Role-Play Enhanced Framework for Big Five Personality Assessment from Counseling Dialogues via 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. We introduce a novel framework that automatically predicts Big Five (OCEAN) personality traits directly from counseling dialogues by combining role-play prompting with questionnaire-based task decomposition. Our framework conditions Large Language Models (LLMs) to simulate client responses to the Big Five Inventory through counseling dialogue context, achieving significant correlations with professional assessments. Through systematic ablation studies on 853 real-world counseling sessions, we demonstrate that our role-play mechanism significantly improves prediction validity by 33.54% and reduces safety rejection rates from 28.09% to 0.31%. Our fine-tuned LLaMA3-8B model achieves a 36.94% improvement over larger models like Qwen1.5-110B while reducing computational requirements by 92.73%. Notably, our framework requires only 30% of dialogue content for reliable predictions, enabling efficient and unobtrusive personality assessment during natural counseling conversations. Our code, models, and data are publicly available to facilitate further research in computational psychometrics.[1]

1 Introduction

Understanding client personalities is fundamental to effective psychological counseling, as personality traits significantly influence treatment outcomes and guide counseling approach selection (Gordon and Toukmanian, 2002; Anestis et al., 2021). While practitioners commonly use self-report instruments like the Big Five Inventory (BFI) (John et al., 1991), these traditional assessment methods face considerable limitations. The time-consuming nature of questionnaires can disrupt counseling flow, and responses are potentially subject to social desirability or self-presentation bias (Chernyshenko et al., 2001; McCrae and Weiss, 2007; Khorramdel and von Davier, 2014), compromising assessment accuracy and treatment effectiveness. This underscores the pressing need for automated, unobtrusive, and effective methods of personality prediction in psychometrics, a challenge that modern computational approaches may be uniquely positioned to address.

Recent advances in Large Language Models (LLMs) (OpenAI, 2023; Bai et al., 2023; Gemini-Team, 2024) demonstrate remarkable capabilities in understanding human behavior through text analysis, contextual reasoning, and nuanced role-playing (Ng et al., 2024). These capabilities present a promising solution to the inherent limitations of traditional personality assessments in psycho-counseling. Specifically, LLMs could analyze the rich behavioral information naturally embedded in counseling dialogues to predict personality traits, potentially offering an unobtrusive and bias-free alternative to self-report methods. However, despite this potential to transform personality assessment in counseling settings, the application of LLMs for OCEAN trait [2] prediction from counseling dialogues remains largely unexplored, presenting a crucial gap in both computational linguistics and psychometrics research.

Research Questions Given the potential of LLMs in understanding human behavior, we investigate their capability to predict personality traits through two key hypotheses:

H1 LLMs can effectively simulate client behavior through dialogue conditioning, enabling accurate personality assessment.

H2 LLMs can extract behavioral indicators from

[1] https://anonymous.4open.science/r/BigFive-LLM-Predictor-5B41/

[2] The acronym "OCEAN" stands for 5 traits of BFI: **O**pen-Mindedness, **C**onscientiousness, **E**xtraversion, **A**greeableness, and **N**egative Emotionality. Same in the following tables.

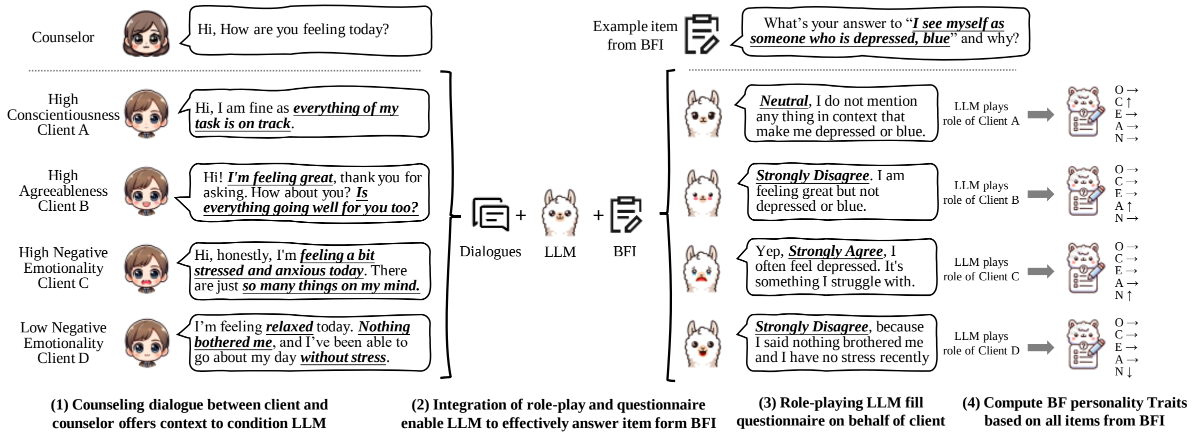

Counselor
Hi, How are you feeling today?
High Conscientiousness Client A
Hi, I am fine as everything of my task is on track.
High Agreeableness Client B
Hi! I'm feeling great, thank you for asking. How about you? Is everything going well for you too?
High Negative Emotionality Client C
Hi, honestly, 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: **An Illustration of Our Framework for Predicting OCEAN Traits from Counseling Dialogues.** Our framework consists of three integral steps: conditioning the LLM on the counseling dialogues, prompting the LLM with role-play and questionnaires, and having the LLM complete the questionnaire on behalf of the client to predict their OCEAN traits.

dialogues to make relevant personality predictions.

These hypotheses address fundamental questions about LLMs' ability to understand and analyze human personality traits in counseling contexts, with implications for both computational linguistics and psychometrics research.

Approach In this study, we test the hypothesis through three key phases:

1. Developing and validating a novel framework that combines role-playing and questionnaire prompting to enable unobtrusive personality assessment from counseling dialogues.
2. Conducting rigorous ablation studies to quantify the impact of critical factors including role alignment, dialogue context length, and model architectures on prediction accuracy.
3. Optimizing model performance through combined fine-tuning strategies, incorporating Direct Preference Optimization (DPO) and Supervised Fine-Tuning (SFT) to enhance prediction validity.

To evaluate how well that role-play LLM aligned with human behavior, we compare LLM-predicted OCEAN traits against self-reported assessments from our participant pool.

Findings We evaluated our framework using 853 real-world counseling sessions from 83 clients. Statistical analysis revealed significant correlations ($p < 0.001$) across all OCEAN traits, with Pearson Correlation Coefficients (PCC) ranging from 0.448 to 0.692. Through systematic experimentation, we identified two key factors enhancing prediction accuracy: role alignment through effective prompting and questionnaire-based trait assessment. Notably, our framework achieved reliable predictions using only 30% of session content, significantly reducing computational requirements. Our fine-tuned Llama3-8B model demonstrated a 36.94% improvement in prediction validity over the state-of-the-art Qwen1.5-110B model, while requiring only 7.27% of the computational resources, making it both more effective and more efficient for practical applications.

Contributions We advance both computational linguistics and psychometrics through three key contributions:

1. **Novel Framework for Personality Assessment:** We introduce a role-play-driven framework that automatically predicts OCEAN traits from counseling dialogues. By decomposing complex personality assessment into interpretable sub-tasks via BFI questionnaires, our approach achieves strong correlations with human assessments (PCC: 0.448-0.692) across all traits.
2. **Systematic Analysis of Role-Play Impact:** Through comprehensive experiments on 853 counseling sessions, we demonstrate that:
   - Client role enhances validity by 33.54%
   - 30% dialogue is enough for reliable predictions
   - Combined role-play and questionnaire prompting reduces safety rejection from 28.09% to 0.31%
3. **Efficient Model Optimization:** Our fine-tuned LLaMA3-8B model demonstrates both superior performance and computational efficiency:
   - Surpasses Qwen1.5-110B by 36.94% in validity
   - Reduces GPU requirements by 92.73%
   - Improves throughput by 3.4x (6.87 vs 2 req/sec)

To facilitate reproducibility and advancement of computational psychometrics, we release our code, models, and evaluation framework.

2 Related Work

Automatic Personality Assessment Recent studies have explored personality assessment using LLMs, primarily focusing on the Myers-Briggs Type Indicator (MBTI) (Myers, 1962). For instance, Rao et al. (2023) demonstrated promising results in generating MBTI-based personality assessments using ChatGPT. However, the BFI offers superior validity and reliability compared to MBTI (John et al., 1991), suggesting the need to extend LLM-based assessment to OCEAN traits.

While some researchers have attempted automatic OCEAN trait prediction using traditional approaches, such as LSTM networks (Sun et al., 2018), language model embeddings (Mehta et al., 2020), and pre-trained models (Christian et al., 2021), these studies focused primarily on essay datasets and social media posts. The application of LLMs for predicting OCEAN traits directly from counseling dialogues remains largely unexplored, despite its potential significance for psychocounseling. This research gap motivates our development of an effective framework for OCEAN trait prediction in counseling settings.

Prompting Strategies Advanced prompting strategies are crucial for maximizing LLM capabilities in personality assessment tasks. Chain-of-Though (Wei et al., 2022) and its variants enhance LLM reasoning by decomposing complex tasks into manageable steps (Singh et al., 2023; Lin et al., 2023; Yao et al., 2023; Besta et al., 2024), suggesting potential applications in personality trait prediction. Similarly, 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), with recent studies demonstrating their effectiveness in complex social tasks (Li et al., 2023; Chen et al., 2024; Wang et al., 2024b; Qian et al., 2024; Kong et al., 2024). Notably, Wang et al. (2024a) explores using role-playing agents to predict personality traits of virtual characters, indicating the potential of this approach for personality assessment. However, despite these promising advances in prompting strategies, their application to predicting OCEAN traits within counseling dialogues remains largely unexplored, presenting a crucial gap in the literature that our research aims to address.

Alignment Strategies Aligning LLMs with human is crucial for optimal performance in personality assessment tasks. Reinforcement Learning from Human Feedback (RLHF) (Ouyang et al., 2022) has shown significant improvements in LLM behavior through preference learning with Proximal Policy Optimization (PPO) (Schulman et al., 2017). To address PPO's complexity and instability, DPO (Rafailov et al., 2023) introduced a parametrized reward function approach. However, recent studies (Feng et al., 2024; Xu et al., 2024) reveal that while DPO effectively reduces dispreferred outputs, it struggles to enhance preferred response generation. Pang et al. (2024) addressed this limitation by combining negative log-likelihood loss with DPO loss. Complementing these approaches, SFT with high-quality data has proven effective for improving generation quality in successful LLMs (Touvron et al., 2023; Liu et al., 2023). Despite these advances in LLM alignment, their potential benefits for predicting OCEAN traits from counseling dialogues remain unexplored, presenting a crucial gap that our research aims to address.

3 Role-Play Enhanced Framework for OCEAN Trait Prediction

Our framework leverages role-play mechanics and questionnaire prompting to enable accurate prediction of OCEAN personality traits from counseling dialogues, comprising two key components: prompting strategy design and LLM conditioning.

3.1 Prompting Strategy Design

The core innovation of our framework lies in its structured prompting methodology that combines counseling role-play with validated psychological assessments. This approach enables robust personality trait prediction through three integrated components:

1. **Role-Based Conditioning:** We establish explicit counseling roles (client, counselor) with well-defined interaction parameters to simulate authentic counseling dialogues. This enables precise behavioral conditioning of the LLM through contextualized simulation of counseling dynamics.
2. **Context Integration:** Historical counseling sessions provide rich behavioral data, allowing extraction of personality-relevant patterns while maintaining temporal and contextual consistency. This ensures the LLM's predictions are grounded in actual counseling interactions.
3. **Structured Assessment:** We decompose complex personality prediction into discrete compo-

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:

Figure 2: **Example prompt template for the client role.** This template structures the conversation flow and questionnaire format for personality assessment.

nents using validated BFI questionnaire items. This ensures alignment between LLM outputs and established psychological metrics while enabling systematic evaluation.

A typical prompt in client's aspective is structured in Figure 2, which guides the conversation flow and questionnaire format for personality assessment. The counselor's prompt is similar to the client's, with the roles reversed to simulate the counselor's perspective.

3.2 LLM Conditioning for OCEAN trait Prediction

To formalize our approach for personality trait prediction, we frame the task as a conditional language modeling problem that maps counseling dialogue context and standardized questionnaire items to trait predictions. Formally, let x_{context} denote the historical counseling dialogues and questionnaire represent BFI items embedded in the prompt template. The prediction process can be expressed as $y_{\text{trait}} = \text{LLM}(x_{\text{context}}, \text{questionnaire})$, where y_{trait} represents the LLM's generated response containing both a numerical choice and supporting rationale for each BFI item. The numerical choices are extracted using pattern matching and aggregated following the standardized BFI scoring protocol (Soto and John, 2017) to compute the final OCEAN trait.

The efficacy of this prediction framework depends on several key factors including the model architecture, configuration parameters, and granularity of dialogue context. We systematically evaluate the impact of these factors through comprehensive experiments detailed in the Section 4.4.

4 Experiments

We conducted experiments using real-world counseling dialogues to evaluate our framework through three research questions: 1) Can LLMs predict OCEAN traits from counseling dialogues? 2) What influences prediction validity? 3) Does aligning LLMs improve prediction performance? The results validate both theoretical foundations and practical applications.

4.1 Data Collection and Preprocessing

We gathered text-based counseling conversations between professional counselors and actual clients from an online Chinese text-based psycho-counseling platform. Our study analyzed 853 counseling dialogues collected from a diverse participant pool comprising 82 adult clients and 9 professional counselors. The client group consisted of 55 females with ages ranging from 19 to 54 years (M=27.62, SD=5.94), while the counselor group included 7 females with ages ranging from 25 to 45 years (M=34.67, SD=7.45), as summarized in Table 5. To establish baseline personality profiles, all clients completed the Chinese version of BFI-2 (Soto and John, 2017) before their initial counseling sessions. We preprocessed the counseling dialogues by anonymizing all personal information and removing irrelevant content, ensuring data privacy and confidentiality.

4.2 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, Mean Absolute Error (MAE) is included for a detailed analysis of prediction errors.
2. Content Validity examines the justification behind predictions. By analyzing predictions with

Role	O	C	E	A	N
client	0.455***	0.463***	0.521***	0.334**	0.354**
counselor	0.314**	0.354**	0.488***	0.050	0.422***
observer	0.375**	0.341**	0.436***	0.378**	0.400***
no-role	0.292*	0.332**	0.391***	0.257*	0.324**

Table 1: **PCC Analysis Across Role Assignments for OCEAN Trait Prediction.** Our framework evaluation demonstrates a clear hierarchy of prediction validity across roles: client (avg. PCC=0.426) achieved optimal performance, followed by observer (0.386), counselor (0.326), and no-role conditions (0.319). The superior performance of client and counselor roles validates the importance of in-context role alignment for personality assessment. Significance levels: * ($p < 0.05$), ** ($p < 0.01$), and *** ($p < 0.001$).

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 Following established psychometric principles, reliability is evaluated and detailed in Appendix A.3 for space constraints.

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

To systematically evaluate our hypotheses about LLMs' capability to predict OCEAN traits and offer empirical evidence, we conducted controlled experiments examining different roles and configurations in counseling dialogue analysis. Our investigation focused particularly on testing H1 regarding LLMs' ability to simulate client behavior through dialogue conditioning.

Role Proximity Improves Prediction Validity To evaluate H1 regarding LLMs' ability to simulate client behavior through dialogue conditioning, we conducted controlled experiments examining different role configurations in counseling dialogue analysis. Results in Table 1 demonstrate a clear performance hierarchy, with client-role predictions achieving significantly higher correlations across all OCEAN traits ($p < 0.01$). Notably, the client role outperformed other conditions by substantial margins: 10.36% over observer, 30.67% over counselor, and 33.54% over no-role baselines. This pattern aligns with psychology research suggesting that increased role proximity enables more nuanced understanding of behavioral patterns, providing strong empirical support for our framework's role-based approach to personality assessment.

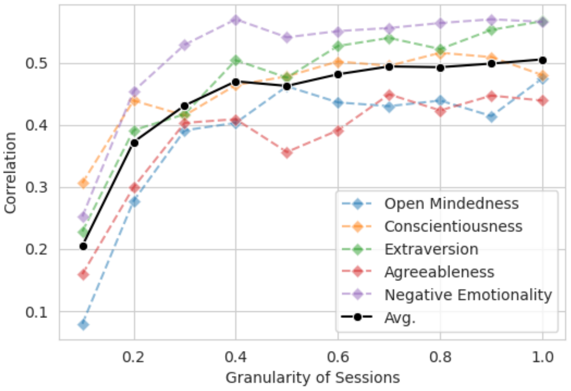


Figure 3: **PCC Changes Across Different Granularities of Dialogue Session.** The plots illustrate that the PCC increases rapidly up to 30% of the dialogue context, beyond which the increase is slower.

30% Dialogue is Enough for Reliable Prediction To determine the minimum dialogue context necessary for valid prediction, we conducted systematic ablation studies examining prediction performance across varying dialogue lengths (10-100%). Our analysis revealed a critical threshold at 30% of session content, as shown in Figure 3. Below this threshold, prediction validity was unstable with non-significant correlations ($p > 0.05$). Above 30%, both validity ($PCC > 0.4$) and statistical significance ($p < 0.01$) stabilized, with minimal additional improvements from including more context. This finding provides crucial empirical evidence that personality traits can be reliably assessed from partial dialogues, contributing to our understanding of personality manifestation in conversation.

Greater Model Capacity Enhances Prediction Model capacity emerges as a critical factor influencing prediction validity in our framework. Through systematic evaluation of 23 state-of-the-art LLMs, followed by focused analysis of the Qwen1.5 series (4B-110B parameters), we demonstrate a strong relationship between model size and prediction performance. As shown in Table 2, larger models consistently achieve higher and statistically significant correlations across all personality dimensions, indicating enhanced capability to comprehend complex psychological patterns in dialogues. This positive correlation between model size and prediction validity, visualized in Figure 4, not only validates our framework's effectiveness across different model scales but also underscores the importance of LLM capacity in personality trait prediction.

Synergy of Role-play and Questionnaires Prompting To evaluate our prompting strategies, we conduced ablation which compared four ap-

Model	O	C	E	A	N	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
Qwen2.5-72B-Chat (Bai et al., 2023)	0.406***	0.313**	0.433***	0.323**	0.410***	0.377
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 2: **PCC of Various LLMs for Predicting OCEAN traits.** Bold values indicate highest PCC per dimension. Our fine-tuned Llama-3-8b-BFI model achieved superior performance across all traits, surpassing both larger open-source models (Qwen1.5-110B-Chat) and proprietary models (Gemini-1.5-Pro). Despite its smaller size, the model demonstrated significantly higher PCC values, validating both our framework's effectiveness and our fine-tuning approach.

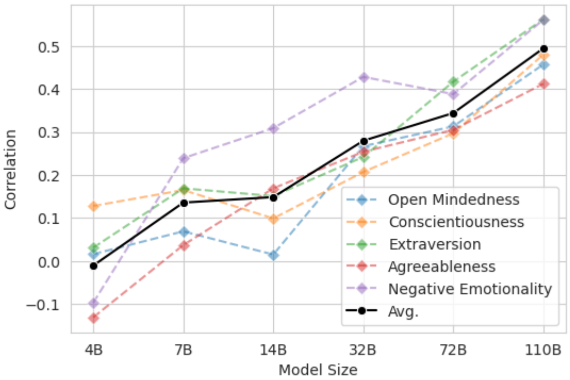


Figure 4: **Relationship between Model Size and Personality Prediction Performance.** Analysis of the Qwen1.5 series shows a positive correlation between model size and prediction accuracy. Notably, only the larger models (Qwen1.5-110B-Chat: PCC=0.425; Qwen1.5-72B-Chat: PCC=0.397) achieve statistically significant results ($p < 0.01$), suggesting that effective zero-shot personality prediction requires substantial computational resources characteristic of large language models.

Method	O	C	E	A	N
Direct Predicting	0.267*	0.167	0.190	0.091	0.142
+ *Role-Play*	0.006	0.162	-0.096	0.227	-0.028
+ *Questionnaire*	0.292*	0.332**	0.391***	0.257*	0.324**
+ *Both (Ours)*	0.455***	0.463***	0.521***	0.334**	0.354**

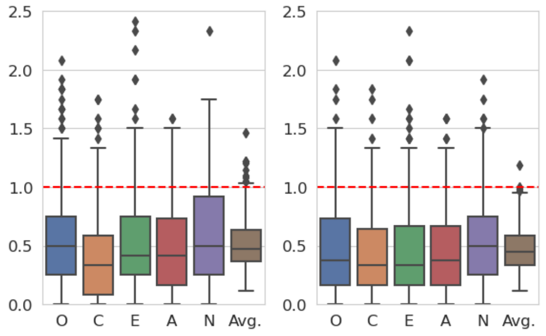
Table 3: **PCC Analysis of Method Combinations for OCEAN Trait Prediction.** Combined role-play and questionnaire prompting achieves optimal prediction (PCC=0.426), improving over questionnaire-only (0.319) by +33.54% and direct prediction (0.172) by +147.67%. Role-play-only's low performance (0.054) stems from safety rejection issues (28.09% rate, Section 4.4), which our combined approach resolves (Section 4.5).

proaches: direct prediction (baseline), role-play or questionnaire prompting only, and their combination. Results in Table 3 show that the combination yielded optimal validity (PCC=0.426) across all traits, outperform baseline by +147.67%, aligning with Item Response Theory principles (Reise and Waller, 2009; Embretson and Reise, 2013).

These experiments conclusively demonstrate the feasibility of using LLMs to predict OCEAN traits from counseling dialogues, addressing RQ1. The results underscore three key factors in enhancing prediction validity: effective role-play implementation, structured questionnaire integration, and sufficient model capacity.

4.4 RQ2: What influences the validity of the predictions?

Building upon our validation of H1, we investigated H2 through quantitative and qualitative analysis. Our dual methodology combined statistical validity metrics with content analysis of prediction cases to examine how LLMs identify behavioral patterns in dialogues. This approach revealed key factors impacting prediction accuracy, the mechanisms of behavioral inference, and core limitations. Below we examine prediction outliers, LLM reasoning capabilities, and validity constraints in personality trait prediction from counseling dialogues.



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

Figure 5: **MAE Distribution Analysis for OCEAN Trait Predictions.** The boxplots illustrate prediction error distributions, with MAE=1 (red line) representing one scale level difference, a meaningful threshold for maintaining directional accuracy. Both models demonstrate strong performance with median and upper quartile errors below this threshold. The Llama-3-8b-BFI shows superior error characteristics with fewer outliers compared to Qwen1.5-110B-Chat, validating both our model architecture and fine-tuning approach.

Analyzing Prediction Accuracy through Error Distribution To evaluate prediction accuracy, we analyzed MAE distributions (Figure 5), establishing 1.0 as a meaningful threshold representing one scale level difference. Both models demonstrate strong performance with median and upper quartile errors below this threshold, though our Llama-3-8b-BFI shows superior error characteristics with fewer outliers. Using the IQR method to identify anomalous predictions ($Q1 - 1.5 \times$IQR to $Q3 + 1.5 \times$IQR), we systematically investigated cases with significant deviations from ground truth for deeper insight into prediction limitations.

LLM Demonstrates Sophisticated Reasoning Capabilities Analysis of high-accuracy predictions reveals four key reasoning capabilities essential for valid personality assessment. First, LLMs effectively extract emotional and behavioral information from dialogues (e.g., "I feel melancholy sometimes, especially when facing work stagnation and relationship issues, suggesting *maintaining stable emotions* scores 2"). Second, they employ logical reasoning based solely on dialogue content (e.g., "Our talk doesn't cover personal artistic interests, thus the score of *loving art* is 3"). Third, LLMs demonstrate contextual adaptation through comprehensive assessments (e.g., "In our conversation, I shared personal growth experiences, indicating *willing to trust others* scores 4"). Finally, they maintain objectivity while recognizing situational nuances (e.g., "although I consider myself talkative, the dialogue reveals anxiety...*feeling anxious* scores 4"). These sophisticated reasoning capabilities significantly enhance the validity of OCEAN trait predictions, as evidenced by the strong correlations reported in Table 3.

Error Analysis To address the universality of our predictive framework, we also explored biases at the client level, particularly by identifying outliers. Using the IQR depicted in Figure 5, 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 from their self-reported personality, potentially affecting the validity of the prediction.

LLMs' Limitations Pose Challenges to Prediction Validity Analysis of GPT-4-turbo's predictions revealed three key limitations affecting OCEAN trait prediction validity: emotional misinterpretation, contextual oversimplification, and safety constraints.

First, LLMs frequently misinterpret emotional and cognitive states in counseling dialogues. For example, when a client demonstrated positive resilience, the LLM incorrectly concluded "I feel depressed and frustrated" based on mere mention of setbacks. Second, LLMs tend to oversimplify behavioral patterns, ignoring nuanced contextual cues. This was evident when an LLM characterized a selectively expressive introvert as categorically "quiet" based on limited dialogue samples. Third, LLMs sometimes misattribute client motivations, as when interpreting statements about anxiety over others' evaluations literally, despite the client's admission of intentional exaggeration for effect.

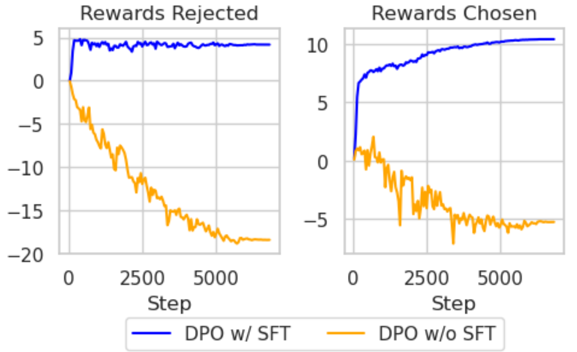


Figure 6: **DPO fine-tuning rewards with and without SFT.** Incorporating SFT during DPO fine-tuning leads to consistent reward decreases, while DPO alone shows increased and stabilized rewards. The more pronounced changes in "rejected" versus "chosen" rewards align with previous findings (Feng et al., 2024; Xu et al., 2024; Pang et al., 2024), demonstrating the effectiveness of our alignment strategy.

Additionally, safety rejection (e.g., "I am a AI, I have no personality ...") poses a significant challenge to prediction validity. Analysis of Qwen1.5-110B-Chat showed varying rejection rates: 0.2% in direct prediction, 28.09% with role-play alone, and 0.31% with combined role-play and questionnaire approaches (Table 3). These findings underscore the importance of our integrated approach in mitigating both interpretative limitations and safety rejections, as further explored in Section 4.5.

4.5 RQ3: Does aligning LLMs improve OCEAN trait prediction?

Building on our finding that role proximity enhances prediction validity, we investigated whether explicitly aligning LLMs with the OCEAN prediction task could further improve both effectiveness and efficiency. Prior work suggests that alignment through fine-tuning can help bridge the gap between pre-training objectives and downstream tasks. To evaluate this hypothesis, we divided our dataset into training and validation sets (70/30 split), yielding 611 dialogues for training and 242 for validation.

Alignment Strategy To optimize our model for personality trait prediction, we developed an alignment strategy integrating DPO (Rafailov et al., 2023) with SFT, treating self-reported traits as structured signals for robust inference rather than absolute ground truth. This approach leverages both the preference-based nature of BFI assessment and supervised learning benefits (Pang et al., 2024), using Meta-Llama-3-8B-Instruct (Meta, 2024) as our base model for its optimal efficiency-performance balance. Training data pairs were constructed by extracting model responses from our comparative analysis (Table 2), with minimal-error responses serving as "chosen" examples and maximal-error responses as "rejected" examples. The integration of SFT with DPO enhances prediction robustness while maintaining high-quality generations. Hyperparameters are detailed in Table 13.

Alignment Enhances Prediction Validity and Efficiency Our alignment strategy combines DPO with SFT to optimize model performance. As shown in Figure 6, DPO without SFT led to declining rewards for both responses, while incorporating SFT stabilized and improved rewards during training. Quantitative analysis reveals that DPO with SFT achieved a significantly higher average PCC of 0.582 compared to 0.563 without SFT ($p < 0.01$, Table 7). The aligned Llama-3-8b-BFI model demonstrates substantial improvements in both prediction validity and computational efficiency, with validation results showing a 130.95% increase in PCC over the base model ($p < 0.001$) and a 36.94% improvement over the state-of-the-art Qwen1.5-110B-Chat. Notably, while Qwen1.5-110B-Chat requires 8 A100 GPUs to process 2 requests per second, our model achieves 6.87 requests per second on a single A100 GPU. This 92.73% reduction in hardware requirements while maintaining superior prediction validity establishes our model as a practical and efficient tool for computational psychology research.

5 Conclusion

This study validates the capability of LLMs to predict OCEAN traits from counseling dialogues through an innovative framework integrating role-play and questionnaire-based prompting. Our fine-tuned Llama3-8B model achieves a 130.95% increase in prediction validity while reducing computational requirements by 92.73% compared to state-of-the-art models. The framework's ability to generate reliable predictions from just 30% of dialogue content demonstrates its practical viability for real-world counseling applications.

Our work advances both computational linguistics and psychometrics by enabling automated personality assessment, representing a significant step towards democratizing mental health support. Future research can explore cross-cultural applications, refine alignment strategies, and investigate the framework's scalability to larger datasets.

Ethical Considerations

Scope and Applicability This research advances automatic psychometric assessment framework for research purpose. While our framework demonstrates strong correlations with established measures, it requires further validation before clinical deployment. The technology should be restricted to research contexts and **not used for clinical assessment or automated debiasing** at this stage. Additional studies are needed to evaluate potential biases, establish clinical validity, and develop appropriate guidelines for responsible implementation.

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 CNY 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

While our study demonstrates significant advancements in computational psychometrics, we acknowledge certain limitations and outline our efforts to address them:

Limited Direct Benchmarks Our novel framework represents one of the first attempts to predict OCEAN traits from counseling dialogues using LLMs. While the absence of direct benchmarks reflects the innovative nature of our work, we have rigorously validated our approach through systematic ablation studies and statistical analyses. Our open-source framework provides a foundation for developing standardized evaluation metrics in this emerging field.

Dataset Considerations Our dataset of 853 counseling dialogues from 82 clients, while statistically significant for model fine-tuning, could benefit from greater scale. We have mitigated this limitation by employing nine professional counselors, implementing robust anonymization protocols, and validating results through extensive ablation studies. Our framework's strong performance on this focused dataset (PCC=0.582) suggests promising scalability to larger collections.

Ground Truth Methodology While relying on self-reported BFI scores as ground truth follows established practice in personality psychology, we acknowledge potential self-presentation effects. We addressed this through strict anonymity protocols and temporal separation between counseling and assessment. Our framework's significant correlations with these standardized measures ($p < 0.001$) demonstrate its validity within accepted psychological assessment paradigms.

Cultural Context Our current focus on Chinese-speaking participants reflects a deliberate choice to develop and validate our framework within a well-defined cultural context. We have demonstrated the framework's effectiveness through statistically significant results across all OCEAN traits. Future work can build on this foundation to explore cross-cultural applications while maintaining the robust methodology established in this study.

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. *Preprint*, arXiv:2403.04652.

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. *Preprint*, arXiv:2309.16609.

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.

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.

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. *Preprint*, arXiv:2403.17297.

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. *Preprint*, arXiv:2309.17288.

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.

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. *Preprint*, arXiv:2401.02954.

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. *Preprint*, arXiv:2405.04434.

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. *Preprint*, arXiv:2404.04626.

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

Gemini-Team. 2024. Gemini: A family of highly capable multimodal models. *Preprint*, arXiv:2312.11805.

Gemma-Team. 2024. Gemma: Open models based on gemini research and technology. *Preprint*, arXiv:2403.08295.

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*.

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. *Preprint*, arXiv:2308.07702.

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. *Preprint*, arXiv:2303.17760.

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. *Preprint*, arXiv:2404.13957.

OpenAI. 2023. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

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. *Preprint*, arXiv:2404.19733.

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. *Preprint*, arXiv:2307.07924.

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. *Preprint*, arXiv:2305.18290.

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. *Preprint*, arXiv:1707.06347.

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. *Preprint*, arXiv:2307.09288.

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. *Preprint*, arXiv:2310.17976.

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. *Preprint*, arXiv:2310.00746.

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. *Preprint*, arXiv:2404.10719.

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. *Preprint*, arXiv:2309.10305.

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 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

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.**

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 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 Section 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.

A.1.3 Justification for choosing BFI-2 over MBTI

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; Diekmann 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, along with associated ethical safeguards. The preprocessing pipeline includes: 1. Data Collection, 2. Data Cleaning, 3. Anonymization, 4. Template Generation, and 5. Tokenization.

Data Collection: We gathered 853 text-based counseling sessions through our platform, featuring dialogues between certified counselors and clients. Participant recruitment occurred through advertising, word-of-mouth, and snowball sampling, with rigorous screening in two stages:

- **Stage 1:** 1,032 initial candidates completed the Symptom Checklist-90 (SCL-90) to screen for severe psychological conditions (T scores < 2.5).

- **Stage 2:** Clinical psychologists conducted phone interviews with 148 eligible candidates, ultimately selecting 82 adult participants (55 females, ages 19-54, M = 27.62, SD = 5.94) based on mental health status and counseling suitability.

Selected participants received compensation and provided informed consent for non-profit academic use. Sessions covered diverse counseling topics (mental health, relationships, personal development) to enhance ecological validity, with text-based format ensuring accurate transcription.

Data Cleaning: We conducted thorough cleaning to eliminate illegal characters and extraneous information while preserving linguistic patterns essential for OCEAN trait prediction.

Anonymization: We removed personally identifiable information (PII) from 242 dialogues, including names, locations, and identifying details. Clients retain permanent redaction rights through platform contact, with immediate action taken on any removal request.

Ethical Safeguards: Three-layer protections ensure responsible data handling:

- **Data Access:** Sharing requires IRB approval and data use agreements for *bona fide* research

- **Counselor Expertise:** All counselors are certified professionals trained in distress management and client safety protocols

- **Client Protection:** Participants received platform resources and referral options for additional support

Template Creation: We developed prompt templates simulating counselor-client interactions (detailed in Section 3.1 and Appendix A.4), enabling LLMs to generate BFI-2 responses for OCEAN trait inference.

Tokenization: Dialogues were tokenized using each LLM's native tokenizer, maintaining consistency with instructional fine-tuning through chat template application.

A.3 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.

	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.3.1 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.4 Prompts Used in Our Framework

As discussed in Section 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.4.1 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:

A.4.2 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.5 Ablation Study

A.5.1 Further Analysis of Role-Play and Questionnaire Prompting

The effectiveness of our proposed framework hinges on the synergistic integration of role-play and questionnaire-based prompting. This subsection addresses these points through detailed analysis and additional experimental results.

Synergistic Effect of Role-Play and Questionnaires Table 3 and Table 1 in the main paper demonstrates that the Role-Play Only method, when used in isolation, yields lower performance than the Questionnaire Only approach, and even negative performance in some cases. However, this is not indicative of its ineffectiveness within our framework. Indeed, we hypothesize that the complexity of directly predicting personality traits from counseling dialogues limits the standalone efficacy of role-play. As shown in Table 3, integrating role-play with the questionnaire yields optimal prediction validity (PCC=0.582 for Llama-3-8B-BFI), outperforming questionnaire alone by a margin of +33.54% and baseline direct prediction by +147.67%, indicating a significant synergistic effect. To further investigate the mechanism behind this, we conducted additional experiments finetuning Llama-3-8B with only the questionnaire, without role-play, and summarized the results in Table 8.

As demonstrated in Table 8, the Llama-3-8B model fine-tuned without the questionnaire (Llama-3-8B-no-roleplay) has poor performance in OCEAN traits prediction, regardless of role assignment. Even the no-role condition with limited questionnaire input only achieved an average PCC of 0.102, which is only slightly better than the original Llama-3-8B-model-BFI model without any fine-tuning(0.050 in Tab.3). However, when integrated with the questionnaire and fine-tuned with our method, the same Llama-3-8B model achieve significant improvement, with average PCC of 0.476 (client role) and 0.514 (no-role), indicating that the role-play prompt is indeed effective, and the questionnaire is essential for achieving optimal performance. These results suggest that the questionnaires facilitate a task decomposition and the role-play serves to create a context for prediction in our framework.

Model Safety Rejection and the Importance of Combined Approach Furthermore, the low per-

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 7: **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.

Model	Questionnaire	Role	O	C	E	A	N	Avg.
Llama-3-8B-no-roleplay	No	client	-0.035	0.068	0.055	-0.119	0.034	0.001
		no-role	-0.004	0.299*	0.272*	0.136	-0.191	0.102
	Yes	client	0.484***	0.556***	0.446***	0.301*	0.595***	0.476
		no-role	0.656***	0.449***	0.561***	0.359**	0.547***	0.514

Table 8: **PCC of Llama-3-8B-no-roleplay With and Without Questionnaire.** We evaluated Llama-3-8B with and without role-playing components to test the synergy between the role-play and questionnaire. The lower part shows the result of using questionnaires, the higher performance indicated the importance of questionnaires in our framework.

formance of the Role-Play Only method is partly due to safety rejection. As discussed in Section 4.4 of the main paper, our analysis reveals that role-play prompts alone result in a 28.09% safety rejection rate, where the LLM refuses to respond as instructed. However, when used in conjunction with the questionnaire approach, this rate significantly reduces to 0.31%, as discussed in the main paper, Table 3. This reduction in safety rejection suggests that the questionnaire not only facilitates task decomposition but also helps maintain consistency in the model's behavior during complex interactions.

Framework Effectiveness and Fine-Tuning While it's true that fine-tuning alone provides substantial improvement over baseline LLM performance, it's essential to recognize that our framework is the methodology for task alignment with proper prompts, which is necessary to have effective fine-tuning. Without the framework to guide the model with detailed role-play prompts and a series of BFI items, the performance of fine-tuned LLM is limited to the general capabilities of the model. The role-play framework and questionnaire methodology serve as a bridge between generic LLMs and specific tasks in psychometrics, enabling the models to understand the underlying psychological constructs, and to form a systematic workflow. This framework not only enhances overall performance but provides a generalized framework for different models. As shown in Table 3, our framework enables different models to benefit from the workflow (e.g., Qwen1.5-110B-Chat and deepseek-chat). This aspect underscores the complementary nature of our framework and the fine-tuning approach, as both are necessary for achieving the reported high performance.

In conclusion, although the isolated Role-Play method has suboptimal performance, it is not ineffective within our framework. It works synergistically with the questionnaire to improve the performance in OCEAN traits prediction. In fact, the combination of role-play prompting with questionnaire is the primary driver for high performance in our framework.

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

As mentioned in Section 4.3, 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 Table 11 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.

Role	Model	Open Mindedness	Conscientiousness	Extraversion	Agreeableness	Negative Emotionality	Avg.
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 9: **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.

Method	Model	Open Mindedness	Conscientiousness	Extraversion	Agreeableness	Negative Emotionality	Avg.
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 10: **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 *** ($p < 0.001$).

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 Qwen2-7B-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 Table 12 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, 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 Figure 7.

	Open Mindedness	Conscientiousness	Extraversion	Agreeableness	Negative Emotionality	Avg.
Role						
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 11: **Effect of different roles on the performance of predicting OCEAN traits.**

	Train #	Valid #	Open Mindedness	Conscientiousness	Extraversion	Agreeableness	Negative Emotionality	Avg.
Model								
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 12: **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.

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 Figure 8. This model attained the highest correlation as indicated in Table 2, providing a robust benchmark for the OCEAN traits prediction task.

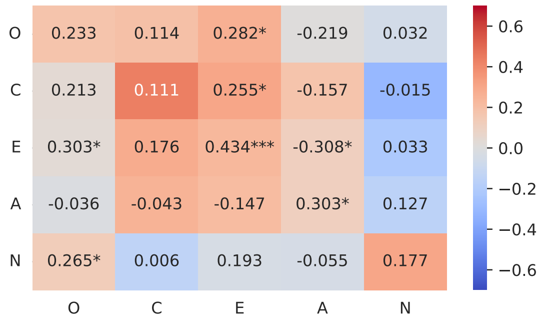


Figure 7: **PCC between predicted and actual OCEAN traits using Meta-Llama-3-8B-Instruct (Meta, 2024).**

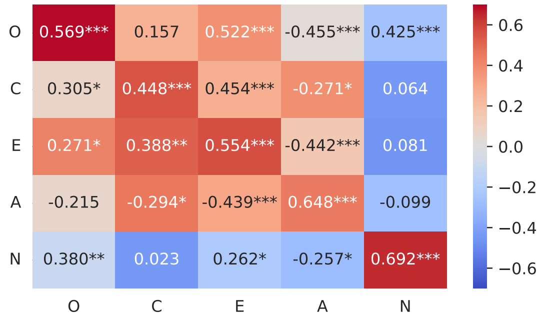


Figure 8: **PCC between predicted and actual OCEAN traits using Llama-3-8b-BFI (Meta, 2024).**

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 Figure 9.

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 4.3 due to its excellent performance and affordable cost. The related correlation results are presented in Figure 10.

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 Figure 11.

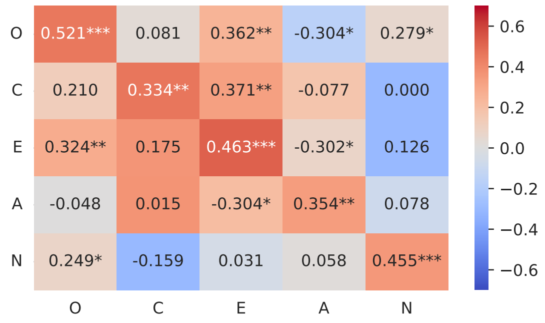


Figure 9: **PCC between predicted and actual OCEAN traits using qwen1.5-110b-chat (Bai et al., 2023).**

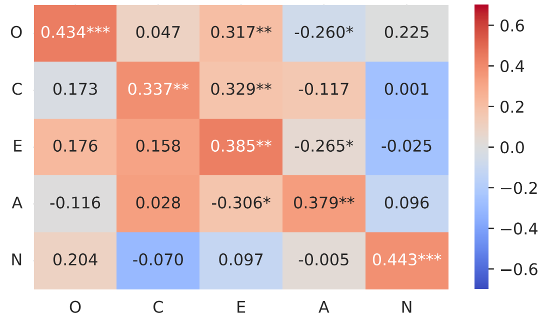


Figure 10: **PCC between predicted and actual OCEAN traits using deepseek-chat (DeepSeek-AI et al., 2024b).**

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 Figure 12.

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.

Table 13 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 13: **Key Hyperparameters for Fine-tuning LLM**

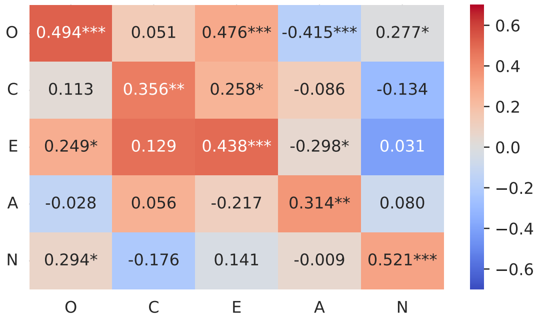


Figure 11: **PCC between predicted and actual OCEAN traits using Gemini-1.5-Pro (Gemini-Team, 2024).**

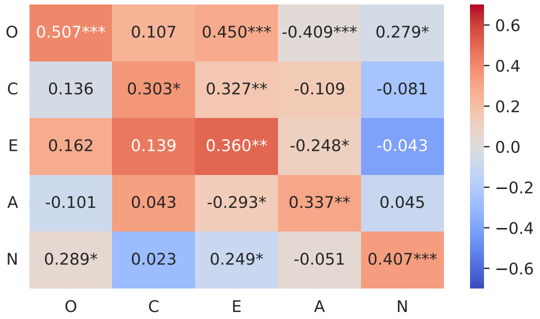


Figure 12: **PCC between predicted and actual OCEAN traits using GPT-4-Turbo (OpenAI, 2023).**