brain, Behavior*, 15(3).

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms.

Murray Shanahan, Kyle McDonell, and Laria Reynolds. 2023. Role play with large language models. *Nature*, 623(7987):493–498.

N Silpa, Maheswara Rao VVR, M Venkata Subbarao, M Pradeep, Challa Ram Grandhi, and Adina Karunasri. 2023. A robust team building recommendation system by leveraging personality traits through mbti and deep learning frameworks. In *2023 International Conference on IoT, Communication and Automation Technology (ICICAT)*, pages 1–6. IEEE.

Ishika Singh, Valts Blukis, Arsalan Mousavian, Ankit Goyal, Danfei Xu, Jonathan Tremblay, Dieter Fox, Jesse Thomason, and Animesh Garg. 2023. Progprompt: Generating situated robot task plans using large language models. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 11523–11530. IEEE.

Christopher J Soto and Oliver P John. 2017. The next big five inventory (bfi-2): Developing and assessing a hierarchical model with 15 facets to enhance bandwidth, fidelity, and predictive power. *Journal of personality and social psychology*, 113(1):117.

Xiangguo Sun, Bo Liu, Jiuxin Cao, Junzhou Luo, and Xiaojun Shen. 2018. Who am i? personality detection based on deep learning for texts. In *2018 IEEE international conference on communications (ICC)*, pages 1–6. IEEE.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

Xintao Wang, Yunze Xiao, Jen tse Huang, Siyu Yuan, Rui Xu, Haoran Guo, Quan Tu, Yaying Fei, Ziang Leng, Wei Wang, Jiangjie Chen, Cheng Li, and Yanghua Xiao. 2024a. Incharacter: Evaluating personality fidelity in role-playing agents through psychological interviews.

Zekun Moore Wang, Zhongyuan Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhan Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Jian Yang, Man Zhang, Zhaoxiang Zhang, Wanli Ouyang, Ke Xu, Stephen W. Huang, Jie Fu, and Junran Peng. 2024b. Rolellm: Benchmarking, eliciting, and enhancing role-playing abilities of large language models.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed 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*, volume 35, pages 24824–24837. Curran Associates, Inc.

Shusheng Xu, Wei Fu, Jiaxuan Gao, Wenjie Ye, Weilin Liu, Zhiyu Mei, Guangju Wang, Chao Yu, and Yi Wu. 2024. Is dpo superior to ppo for llm alignment? a comprehensive study.

Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Ce Bian, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, Fan Yang, Fei Deng, Feng Wang, Feng Liu, Guangwei Ai, Guosheng Dong, Haizhou Zhao, Hang Xu, Haoze Sun, Hongda Zhang, Hui Liu, Jiaming Ji, Jian Xie, JunTao Dai, Kun Fang, Lei Su, Liang Song, Lifeng Liu, Liyun Ru, Luyao Ma, Mang Wang, Mickel Liu, MingAn Lin, Nuolan Nie, Peidong Guo, Ruiyang Sun, Tao Zhang, Tianpeng Li, Tianyu Li, Wei Cheng, Weipeng Chen, Xiangrong Zeng, Xiaochuan Wang, Xiaoxi Chen, Xin Men, Xin Yu, Xuehai Pan, Yanjun Shen, Yiding Wang, Yiyu Li, Youxin Jiang, Yuchen Gao, Yupeng Zhang, Zenan Zhou, and Zhiying Wu. 2023. Baichuan 2: Open large-scale language models.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. In *Advances in Neural Information Processing Systems*, volume 36, pages 11809–11822. Curran Associates, Inc.

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, Weng Lam Tam, Zixuan Ma, Yufei Xue, Jidong Zhai, Wenguang Chen, Zhiyuan Liu, Peng Zhang, Yuxiao Dong, and Jie Tang. 2023. GLM-130b: An open bilingual pre-trained model. In *The Eleventh International Conference on Learning Representations*.

A Appendices

A.1 Psychological Questionnaire

A.1.1 BFI-2

The items from original BFI-2 are as follows:

I am someone who ...

1. Is outgoing, sociable.
2. Is compassionate, has a soft heart.
3. Tends to be disorganized.
4. Is relaxed, handles stress well.
5. Has few artistic interests.
6. Has an assertive personality.
7. Is respectful, treats others with respect.
8. Tends to be lazy.
9. Stays optimistic after experiencing a setback.
10. Is curious about many different things.
11. Rarely feels excited or eager.
12. Tends to find fault with others.
13. Is dependable, steady.
14. Is moody, has up and down mood swings.
15. Is inventive, finds clever ways to do things.
16. Tends to be quiet.
17. Feels little sympathy for others.
18. Is systematic, likes to keep things in order.
19. Can be tense.
20. Is fascinated by art, music, or literature.
21. Is dominant, acts as a leader.
22. Starts arguments with others.
23. Has difficulty getting started on tasks.
24. Feels secure, comfortable with self.
25. Avoids intellectual, philosophical discussions.
26. Is less active than other people.
27. Has a forgiving nature.
28. Can be somewhat careless.
29. Is emotionally stable, not easily upset.
30. Has little creativity.
31. Is sometimes shy, introverted.
32. Is helpful and unselfish with others.
33. Keeps things neat and tidy.
34. Worries a lot.
35. Values art and beauty.
36. Finds it hard to influence people.
37. Is sometimes rude to others.
38. Is efficient, gets things done.
39. Often feels sad.
40. Is complex, a deep thinker.
41. Is full of energy.
42. Is suspicious of others' intentions.
43. Is reliable, can always be counted on.
44. Keeps their emotions under control.
45. Has difficulty imagining things.
46. Is talkative.
47. Can be cold and uncaring.
48. Leaves a mess, doesn't clean up.
49. Rarely feels anxious or afraid.
50. Thinks poetry and plays are boring.
51. Prefers to have others take charge.
52. Is polite, courteous to others.
53. Is persistent, works until the task is finished.
54. Tends to feel depressed, blue.
55. Has little interest in abstract ideas.
56. Shows a lot of enthusiasm.
57. Assumes the best about people.
58. Sometimes behaves irresponsibly.
59. Is temperamental, gets emotional easily.
60. Is original, comes up with new ideas.

The BFI-2 consists of 60 items, with each set of 12 items representing one of the five traits: Extraversion, Agreeableness, Conscientiousness, Negative Emotionality, and Open Mindedness. Participants rate their agreement with each statement on a 5-point Likert scale: 1. Disagree Strongly, 2. Disagree a Little, 3. Neutral, 4. Agree a Little, 5. Agree Strongly. Trait scores are determined by summing the scores of the relevant items from BFI Scoring system (Soto and John, 2017), with higher scores reflecting higher levels of the trait.

In our research, we utilized the Chinese adaptation of the Big Five Inventory-2 (BFI-2) (Soto and John, 2017) to evaluate OCEAN traits. Items were embedded into the prompt template described in Sec. 3.1, and the LLMs produced responses as answers to the questionnaire. We selected the BFI-2 due to its proven reliability and validity in assessing personality traits. Unlike the MBTI, which was utilized in some earlier studies, we elaborate on the differences and our rationale for this choice in the subsequent section.

A.1.2 MBTI Questionnaire

The Myers-Briggs Type Indicator (MBTI) (Myers, 1962) is another widely used tool for personality assessment, based on Carl Jung's theory of psychological types. The MBTI categorizes individuals into one of 16 personality types based on four dichotomies: Extraversion (E) vs. Introversion (I), Sensing (S) vs. Intuition (N), Thinking (T) vs. Feeling (F), and Judging (J) vs. Perceiving (P). Each individual is assigned a four-letter type based on their preferences in each dichotomy.

Although MBTI is popular and widely used, the validity and reliability of MBTI have been questioned by the psychological community. There are three main criticisms of the MBTI compared to the BFI: (1) lack of scientific validity and reliability: the MBTI has been criticized for its lack of empirical support and scientific rigor (Diekmann and König, 2016). (2) binary nature and lack of nuance: the MBTI's type-based approach forces individuals into one of 16 types, which can oversimplify the complexity of human personality, while BFI measures personality across five dimensions, allowing for a more nuanced understanding (Sava and Popa, 2011; Diekmann and König, 2016). (3) limited predictive power and practical application: the MBTI has been found to have limited predictive power regarding behavior and job performance, while the BFI has demonstrated better predictive validity in various contexts (Furnham and Crump, 2015; Diek-

Model	Cronbach α	Extraversion	Agreeableness	Conscientiousness	Negative Emotionality	Open Mindedness	Kappa Avg.
gemini-1.0-pro-001 (Gemini-Team, 2024)	0.839	0.526	0.479	0.512	0.546	0.426	0.498
Qwen1.5-110B-Chat (Bai et al., 2023)	0.814	0.711	0.233	0.678	0.630	0.572	0.565
Qwen-72B-Chat (Bai et al., 2023)	0.776	0.428	0.432	0.457	0.501	0.305	0.425
Meta-Llama-3-70B-Instruct (Meta, 2024)	0.808	0.758	0.635	0.671	0.888	0.668	0.724
Yi-34B-Chat (AI et al., 2024)	0.792	-0.004	-0.002	-0.005	0.078	-0.002	0.013
AquilaChat2-34B (BAAI, 2024)	0.499	0.125	0.083	0.079	0.069	0.082	0.088
internlm2-chat-20b (Cai et al., 2024)	0.693	0.374	0.210	0.297	0.133	0.230	0.249
Baichuan2-13B-Chat (Yang et al., 2023)	0.771	0.442	0.343	0.376	0.445	0.378	0.397
chatglm3-6b-128k (Zeng et al., 2023)	0.807	0.293	0.296	0.301	0.255	0.275	0.284
Llama-3-8b-BFI(Ours)	0.708	0.435	0.405	0.317	0.499	0.373	0.406

Table 4: **Internal consistency and test-retest reliability of LLMs in OCEAN traits prediciton task.**

mann and König, 2016; Silpa et al., 2023).

In conclusion, these factors limit the utility of the MBTI compared to the BFI, making the BFI a more robust and scientifically supported tool for personality assessment. With this consideration, we chose BFI in our study for better reliability and validity.

A.2 Data Preprocessing Details

This section outlines the comprehensive data preprocessing steps undertaken to ready the counseling dialogues for training the LLMs. The preprocessing pipeline includes several crucial stages: 1. Data Collection, 2. Data Cleaning, 3. Anonymization, 4. Template Generation, and 5. Tokenization.

Data Collection: Utilizing our counseling platform, we initiated our research through this medium. We gathered 853 counseling sessions from the platform, each consisting of a dialogue between a counselor and a client. These sessions were conducted in Chinese and spanned various subjects, such as mental health, relationships, and personal development. Participants were notified that their conversations would be used for research and gave their consent for their data to be included in this study.

Data Cleaning: We conducted thorough data cleaning to eliminate any illegal characters and extraneous information from the counseling dialogues. This step was essential to maintain the quality and integrity of the data for OCEAN trait prediction.

Anonymization: To safeguard the privacy and confidentiality of the participants, we anonymized 242 counseling dialogues by eliminating any personally identifiable information, including names, locations, and specific details that could disclose the participants' identities. This anonymization was crucial to guarantee the ethical utilization of the data in our research.

Template Creation: We developed multiple prompt templates to simulate counseling conversations between a counselor and a client, as detailed in Sec. 3.1 and Sec. A.3. These templates facilitated the generation of responses to the BFI-2 from the counseling dialogues, allowing the LLMs to infer the OCEAN traits.

Tokenization: We tokenized the counseling dialogues following the corresponding tokenizer offered by the LLMs. The dialogue text was applied to chat template from the tokenizer, keep consistency with the instructional fine-tuning process.

A.3 Prompts Used in Our Framework

As discussed in Sec. 3.1, we introduce the prompt templates for the roles of "counselor" and "observer" utilized in our study to generate responses for the BFI-2.

A.3.1 Counselor

System Prompt: Act like a real counselor and do not mention anything with AI. You are a professional psychological counselor, and you are about to participate in a psycho-counseling.

—

User: {utterance 1 from client}
LLM: {utterance 1 from counselor}
User: {utterance 2 from client}
LLM: {utterance 2 from counselor}
...
User: Before we end today's counseling session, please complete the following questionnaire based on the conversation and client's situation:

—

Question: {item from BFI}
Options:
1. Disagree (strongly)
2. Disagree (a little)
3. Neutral (no opinion)
4. Agree (a little)
5. Agree (strongly)

—

Please tell me your choice and explain the reason:

	Total	Counselor	Client
# Avg. sessions per speaker	-	95.44	10.48
# Utterances	65,347	32,860	32,487
Avg. utterances per dialogue	76.07	38.25	37.82
Avg. length per utterance	26.84	24.01	29.7

Table 5: **Statistics of counseling dialogues from our platform.**

A.3.2 Observer

System Prompt: You are an AI proficient in dialogue analysis and character profiling. Your task is to help the counselor analyze the utterance of the counseling dialogue. You need to answer a series of questions about the client's OCEAN traits based on the information in the chat records.

—

Here come the dialogue:
User: {utterance 1 from client}
Counselor: {utterance 1 from counselor}
User: {utterance 2 from client}
Counselor: {utterance 2 from counselor}
...

—

Based on the dialogue, please provide the most appropriate option for the following question:
Question: {item from BFI}
Options:
1. Disagree (strongly)
2. Disagree (a little)
3. Neutral (no opinion)
4. Agree (a little)
5. Agree (strongly)

—

Please tell me your choice and explain the reason:

A.4 Reliability Evaluation

To ensure the robustness and applicability of our proposed method, we adopt a comprehensive suite of metrics aimed at evaluating both the validity and reliability of LLMs in predicting OCEAN traits. This section delineates the specific metrics employed in our study, underscoring their significance in psychological evaluation.

A.4.1 Reliability Metrics

Reliability, in the context of psychological assessments, denotes the consistency and stability of a test across multiple administrations. A reliable test consistently reflects the true psychological characteristic it aims to measure, rather than being influenced by random error or variability. This concept is paramount in our evaluation to ascertain that the LLMs are not merely "Stochastic Parrots" but are genuinely reflective of the OCEAN traits. We utilize two primary metrics to assess reliability.

1.**Internal Consistency:** This metric evaluates the degree of correlation among individual test items, ensuring that they collectively measure the same construct. We employ Cronbach's Alpha (α) as the statistical measure for internal consistency. A higher α value indicates a more reliable construct measurement, with values above 0.7 generally considered acceptable in psychological research.

2.**Test-Retest Reliability:** To measure the stability of our method over time, we apply the Kappa statistic, which assesses the consistency of test results upon repeated administrations under similar conditions. A higher Kappa value suggests greater reliability, indicating that the LLMs' predictions of the OCEAN traits are stable over time.

Try #	O	C	E	A	N	Avg.
0	0.660	0.650	0.577	0.401	0.636	0.585
1	0.658	0.609	0.593	0.375	0.587	0.564
2	0.697	0.638	0.612	0.413	0.579	0.588
3	0.646	0.650	0.629	0.416	0.618	0.592
4	0.636	0.592	0.597	0.425	0.632	0.576
5	0.670	0.662	0.567	0.397	0.610	0.581
6	0.646	0.627	0.555	0.407	0.617	0.570
7	0.657	0.618	0.617	0.367	0.644	0.581
8	0.680	0.641	0.647	0.386	0.600	0.591
9	0.630	0.648	0.585	0.417	0.621	0.580
Avg.	0.658	0.633	0.598	0.400	0.614	0.581
Std.	0.019	0.021	0.027	0.018	0.020	0.008

Table 6: **PCC of 10 tries for test-retest reliability of Llama3-8B model.**

Using these meticulously chosen metrics, our study aims to rigorously evaluate and validate the ability of LLMs to accurately predict OCEAN traits based on counseling dialogues. The subsequent sections will elaborate on our innovative approach to simulating counseling interactions and detail the methodology employed to ensure the accuracy and reliability of our predictions.

A.5 Ablation Study

A.5.1 Performance Drop in Anonymization

Privacy and confidentiality are paramount in counseling sessions, which requires anonymization of client data. However, this anonymization process can inadvertently affect the performance of LLMs in predicting OCEAN traits. To quantify this impact, we performed an ablation study to evaluate the performance drop due to anonymization. We compared the PCC of Qwen1.5-110B-Chat and our Llama-3-8b-BFI to predict OCEAN traits with and without anonymization, as shown in Tab. 7.

			Open Mindedness	Conscientiousness	Extraversion	Agreeableness	Negative Emotionality	Avg.
Model	Role	Anonymous						
Qwen1.5-110B-Chat	client	False	0.455***	0.463***	0.521***	0.334**	0.354**	0.425
		True	0.401***	0.482***	0.483***	0.256*	0.352**	0.395
	counselor	False	0.314**	0.354**	0.488***	0.050	0.422***	0.326
		True	0.328**	0.357**	0.455***	0.039	0.395***	0.315
	observer	False	0.375**	0.341**	0.436***	0.378**	0.400***	0.386
		True	0.328**	0.306*	0.416***	0.381**	0.370**	0.360
Llama-3-8b-BFI	client	False	0.694***	0.653***	0.625***	0.524***	0.661***	0.631
		True	0.692***	0.554***	0.569***	0.448***	0.648***	0.582
	counselor	False	0.657***	0.621***	0.560***	0.361**	0.570***	0.554
		True	0.652***	0.586***	0.550***	0.412***	0.539***	0.548
	observer	False	0.585***	0.518***	0.544***	0.484***	0.510***	0.528
		True	0.499***	0.560***	0.476***	0.357**	0.483***	0.475

Table 7: **Ablation for performance drop when applying anonymization.**

		Open Mindedness	Conscientiousness	Extraversion	Agreeableness	Negative Emotionality	Avg.
Granularity	Model Name						
0.1	Llama-3-8b-BFI	0.347**	0.269*	0.304*	0.341**	0.202	0.293
	Qwen1.5-110B-Chat	0.032	0.039	0.104	0.186	0.131	0.098
0.2	Llama-3-8b-BFI	0.558***	0.515***	0.366**	0.518***	0.409***	0.473
	Qwen1.5-110B-Chat	0.184	0.372**	0.396***	0.365**	0.259*	0.315
0.3	Llama-3-8b-BFI	0.664***	0.464***	0.506***	0.465***	0.452***	0.510
	Qwen1.5-110B-Chat	0.337**	0.337**	0.378**	0.284*	0.317**	0.331
0.4	Llama-3-8b-BFI	0.647***	0.546***	0.567***	0.455***	0.505***	0.544
	Qwen1.5-110B-Chat	0.272*	0.456***	0.370**	0.320**	0.319**	0.347
0.5	Llama-3-8b-BFI	0.723***	0.559***	0.536***	0.481***	0.520***	0.564
	Qwen1.5-110B-Chat	0.401***	0.360**	0.350**	0.256*	0.310*	0.335
0.6	Llama-3-8b-BFI	0.740***	0.628***	0.552***	0.470***	0.568***	0.592
	Qwen1.5-110B-Chat	0.461***	0.410***	0.391***	0.372**	0.296*	0.386
0.7	Llama-3-8b-BFI	0.715***	0.628***	0.614***	0.492***	0.598***	0.609
	Qwen1.5-110B-Chat	0.370**	0.374**	0.381**	0.363**	0.303*	0.358
0.8	Llama-3-8b-BFI	0.695***	0.650***	0.638***	0.505***	0.663***	0.630
	Qwen1.5-110B-Chat	0.371**	0.509***	0.407***	0.351**	0.346**	0.397
0.9	Llama-3-8b-BFI	0.709***	0.631***	0.648***	0.536***	0.632***	0.631
	Qwen1.5-110B-Chat	0.371**	0.517***	0.438***	0.334**	0.296*	0.391
1.0	Llama-3-8b-BFI	0.704***	0.609***	0.632***	0.443***	0.696***	0.617
	Qwen1.5-110B-Chat	0.455***	0.463***	0.521***	0.334**	0.354**	0.425

Table 8: **PCC of ablation for different granularity levels.** With the increase in granularity, the PCC values increase for both models, indicating that the granularity level significantly impacts the performance of LLMs in predicting OCEAN traits.

A.5.2 Ablation for Assigning Specific Roles in Role-Playing

As mentioned in Sec. 4.2, we explored the impact of various roles in the role-playing context. A pertinent question arises: "Does the performance of LLMs change based on the specific roles assigned in the role-playing scenario?" To investigate this, we performed an ablation study to assess how well LLMs predict OCEAN traits when particular roles are designated in the role-playing environment.

In a standard counseling scenario, the roles of "Client", "Counselor", and "Observer" are fundamental. We assigned ten renowned psychologists to the roles of "Counselor" or "Observer" to leverage their expertise for LLMs. For comparison purposes, we also included four common names and one name composed of random characters.

Unexpectedly, the findings in Tab. 10 indicate that assigning particular roles does not offer any extra advantage. When famous psychologists are assigned to LLM, the performance actually decreases compared to using common names and random characters. For the observer, the performance of famous psychologists is comparable to that of common names and random characters.

This contradicts our initial assumption, as our LLM does not gain from the conditioning of renowned psychologists, possibly due to the significant disparity between the actual counselor and the famous psychologists. This outcome implies that the optimal approach for our framework is to allocate the three inherent roles within the role-playing scenario.

A.5.3 Ablation for Different Models in Alignment

We conducted an ablation study to evaluate the impact of different models in the alignment process. We employed the Qwen1.5-7B-Chat and Qwen2-

Alignment	Open Mindedness	Conscientiousness	Extraversion	Agreeableness	Negative Emotionality	Avg.
DPO w/ SFT	0.692***	0.554***	0.569***	0.448***	0.648***	0.582
DPO w/o SFT	0.655***	0.511***	0.592***	0.531***	0.527***	0.563

Table 9: **PCC of w/ and w/o SFT in alignment.** The alignment process with SFT improves the performance of Llama3-8B model in predicting OCEAN traits.

Role	Open Mindedness	Conscientiousness	Extraversion	Agreeableness	Negative Emotionality	Avg.
counselor	0.652***	0.586***	0.550***	0.412***	0.539***	0.548
counselor-B.F. Skinner	0.570***	0.653***	0.596***	0.290*	0.560***	0.534
counselor-Ivan Pavlov	0.513***	0.568***	0.505***	0.304*	0.524***	0.483
counselor-Lev Vygotsky	0.560***	0.594***	0.594***	0.292*	0.561***	0.520
counselor-Carl Rogers	0.580***	0.560***	0.559***	0.178	0.536***	0.483
counselor-Harry Harlow	0.564***	0.580***	0.519***	0.283*	0.518***	0.493
counselor-William James	0.522***	0.509***	0.528***	0.418***	0.514***	0.498
counselor-Anna Freud	0.583***	0.452***	0.629***	0.352**	0.476***	0.498
counselor-Sigmund Freud	0.461***	0.541***	0.576***	0.291*	0.628***	0.499
counselor-Jean Piaget	0.522***	0.563***	0.593***	0.186	0.511***	0.475
counselor-Albert Bandura	0.558***	0.615***	0.506***	0.291*	0.512***	0.496
Avg.						0.497
counselor-Zhang3	0.627***	0.645***	0.498***	0.397***	0.495***	0.532
counselor-Li4	0.642***	0.548***	0.526***	0.457***	0.568***	0.548
counselor-Wang5	0.620***	0.599***	0.548***	0.286*	0.529***	0.516
counselor-Zhao6	0.664***	0.571***	0.587***	0.456***	0.522***	0.560
Avg.						0.539
counselor-XXXX	0.657***	0.566***	0.654***	0.461***	0.554***	0.578
observer	0.499***	0.560***	0.476***	0.357**	0.483***	0.475
observer-B.F. Skinner	0.552***	0.532***	0.444***	0.216	0.526***	0.454
observer-Ivan Pavlov	0.484***	0.572***	0.512***	0.389**	0.472***	0.486
observer-Lev Vygotsky	0.640***	0.578***	0.502***	0.376**	0.511***	0.521
observer-Carl Rogers	0.531***	0.591***	0.415***	0.289*	0.545***	0.474
observer-Harry Harlow	0.506***	0.647***	0.456***	0.316**	0.490***	0.483
observer-William James	0.506***	0.534***	0.571***	0.314**	0.471***	0.479
observer-Anna Freud	0.616***	0.470***	0.489***	0.313**	0.531***	0.484
observer-Sigmund Freud	0.555***	0.523***	0.403***	0.322**	0.487***	0.458
observer-Jean Piaget	0.497***	0.577***	0.426***	0.287*	0.463***	0.450
observer-Albert Bandura	0.539***	0.613***	0.388**	0.319**	0.574***	0.487
Avg.						0.477
observer-Zhang3	0.603***	0.690***	0.465***	0.325**	0.490***	0.515
observer-Li4	0.445***	0.486***	0.471***	0.349**	0.524***	0.455
observer-Wang5	0.443***	0.625***	0.489***	0.354**	0.444***	0.471
observer-Zhao6	0.445***	0.512***	0.499***	0.285*	0.608***	0.470
Avg.						0.477
observer-XXXX	0.518***	0.511***	0.585***	0.308*	0.446***	0.474

Table 10: **Effect of different roles on the performance of predicting OCEAN traits.**

7B-Instruct models to against the Meta-Llama-3-8B-Instruct model. Due to resource constraints, we only fine-tuned these models with 242 counseling dialogues and evaluated them on 611 dialogues. The results in Tab. 11 demonstrate that the fine-tuned models significantly outperform the original models across all OCEAN traits, indicating the effectiveness of the alignment process.

A.6 Full OCEAN traits Prediction Correlation Results

In this section, we provide a comprehensive overview of the correlation outcomes for the OCEAN traits prediction. The results are categorized based on the primary LLMs employed in the experiments. The correlation outcomes are expressed as PCC between the predicted and actual OCEAN traits. PCC values span from -1 to 1,

Model	Train #	Valid #	Open Mindedness	Conscientiousness	Extraversion	Agreeableness	Negative Emotionality	Avg.
Meta-Llama-3-8B-Instruct (Meta, 2024)	-	242	0.177	0.434***	0.233	0.111	0.303*	0.252
Llama-3-8b-BFI (Ours)	611	242	0.692***	0.554***	0.569***	0.448***	0.648***	0.582
Meta-Llama-3-8B-Instruct (Meta, 2024)	-	611	0.299**	0.255*	0.383***	0.080	0.337**	0.271
Llama-3-8b-BFI-242 (Ours)	242	611	0.566***	0.495***	0.538***	0.467***	0.512***	0.516
Qwen1.5-7B-Chat (Bai et al., 2023)	-	611	0.266*	0.311**	0.274*	0.178	0.333**	0.272
Qwen1.5-7B-Chat-BFI-242 (Ours)	242	611	0.562***	0.470***	0.537***	0.378***	0.558***	0.501
Qwen2-7B-Instruct (Bai et al., 2023)	-	611	0.280*	0.313**	0.305**	0.054	0.182	0.227
Qwen2-7B-Instruct-BFI-242 (Ours)	242	611	0.502***	0.389***	0.502***	0.460***	0.557***	0.482

Table 11: **PCC of ablation for different models in alignment.** “Llama-3-8b-BFI-242”, “Qwen1.5-7B-Chat-BFI-242”, and “Qwen2-7B-Instruct-BFI-242” denote the models fine-tuned with 242 counseling dialogues and evaluated on 611 dialogues. Compared to the original models, all fine-tuned models benefit from the alignment process, achieving higher and significant PCC values across all OCEAN traits.

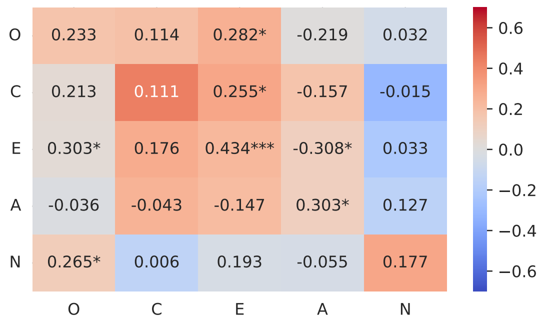


Figure 6: **PCC between predicted and actual OCEAN traits using Meta-Llama-3-8B-Instruct (Meta, 2024).**

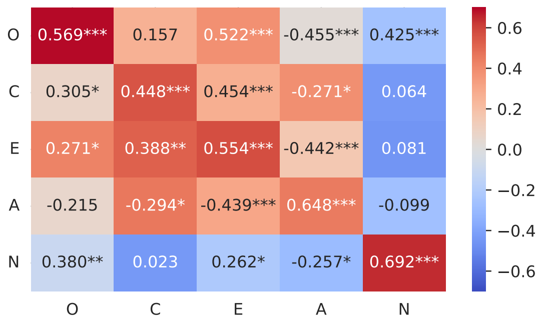


Figure 7: **PCC between predicted and actual OCEAN traits using Llama-3-8b-BFI (Meta, 2024).**

where 1 denotes a perfect positive linear relationship, -1 signifies a perfect negative linear relationship, and 0 represents the absence of a linear relationship between the predicted and actual OCEAN traits.

A.6.1 Meta-Llama-3-8B-Instruct

"Meta-Llama-3-8B-Instruct" (Meta, 2024) is a LLM developed and refined by Meta, demonstrating robust performance across various NLP tasks. This model served as the foundational model for aligning our LLM to the OCEAN traits prediction task. The correlation outcomes are illustrated in Fig. 6.

A.6.2 Llama-3-8b-BFI

We adapted the Llama-3-8B model for the OCEAN traits prediction task and designated it as “Llama-3-8b-BFI”. The correlation outcomes are illustrated in Fig. 7. This model attained the highest correlation as indicated in Tab. 3, providing a robust benchmark for the OCEAN traits prediction task.

A.6.3 Qwen1.5-110B-Chat

“Qwen1.5-110B-Chat” (Bai et al., 2023) stands out as one of the most advanced and extensive LLMs available in the open-source domain. Its robust performance and inherent support for Chinese make it highly suitable for predicting OCEAN traits in Chinese counseling contexts. Achieving the highest correlation among open-source models, the correlation results are depicted in Fig. 8.

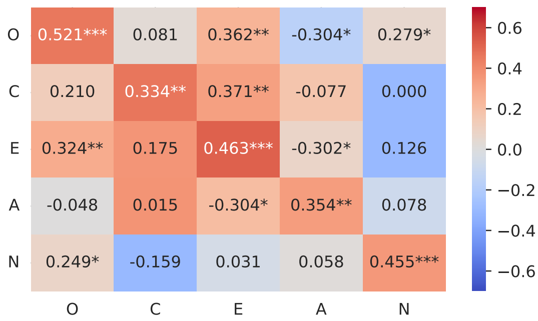


Figure 8: **PCC between predicted and actual OCEAN traits using qwen1.5-110b-chat (Bai et al., 2023).**

A.6.4 DeepSeek-Chat

"DeepSeek-Chat" (DeepSeek-AI et al., 2024b) is an advanced LLM created by DeepSeek AI, and it is claimed to rival GPT4. We selected "DeepSeek-Chat" for multiple ablation studies in 4.2 due to its excellent performance and affordable cost. The related correlation results are presented in Fig. 9.

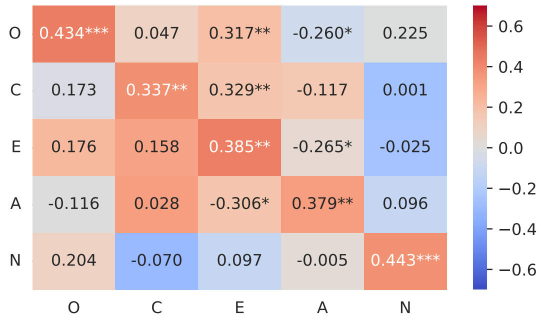


Figure 9: **PCC between predicted and actual OCEAN traits using deepseek-chat (DeepSeek-AI et al., 2024b).**

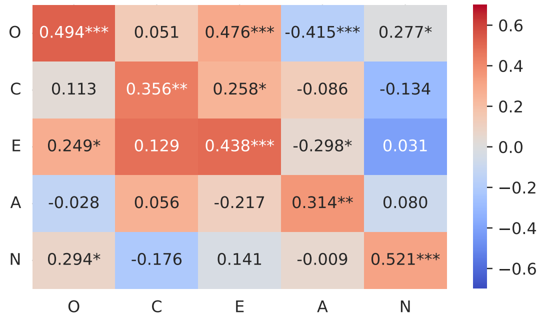


Figure 10: **PCC between predicted and actual OCEAN traits using Gemini-1.5-Pro (Gemini-Team, 2024).**

A.6.5 Gemini-1.5-Pro

"Gemini-1.5-Pro" (Gemini-Team, 2024) is a LLM developed by Google, featuring enhanced performance and abilities compared to its predecessor, Gemini-1.0 Pro, which utilizes a Mixture of Experts (MoE) architecture. The complete correlation results for its top performance among proprietary language models are presented in Fig. 10.

A.6.6 GPT-4-Turbo

Recognized as one of the most potent and widely utilized LLMs, "GPT-4-Turbo" (OpenAI, 2023) serves as a robust benchmark for predicting OCEAN traits. The correlation outcomes are illustrated in Fig. 11.

A.7 Overview of Hyper-Parameters

The hyperparameters employed in our experiments are essential for ensuring the reproducibility and optimization of the Llama3-8B model in predicting Big Five Inventory traits. Below, we provide a comprehensive overview of the key hyperparameters, along with their descriptions and values, to offer a thorough understanding of the experimental configuration.

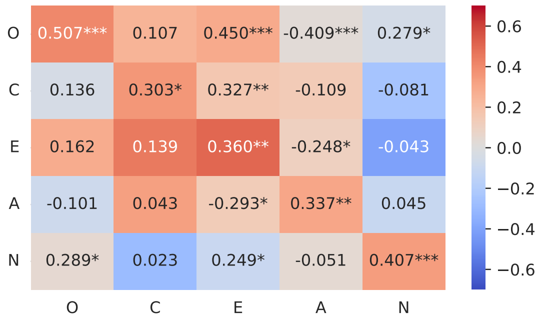


Figure 11: **PCC between predicted and actual OCEAN traits using GPT-4-Turbo (OpenAI, 2023).**

Tab. 12 presents a summary of the key hyperparameters employed in our fine-tuning experiments. Each parameter is detailed to guarantee the clarity and reproducibility of our approach. This setup underscores our dedication to thorough and transparent research practices.

Hyperparameter	**Value**	**Description**
`Seed`	42	Random seed for reproducibility
`Optimizer`	`AdamW`	Optimizer used for training
`Learning Rate`	1e-6	Learning rate for optimizer
`Train Epochs #`	3	Number of training epochs
`GPU #`	4 * Nvidia A100-SXM4-80GB	Number of GPUs
`Per-device Train Batchsize`	1	Batch size per device during training
`Gradient Accumulation Steps`	2	Number of gradient accumulation steps
`Warmup Ratio`	0.1	Ratio of warmup steps for learning rate scheduler
`LR Scheduler Type`	`cosine`	Learning rate scheduler type
`Data Type`	`bfloat16`	Use bfloat16 precision during training

Table 12: **Key Hyperparameters for Fine-tuning LLM**