Human-AI Collaboration Enables GPT-4 to Achieve Human-Level User Feedback in Emotional Support Conversations: Integrative Modeling and Prompt Engineering Approaches

Yinghui Huang, Lie Li, Wanghao Dong, Yuhang Dong, Yingdan Huang, Hui Liu

Submitted to: JMIR Mental Health
on: October 17, 2024

Disclaimer: © The authors. All rights reserved. This is a privileged document currently under peer-review/community review. Authors have provided JMIR Publications with an exclusive license to publish this preprint on it's website for review purposes only. While the final peer-reviewed paper may be licensed under a CC BY license on publication, at this stage authors and publisher expressively prohibit redistribution of this draft paper other than for review purposes.

Table of Contents

Human-AI Collaboration Enables GPT-4 to Achieve Human-Level User Feedback in Emotional Support Conversations: Integrative Modeling and Prompt Engineering Approaches

Yinghui Huang[1] PhD, Prof Dr; Lie Li[2]; Wanghao Dong[3]; Yuhang Dong[4]; Yingdan Huang[5]; Hui Liu[1] PhD

[1]Key Laboratory of Adolescent Cyberpsychology and Behavior (Ministry of Education) Wuhan CN
[2]School of Electrical and Electronic Engineering,Nanyang Technological University Singapore SG
[3]Key Laboratory of Adolescent Cyberpsychology and Behavior (Ministry of Education) Luoyu Road?Hongshan District Wuhan CN
[4]School of Management,Wuhan University of Technology Wuhan CN
[5]Department of Lymphoma Medicine, Hubei Cancer Hospital, Tongji Medical College, Huazhong University of Science and Technology Wuhan CN

Corresponding Author:
Hui Liu PhD
Key Laboratory of Adolescent Cyberpsychology and Behavior (Ministry of Education)
152 Luoyu Road?Hongshan District
Wuhan
CN

Abstract

Background: Emotional support is crucial in enhancing social interactions, facilitating psychological interventions, and improving customer service outcomes by addressing individuals' emotional needs. The emergence of large language models (LLMs) offers potential for delivering emotional support on a large scale, but their effectiveness compared to human counselors has not been well understood. Evaluating and enhancing the emotional support capabilities of LLMs through targeted user-centered strategies is crucial for their successful real-world integration.

Objective: This study aims to evaluate the emotional support capabilities of LLMs, specifically GPT-4o, and to introduce an integrative automatic evaluation framework focused on user perceived feedback (UPF). The framework seeks to enhance LLM performance in emotional support conversations (ESCs) by identifying psycholinguistic clues as intrinsic evaluation metrics and utilizing a customized Chain-of-Thought (CoT) prompting strategy.

Methods: The study utilized a dataset of ESCs from human counselors to develop an explanatory predictive model using explainable artificial intelligence methods, following an integrative modeling paradigm rooted in computational social science. This model was designed to evaluate and interpret UPF scores for GPT-4o. Additionally, Hill's three-stage model of helping was integrated into a manually customized CoT prompting framework to evaluate GPT-4o's performance in ESCs.

Results: GPT-4o achieved high UPF scores, demonstrating relative stability in performance, but it still significantly lags behind human counselors overall (Cliff's Delta = 0.087, P < 0.001). The evaluation framework identified 41 distinct linguistic clues related to emotional expression, social dynamics, cognitive processes, linguistic style, and decision-making stages, enhancing the understanding of both processes and outcomes in ESCs. Notably, GPT-4o's UPF scores significantly improved with the use of manually customized COT prompts (Cohen's d = 0.378, P < 0.001), showing no significant difference from the average performance of human counselors overall (Cliff's Delta = -0.014, P= 0.47). However, the COT prompts demonstrated a considerable advantage in specific emotion categories such as fear (Cliff's Delta = -0.23, P = 0.002), sadness (Cliff's Delta = -0.105, P = 0.012), and issues related to breakups with partners (Cliff's Delta = -0.06, P = 0.254). Compared to human counselors, GPT-4o is effective in reducing negative language and conveying emotional tone, but its overemphasis on emotional content weakens its causal reasoning, engagement prompting, and cognitive depth, limiting its ability to handle complex questions and scenarios.

Conclusions: This study offers preliminary evidence of GPT-4o's emotional support capabilities and introduces a UPF-centered integrative evaluation framework for ESCs. The findings suggest a cautiously optimistic outlook for applying advanced LLMs in emotional support services, though significant challenges persist, particularly in deepening conversational exploration and personalizing language. The proposed framework emphasizes the integration of human expertise into LLMs, enhancing their

efficacy and contributing to developing trustworthy AI-based emotional support services.

(JMIR Preprints 17/10/2024:67653)
DOI: https://doi.org/10.2196/preprints.67653

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

Human-AI Collaboration Enables GPT-4 to Achieve Human-Level User Feedback in Emotional Support Conversations: Integrative Modeling and Prompt Engineering Approaches

Original Paper

Yinghui Huang[a,b+], Lie Li [c+], Wanghao Dong [d,e], Yuhang Dong[d,e], Yingdan Huang [f], Hui Liu [d, e]

a. Research Institute of Digital Governance and Management Decision Innovation, Wuhan University of Technology, 122 Luoshi Road, Wuhan, Hubei Province, China, 430070

b. School of Management, Wuhan University of Technology, 122 Luoshi Road, Wuhan, Hubei Province, China, 430070

c. School of Electrical and Electronic Engineering, Nanyang Technological University, Block S2.1, 50 Nanyang Avenue, Singapore, 639798

d. Key Laboratory of Adolescent Cyberpsychology and Behavior (Ministry of Education), 152 Luoyu Road, Hongshan District, Wuhan, Hubei Province, 430079

e. School of Psychology, Central China Normal University, 152 Luoyu Road, Hongshan District, Wuhan, Hubei Province, 430079

f. Department of Lymphoma Medicine, Hubei Cancer Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan 430079, China

+Yinghui Huang and Lie Li are co-first authors.

Abstract

Background: Emotional support is crucial in enhancing social interactions, facilitating psychological interventions, and improving customer service outcomes by addressing individuals' emotional needs. The emergence of large language models (LLMs) offers potential for delivering emotional support on a large scale, but their effectiveness compared to human counselors has not been well understood. Evaluating and enhancing the emotional support capabilities of LLMs through targeted user-centered strategies is crucial for their successful real-world integration.

Objective: This study aims to evaluate the emotional support capabilities of LLMs, specifically GPT-4o, and to introduce an integrative automatic evaluation framework focused on user perceived feedback (UPF). The framework seeks to enhance LLM performance in emotional support conversations (ESCs) by identifying psycholinguistic clues as intrinsic evaluation metrics and utilizing a customized Chain-of-Thought (CoT) prompting strategy.

Methods: The study utilized a dataset of ESCs from human counselors to develop an explanatory predictive model using explainable artificial intelligence methods, following an integrative modeling paradigm rooted in computational social science. This model was designed to evaluate and interpret UPF scores for GPT-4o. Additionally, Hill's three-stage model of helping was integrated into a manually customized CoT prompting framework to evaluate GPT-4o's performance in ESCs.

Results: GPT-4o achieved high UPF scores, demonstrating relative stability in performance, but it still significantly lags behind human counselors overall (Cliff's Delta = 0.087, $P < 0.001$). The evaluation framework identified 41 distinct linguistic clues related to emotional expression, social dynamics, cognitive processes, linguistic style, and decision-making stages, enhancing the understanding of both processes and outcomes in ESCs. Notably, GPT-4o's UPF scores significantly improved with the use of manually customized COT prompts (Cohen's d = 0.378, $P < 0.001$), showing no significant difference from the average performance of human counselors overall (Cliff's Delta = -0.014, P= 0.47). However, the COT prompts demonstrated a considerable advantage in specific emotion categories such as fear (Cliff's Delta = -0.23, P = 0.002), sadness (Cliff's Delta = -0.105, P = 0.012), and issues related to breakups with partners (Cliff's Delta = -0.06, P = 0.254). Compared to human counselors, GPT-4o is effective in reducing negative language and conveying emotional tone, but its overemphasis on emotional content weakens its causal reasoning, engagement prompting, and cognitive depth, limiting its ability to handle complex questions and scenarios.

Conclusions: This study offers preliminary evidence of GPT-4o's emotional support capabilities and introduces a UPF-centered integrative evaluation framework for ESCs. The findings suggest a cautiously optimistic outlook for applying advanced LLMs in emotional support services, though significant challenges persist, particularly in deepening conversational exploration and personalizing language. The proposed framework emphasizes the integration of human expertise into LLMs, enhancing their efficacy and contributing to developing trustworthy AI-based emotional support services.

Keywords: Emotional Support Conversations; User-perceived Feedback; GPT-4o; Prompt Engineering; Explainable Machine Learning; Integrative Modeling

1. Introduction

The global mental health sector faces a significant resource shortage, marked by too few professional providers and various geographic and economic barriers limiting service access. Consequently, many individuals cannot receive the support they need. Underdiagnosis rates near 90% in some low-income regions due to limited access to professional care[1]. Emotional support, a key psychological intervention, helps individuals alleviate stress by providing empathy, affirmation, and encouragement, thereby promoting mental health and social adaptation[2]. Emotional support mitigates negative emotions, enhances psychological resilience and self-confidence, and facilitates emotional communication and social adaptation[2–6]. Despite its importance, the complexity and specialized nature of emotional support make expert resources scarce and services expensive. As a result, many facing life stressors cannot access timely and effective emotional support to mitigate negative emotions[7].

In this context, artificial intelligence (AI) systems like GPT-4 have garnered significant attention for their potential to manage complex human-computer interactions, especially in emotional support conversations (ESCs). Existing technologies have made progress in simulating emotional support dialogues[8–10]. However, the advent of GPT-4 marks a significant advancement in AI capabilities, particularly in contextual understanding and language expression[11,12]. Nevertheless, large language models like GPT-4 still struggle to respond effectively to human emotional needs due to limitations in emotional comprehension and expression[13,14]. Given the critical nature of mental health applications, deploying large language models necessitates establishing comprehensive metrics to assess capabilities, identify potential errors, and implement effective feedback mechanisms. Such metrics are crucial for delivering robust, accurate, and reliable healthcare services. Existing AI conversational systems lack user-centered automatic evaluation frameworks. These frameworks should include both intrinsic metrics (focused on dialogue process details) and extrinsic metrics (focused on user impact). Intrinsic metrics help dynamically understand and enhance subtle differences in complex dialogue processes, while extrinsic metrics evaluate their impact on users. Both warrant close attention.

When addressing real emotional support needs, a critical question arises: Can advanced generative AI, such as GPT-4o, match or even approach the level of professional human counselors? This study proposes an adaptive automatic evaluation framework for emotional support conversations (ESCs) specifically targeting generative artificial intelligence (GenAI). This framework is developed through an integrative modeling process proposed in the field of computational social science and incorporates the role of prompt engineering. The study dynamically evaluates user experiences with advanced large language models (LLMs) in ESCs to explore their potential as effective supplements to mental health services. Integrative modeling combines data-driven insights and human expertise, such as psychological theories, to guide AI system design and implementation[15]. This approach enhances the relevance and accuracy of AI responses, increasing resonance with users' emotional states [16]. Prompt engineering bridges the gap between user intent and AI understanding by guiding GPT-4 to generate more accurate responses through well-crafted prompts[17]. Based on the Chain of Thought (CoT) approach, this strategy requires AI to process problems logically, enhancing its ability to handle complex emotional dialogues[17,18]. This study demonstrates how these techniques assess and enhance GPT-4o's ability to understand and respond to users' emotional needs. This, in turn, increases user trust and satisfaction with AI-driven emotional support platforms and addresses the ethical and technical challenges of deploying AI in mental health applications.

2. Literature Review

This section reviews the relevant literature on the research domain—emotional support capability—the research subject—GenAI and prompt engineering—the research problem—the evaluation of generative AI—and the research methodology—integrative modeling.

2.1 Emotional Support Conversation Systems

Emotional support involves helping individuals cope with life stressors through empathy, affirmation, and encouragement, assisting them in understanding and addressing their challenges[2]. It can be conveyed through both verbal and non-verbal behaviors. Research suggests that non-verbal emotional support typically precedes verbal forms, as non-verbal cues play a crucial role in initial emotional bonding and communication[2]. Providers must not only master various emotional support techniques and expression skills but also choose appropriate response strategies based on the individual's specific situation and issue background[19,20]. According to Hill’s three-stage model of helping, providing emotional support generally involves: exploration, assisting the help-seeker in discovering the problem; insight, helping them gain deeper self-understanding; and action, guiding them to decide how to address the problem[21]. These stages may not occur sequentially and can repeat[21]. Therefore, training an emotional support conversation system with targeted responses for each stage is crucial[9,10,22], enabling its application across various scenarios, including mental health support and social interaction[9].

Before the Emotional Support Conversation (ESC) task was proposed, two well-researched dialogue systems were relevant: emotional chatting[22,23] and empathetic responding[24,25]. Emotional chatting requires the system to respond with or to a given emotion, such as happiness or anger[23]. Empathetic responding involves understanding and sharing the user's emotional experience and responding with empathy[24]. Therefore, distinctions exist between emotional support, emotional chatting, and empathetic responding capabilities[9], with emotional support being relatively more in-depth and complex due to its multi-turn dialogue nature. The ESC task aims to reduce help-seekers' emotional stress and assist them in solving their problems. The ESC chatbot, BlenderBotJoint, uses BlenderBot[26] as a foundation, incorporating emotional support by encoding context history and predicting strategy tokens to guide responses. MISC utilizes a commonsense model to infer the help-seeker's immediate mental state and employs a weighted average strategy for response generation[27]. The global-to-Local Hierarchical Graph Network (GLHG) leverages a graph neural network to encode relationships between the help-seeker's global situation and local intentions for guiding responses[28]. However, both MISC and GLHG depend on external knowledge from Commonsense Transformers (COMET), which may not apply to specific domains and requires considerable human effort to develop. In Addition, they are limited to the scope of the current conversation, overlooking abundant prior knowledge in the dataset.

Studies have demonstrated that emotional support systems significantly alleviate psychological stress and improve mental health in multi-turn emotional support conversations[13,29]. They have also enhanced user satisfaction and interaction quality[30,31]. Nonetheless, challenges remain in accurately recognizing complex emotions and selecting the most appropriate emotional support strategies[31]. Moreover, existing systems struggle to fully account for users' personalities, backgrounds, and specific needs. Their inability to provide explanations for responses results in a lack of transparency[30].

2.2 GPT-4o and Prompt Engineering

Developed by OpenAI, GPT-4 significantly advances large language models [32,33]. GPT-4 omni (GPT-4o), a derivative of GPT-4, retains its core architecture while matching GPT-4 Turbo model in text, reasoning, and coding, and surpassing GPT-4 in multilingual, audio, and vision capabilities, with faster response times and greater cost-efficiency[34]. It excels in contextual understanding, accurately comprehending user inputs and generating appropriate responses, even accommodating subsequent corrections[11,35]. This ability is critical for dialogue agents in mental health, where interactions must be contextually relevant rather than relying on predefined content[35,36]. Furthermore, existing studies underscore GPT-4's strong reasoning abilities, allowing it to draw logical conclusions and provide clinically relevant insights. These capabilities enhance the model's interpretability and build greater user trust in the system[37–40]. As a powerful productivity tool, AI offers immense potential in enhancing various aspects of human mental health services[41]. It can assist in evaluating and improving treatment outcomes, from diagnosis to therapy effectiveness. From a counselor's perspective, AI can leverage therapeutic databases to diagnose conditions, predict client outcomes, and offer treatment suggestions[42]. A meta-analysis has shown that AI conversational agents can significantly alleviate depressive symptoms in patients[43]. For help-seekers, AI-powered chatbots can act as 'digital counselors,' engaging with those experiencing mild symptoms and reducing barriers like stigma and social anxiety[44]. Furthermore, AI can automate the evaluation of treatment effectiveness, making it particularly valuable for training novice counselors and tracking therapeutic progress[45,46]. Building on this potential, this study focuses on GPT-4, a large language model that excels in text-based communication. Although GPT-4 continues to evolve, scientifically validating its emotional support capabilities could address several critical needs: alleviating the shortage of professional counselors, balancing support provider skills, lowering psychological barriers to seeking help, and delivering timely and personalized mental health support. However, due to issues like AI hallucinations and output instability, particularly in sensitive tasks, AI should be seen as a support tool for human experts rather than a replacement to avoid the risk of 'AI replacement threats'[47].

GPT-4's response quality heavily relies on user-provided prompts, a process known as prompt engineering[12]. Prompt engineering bridges user intent and model understanding, ensuring precise, relevant, and coherent interactions by conveying user intentions to the language model. It is essential for achieving the desired functionality of large language models[12,48,49]. Effective prompts greatly enhance GPT-4's output quality and relevance, whereas poorly designed prompts can cause user dissatisfaction or incorrect responses. Effective prompt construction considers multiple factors. One is 'clarity and precision,' which involves crafting specific, clear prompts to reduce output uncertainty. Another is 'role clarity,' where the model is assigned a clear identity, like an assistant or expert, to ensure consistent responses[49]. A popular strategy in prompt construction is the Chain of Thought (CoT), which prompts the model to break down problems into logical steps, enhancing its ability to manage complex issues[50]. This approach encourages the model to explicitly generate a reasoning process, often leading to more accurate conclusions[18]. In emotional problem-solving, grounding the Chain-of-Thought (CoT) in a professional theoretical framework yields more satisfactory outputs. While Large Language Models (LLMs) have demonstrated capacities for emotional understanding [39], expression[51], and empathetic responses[52], these capabilities are best harnessed through human-AI collaboration in prompt engineering. Considering the critical role of CoT prompting, it is crucial to evaluate GPT-4's performance differences before and after CoT implementation.

2.3 Evaluation of Large Language Model-Driven Chatbots

Emotional intelligence agents, which can perceive, integrate, understand, and regulate emotionsis a key research focus in dialogue systems[53],[54]. Chatbots offer 24/7 personalized support and a private therapy option, serving as an alternative to traditional therapy—especially where mental health professionals are scarce or for those concerned about stigma[55],[56]. Research on emotional chatting has recently surged[17,22,23,57]. Studies indicate that users sometimes prefer chatbots over human professionals, aiding those hesitant to seek traditional therapy. Additionally, chatbots are seen as less judgmental and biased than humans, promoting self-disclosure and greater conversational flexibility[56,58].

Despite their potential, chatbot adoption in mental health is limited by concerns over information accuracy, technological maturity, ethics, and interaction authenticity[59]. Thus, evaluating chatbot capabilities is essential for their broader application. Given the complexity of human-AI interactions in mental health, understanding evaluation strategies is crucial. Both automated and manual evaluation methods are commonly used, each with its strengths and weaknesses[60].

Manual evaluations include quantitative methods like surveys and scales measuring user satisfaction[61], and qualitative methods such as interviews and focus groups exploring user experiences and perceptions[62]. Manual methods are valued for their flexibility, comprehensiveness, and professionalism, as human evaluators can manage complex situations and notice details that automated methods might miss[60]. In contrast, automated methods are lauded for efficiency, objectivity, and consistency. They can quickly process large data sets while reducing subjectivity, high costs, and inconsistencies associated with human evaluation[60]. They also mitigate ethical risks in the evaluation process[60]. However, automated methods often suffer from limited coverage, inflexibility, and reliance on predefined benchmarks, which can introduce biases and challenges in addressing new or unforeseen issues.

Intrinsic evaluation metrics assess a language model's ability to generate coherent and meaningful sentences according to language rules and patterns[63]. Known for their computational simplicity, these metrics are categorized into general automatic metrics and dialogue-based metrics. General automatic metrics, like BLEU (Bilingual Evaluation Understudy), ROUGE (Recall-Oriented Understudy for Gisting Evaluation), and Perplexity[60], measure precision and F1-score by counting matching word sequences between reference and generated text. Dialogue-based metrics include Match-rate, which measures the percentage of successful diagnoses; Dialogue Accuracy, assessing the chatbot's ability to ask relevant questions; and Average Request Turn, tracking the average number of interactions between user and chatbot[60]. While these metrics provide valuable quantitative tools for evaluating LLMs, they focus on surface-level similarity and language-specific aspects. This focus makes them insufficient for healthcare chatbots, as they fail to capture crucial elements like semantics, context, long-range dependencies, and human perspectives, especially in real-world scenarios[60].

In contrast, extrinsic evaluation metrics assess the model's impact on users and how well it meets their expectations and needs[64]. These metrics measure language model performance by incorporating user perspectives and real-world scenarios[63]. Collected through subjective assessments involving human judgment, extrinsic metrics are divided into general-purpose and health-specific categories[48]. General-purpose human evaluation metrics assess LLM performance across diverse domains[60]. evaluating quality, fluency, relevance, and overall effectiveness, and covering a broad range of real-world topics, tasks, and user needs[65]. Health-specific metrics evaluate how healthcare-oriented LLMs and chatbots process and generate health-related information, emphasizing accuracy, effectiveness, relevance, reliability, timeliness, healthy behaviours, and emotional support[60]. These metrics aim to incorporate context and semantic awareness into extrinsic evaluations of LLMs. However, existing studies often focus on specific metrics, overlooking a comprehensive evaluation of healthcare language models and chatbots. Few studies have explored domain-agnostic metrics that combine intrinsic and extrinsic evaluations for healthcare LLMs. Notably, Laing et al. (2023) introduced a multi-metric approach, evaluating LLMs on accuracy, calibration, robustness, fairness, bias, toxicity, and efficiency[65]. As Abbasian et al. (2024) suggest, a balanced evaluation approach that integrates intrinsic and extrinsic metrics better addresses scientific consensus, potential harms, and user satisfaction[60].

Overall, automated evaluation methods can effectively adapt to the dynamic scenarios required by prompt engineering, adequately covering diverse configurations related to psychological help-seeking users, domains, and task types. Intrinsic metrics emphasize computational efficiency and consistency, quantifying a chatbot's performance in the face of complex and subtle interaction differences. Extrinsic metrics focus on user perspectives and application performance, but relevant research has not yet fully addressed comprehensive, complex, and user-centered evaluation metrics[66]. GGiven that user satisfaction, therapeutic effectiveness, engagement, and reliability are widely used to measure the capabilities of GPT-4[61,62], user acceptance—as reflected in subjective feedback—is closely related to the effectiveness of emotional support[67] and can serve as a direct indicator of emotional support quality. Therefore, integrating intrinsic and extrinsic metrics based on subjective acceptance can provide more accurate evaluations and deeper understanding, offering important insights for optimizing chatbots and improving user satisfaction. Additionally, current evaluation schemes primarily focus on performance metrics—measuring accuracy, reliability, and user interaction experience—but often lack the robustness needed to effectively evaluate LLMs in the complex and nuanced interactions required in mental health applications[68].

2.4 The Application of Explainable AI and Integrative Modeling in Mental Health Assessment

In recent years, significant progress has been made in using AI to understand and evaluate psychotherapy dialogues.Unsupervised learning techniques have been applied to identify clusters in unlabeled patient or therapy data, aiding in the analysis of therapeutic processes and outcomes[69]. Building on these insights, supervised learning techniques are widely used to classify or predict labeled therapeutic processes and outcomes. Machine learning has effectively predicted dropout rates in outpatient psychotherapy. For example, researchers have used ensemble methods, such as random forests and nearest-neighbor modeling, to identify patients at high risk of dropping out, especially those with severe depression[70]. Deep learning models trained on large-scale online cognitive behavioral therapy conversations have been used to classify therapists' verbal behaviors and assess their links to clinical outcomes[71]. Deep learning has also excelled in personalized treatment prediction, achieving up to 80% accuracy in predicting treatment responses for depression[72].

When ML models do not meet any of the criteria imposed to declare them transparent, a separate method must be devised and applied to the model to explain its decisions. The purpose of post-hoc explainability techniques (also referred to as post-modeling explainability), which aim at communicating understandable information about how an already developed model produces its predictions for any given input [73]. Model-agnostic XAI techniques for post-hoc explainability are designed to be plugged into any model with the intent of extracting some information from its prediction procedure [73]. Feature relevance explanation techniques aim to describe the functioning of an opaque model by ranking or measuring the influence, relevance, or importance each feature has in the prediction output by the model to be explained. An amalgam of propositions are found within this category, each resorting to different algorithmic approaches with the same targeted goal, such as SHAP (Shapley Additive Explanations) [74], QII (Quantitative Input Influence) [75], ASTRID (Automatic Structure Identification) [76], and other methods. Additionally, several techniques have excelled in coding and interpret psychotherapy dialogue content. The Linguistic Inquiry and Word Count (LIWC) tool, developed by Pennebaker and colleagues, analyzes the psychological and social functions underlying language use[77]. Studies show that improvements in psychological states correspond with changes in language patterns, indicating therapy effectiveness[78–80]. Similarly, Linguistic Style Matching (LSM) measures the synchrony in linguistic style between conversation partners, with research suggesting that higher language matching is linked to better therapeutic relationships and outcomes[81,82]. Moreover, Latent Dirichlet Allocation (LDA) models, especially labeled-LDA, have been used to automate therapy session coding and predict effectiveness[83]. Furthermore, Hofman and colleagues introduced integrative modeling, a research paradigm that combines explanatory and predictive approaches. This paradigm uses data-driven machine learning for prediction and causal inference methods to ensure model interpretability and reliability[15]. For instance, researchers have used machine learning to predict content popularity in social networks and experimental approaches to verify causal effects, combining prediction and explanation to better understand information diffusion[16]. By enhancing AI transparency and interpretability, XAI and integrative modeling can address social and ethical issues, promoting broader acceptance of mental health[84].

XAI has gained significant attention in psychotherapy due to its vital role in enhancing model transparency and interpretability. By offering real-time explanatory feedback, XAI assists therapists in adjusting their strategies, thereby improving patient outcomes[85]. Utilizing conversational AI tools like chatbots, XAI provides personalized support and conducts real-time analyses of emotional changes, which enhances therapists' understanding of patients' states[40]. Moreover, XAI parses and explains language patterns and emotional shifts during psychotherapy sessions, offering detailed decision paths that help therapists identify therapeutic opportunities and potential risks[86]. Therefore, XAI could enables human-AI collaboration by providing interpretable decision suggestions, assisting clinicians in understanding and trusting AI decisions, thereby improving treatment planning and quality[87,88]. Collectively, XAI and integrative modeling demonstrate great potential for advancing AI applications in clinical practice. By fostering effective human-AI collaboration in psychotherapy, they promote the adoption of these technologies, ultimately enhancing treatment outcomes and patient well-being.

2.5 The Current Study

The global shortage of mental health resources highlights the urgent need to leverage generative AI to meet growing demand. Given the critical role of emotional support conversations in mental health, accurately evaluating generative AI's abilities in ESC is crucial for broader adoption. AI-based automated methods are valued for their efficiency, objectivity, and consistency, enabling rapid processing of large datasets with minimal human bias. Perceived feedback (UPF) is a key user-centered evaluation metric. This study evaluates GPT-4's advanced capabilities in ESCs, posing the first research question (RQ1): Can GPT-4 achieve high user feedback ratings in emotional support conversations, especially compared to human counselors?

Secondly, although AI-based evaluation methods are promising, their limited coverage and flexibility may overlook the complex interactions required in ESCs, leading to biases and challenges in addressing unforeseen issues. Intrinsic metrics evaluate a language model's vocabulary, grammar, and sentence structure, while extrinsic metrics focus on user experience and real-world outcomes, potentially causing significant discrepancies. Using psychological and linguistic clues and explainable machine learning in mental health assessments offers a promising approach to evaluating GPT-4's ability in ESCs. This leads to the second research question (RQ2): How can we integrate internal and external metrics to develop a framework for evaluating GPT-4's performance in user-perceived feedback during ESCs?

Thirdly, the integrative evaluation framework, especially with internal metrics, can effectively guide large language models' performance in specific domains. However, it remains unclear how this framework impacts generative AI capabilities and decision-making despite advances in transparency. Given that prompt engineering is key to enhancing and understanding generative AI performance in specific scenarios, this study proposes the third research question (RQ3): Can customized prompts, based on the integrative evaluation framework, elevate GPT-4's performance to levels comparable to human counselors?

To address the research questions, following the Integrative Modeling process, this study first proposes an AI-based predictive model to assess help-seeker's UPF toward GPT-4o’s responses in ESCs, comparing its performance with that of human counselors. Second, AI-based explanatory modeling methods are employed to identify psychological and linguistic cues related to help-seeker's UPF as intrinsic metrics, thereby developing a integrative evaluation framework to assess GPT-4o's responses in ESCs in detailed, while mitigating privacy risks, offering a more granular and insightful evaluation. Third, building on Hill’s three-stage model of helping, along with the established evaluation framework and AI methods, this study developed a customized CoT prompt, and reevaluated GPT-4's responses.

3. Methods & Experiments

This study utilizes a dataset of ESCs between human counselors and help-seekers. It applies an integrative modeling process from computational social science to evaluate and understand user feedback on GenAI in ESCs. This process involves data collection, preprocessing, feature engineering, constructing an ESC UPF evaluation model, and integrating human expertise through CoT prompt engineering. These steps aim to evaluate GPT-4's emotional support capabilities in ESCs. The study proposes an integrative evaluation framework and conducts a comprehensive comparison with human counselors. The following sections outline the datasets, features, algorithms, and prompt engineering methods used to develop, explain, and evaluate the AI models, along with a summary of the methodology and experimental process of integrative modeling.

3.1 Emotional Support Conversation Dataset

The study used the English ESC datasets developed by Liu et al. (2021)[9]. In their work, Liu and colleagues modified the second stage of Hill's three-stage helping skills model from "insight" to "comfort," focusing on providing support and understanding through empathy. This adjustment was made because "insight" often requires reinterpreting the user's behaviors and feelings, which can be challenging and risky for less experienced support providers. During dataset collection, the researchers provided detailed ESC framework training to support providers, selecting the top 7.8% of applicants who passed the examination. This process resulted in 1,053 high-quality ESCs. The dataset comprises dialogue texts from real counseling scenarios created by professional counselors and help-seekers and includes the following annotations:

(1) Empathy Rating by Help-Seeker: After each session, the help-seeker rated the support provider's empathy and understanding on a scale of 1 to 5, with higher scores indicating greater perceived empathy.

(2) Counselor's Response Strategies and Stages: The specific strategies and stages of helping used by the counselor in their responses.

(3) Help-Seeker's Problem Type and Emotion Category: The help-seeker chose one problem type from five options and one emotion category from seven.

(4) Source of Experience: Indicates whether the help-seeker's situation was based on a current or past life experience.

Descriptive statistics for this datasets are presented in Tables 1 and 2. Each conversation averages 29.8 turns. The most frequently mentioned issues were ongoing depression (30.1%) and work-related crises (24.9%). Feedback for support providers was generally high, with 50.5% rated as 'excellent'. Among support strategies, asking questions (20.9%) and offering advice (15.6%) were the most common.

Table 1: Statistical Data of ESCs

Type	Total Num	Supporter Num	Help-Seeker Num
Number of Conversations	2,016	-	-
Average Duration of Conversations (minutes)	22.6	-	-
Number of Participants	854	425	532
Number of Utterances	31,410	14,855	16,555
Average Length of Conversations (turns)	29.8	14.1	15.7
Average Length of Utterances	17.8	20.2	15.7

Table 2: Statistical Data of Annotations in ESCs

Category		Num	Percentage
Help-Seeker Issues	Ongoing Depression	608	30.1
	Work Crisis	502	24.9
	Breakup with Partner	296	14.7
	Issues with Friends	326	16.2
	Academic Pressure	284	14.1
	Total	2,016	100
Help-Seeker Emotions	Anxiety	590	29.27
	Depression	510	25.3
	Sadness	440	21.8
	Anger	166	8.23
	Fear	154	7.64
	Disgust	62	3.08
	Shame	68	3.37
	Total	2,016	100
Help-Seeker Perceived Feedback Scores	1 (Very Poor)	71	1.1
	2 (Poor)	183	2.9
	3 (Average)	960	15.5
	4 (Good)	1,855	29.9
	5 (Excellent)	3,144	50.5
	Total	6,213	100
Support Strategies Used by Supporters	Asking Questions	3,109	20.9
	Restating	883	5.9
	Reflecting Emotions	1,156	7.8
	Self-Disclosure	1,396	9.4
	Affirmation and Reassurance	2,388	16.1
	Offering Advice	2,323	15.6
	Providing Information	904	6.1
	Other	2,696	18.1
	Total	14,855	100

3.2 Integrative Modeling Approach for ESC user-perceived Feedback

The study first employed NLP, machine learning, and deep learning methods to develop models that predict and explain help-seeker's UPF regarding GPT-4o's emotional support capabilities. Then, based on Hill's helping skills theory, the study designed a Chain-of-Thought (CoT) prompting framework and conducted a detailed human-machine comparative analysis to evaluate and enhance GPT-4o's performance in delivering emotional support.

3.2.1 Automated Evaluation of user-perceived Feedback in ESCs

To validate GPT-4o's emotional support capabilities, we developed an evaluation model for help-seeker's UPF scores in ESCs. This model employs machine learning and deep learning techniques to predict UPF scores based on the dialogue content between seekers and supporters during ESCs. The UPF scores serves as the model's predictive target. The feature set includes linguistic metrics such as LIWC and LSM. Additionally, annotations from ESCs, such as emotion categories and problem types faced by the seeker, are incorporated.

The study utilized several regression algorithms, including Ridge Regression, Random Forests (RF), Extreme Gradient Boosting (XGBoost), and Support Vector Regression (SVR)[89]. We applied Recursive Feature Elimination with 10-fold Cross-Validation combined with the XGBoost model to filter out features that significantly impacted the prediction of perceived feedback scores, simplifying the variable set. We then used Grid Search Cross-Validation to determine the hyperparameters that optimize predictive performance[89].

Additionally, the study employed deep neural network and pre-trained language models, including Bidirectional Encoder Representations from Transformers-Bidirectional Long Short-Term Memory (BERT-BiLSTM)[90], Bidirectional Encoder Representations from Transformers-Bidirectional Long Short-Term Memory-Attention (BERT-BiLSTM-Attention), Robustly Optimized BERT Pretraining Approach (RoBERTa)[91], and eXtreme Language Model (XLNet)[92]. These models, with their distinct architectures and optimization strategies, offer strong semantic understanding and predictive capabilities, making them suitable for UPF evaluation in ESCs.

3.2.2 Explanatory Modeling and Prompt Engineering Methods for ESC user-perceived Feedback

Integrative modeling combines explanatory and predictive approaches to identify causal factors and forecast their impact in new situations. This approach involves evaluating and validating the accuracy of model predictions. This study evaluate GPT-4o's UPF scores in ESCs, identify and estimate the impact of relevant psycholinguistic cues (as intrinsic evaluation metrics) guided by human experience, and further enhance GPT-4's performance through customized prompt engineering.

Specifically, we used explainable machine learning to develop customized prompts for GPT-4o's responses in ESCs, focusing on expressing empathetic language. This prompt engineering, driven by explanatory modeling, uses the UPF scores as the dependent variable and linguistic features from the counselor's responses—such as LIWC and LSM—as input features to build predictive models. The model aims to evaluate how these linguistic features influence UPF scores, providing a scientific basis for constructing effective prompts.

For different stages of the ESC, we first used Recursive Feature Elimination with 10-fold cross-validation combined with XGBoost to filter out features that significantly impacted the prediction of UPF scores, thus simplifying the variable set. Next, we applied Shapley Additive Explanations (SHAP), a cooperative game theory tool, to quantify the impact of each feature on model outputs at different ESC stages, enhancing the transparency of the machine learning models[73,74]. Based on SHAP analysis, the study conducted a sensitivity analysis to determine the impact of features on model outputs and ranked them accordingly.

To prevent issues with model interpretability and generalization due to a large and complex feature subset, The study further refined the predictive model. Specifically, the study used the forward stepwise selection method, starting with an empty feature subset and incrementally adding new most important features that improved model performance (measured by MAPE and RMSE) at each step. This process continued until all top-n important features were added. The study identified a specific number of top-n features, beyond which additional features did not significantly enhance model performance.

The proposed model highlights key linguistic cues influencing UPF scores in ESCs, utilizing SHAP values to understand how the supporter's language impacts UPF predictions. This analysis provided critical guidance for developing GPT-4's response strategies. Drawing from Hill’s three-stage helping skills model, we combined these explainability analyses with human experience in ESCs to create a manual CoT prompt strategy framework, offering key insights to enhance GPT-4's performance. Additionally, we introduced the RTF (Request, Task, and Format)-based CoT prompt framework as a control against the manual CoT prompts mentioned earlier[93]. This structured approach clearly defines the requirements, specific tasks, and expected format of ESC outputs. These steps aim to improve GPT-4o's performance in ESCs, enabling it to better understand and respond to help-seekers' needs, thereby providing more effective emotional support.

In summary, The study first developed a predictive model to evaluate GPT-4o's emotional support performance based on UPF scores. Second, The study performed prompt engineering to enhance GPT-4's emotional support capabilities. This involved identifying key linguistic features influencing help-seeker feedback through XAI and model refinement methods, and designing manual CoT prompts for ESC tasks. Finally, The study quantitatively evaluated GPT-4's UPF scores in ESCs compared to human counselors through statistical analysis. The main research process and methods are illustrated in Figure 1.

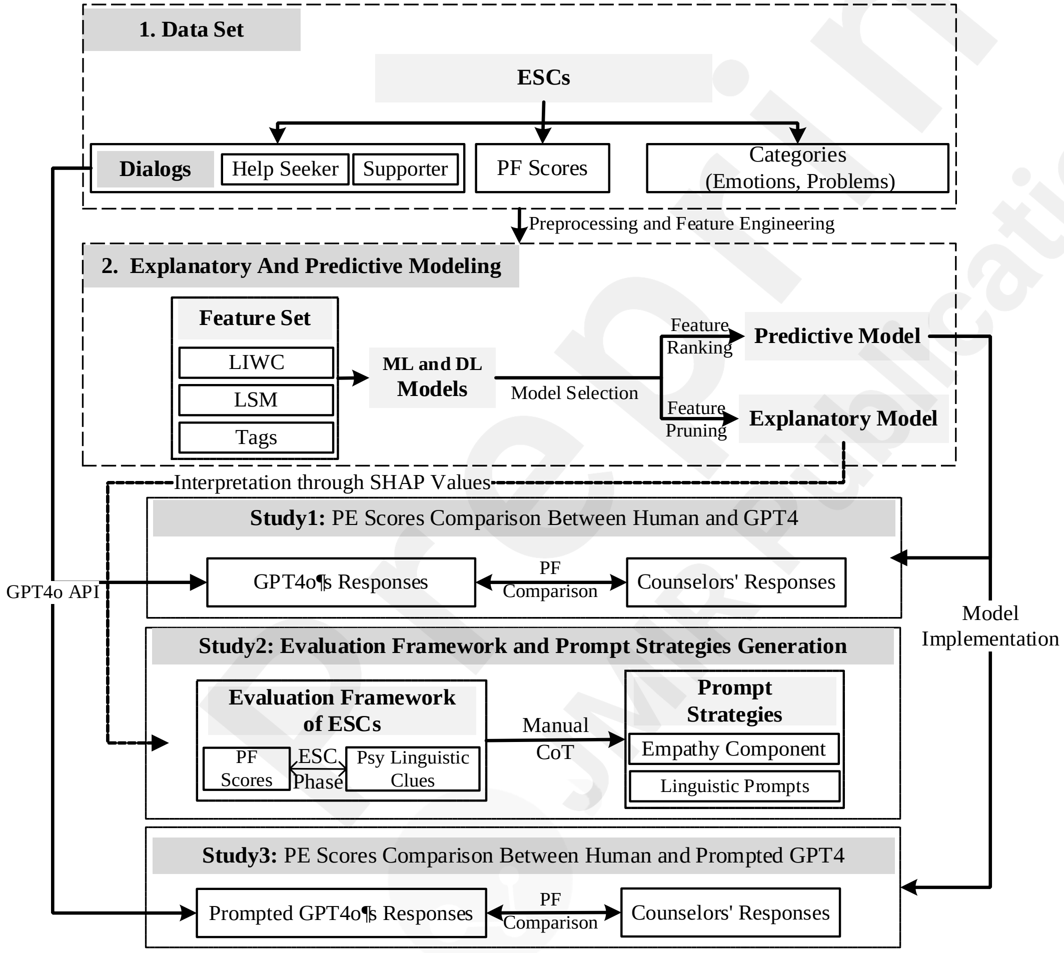


Figure 1. Research Methodology and Process

4. Results

4.1 Analysis of UPF of GPT-4o in ESCs Using Predictive Modeling

In this study, we developed a predictive model to assess UPF scores in ESCs by training and testing various regression algorithms and selecting the one with the best performance. It then evaluated the UPF scores of GPT-4o's responses and compared them with those from human counselors to determine whether GPT-4o's performance in ESCs approaches human-level capabilities.

4.1.1 Performance Analysis of UPF Evaluation Model

We developed various predictive models using a range of regression algorithms, and the performance metrics for each model are presented in Table 3. Among the deep learning models, the BERT-BiLSTM-Attention model achieved a Root Mean Square Error (RMSE) of 0.579 and a Mean Absolute Percentage Error (MAPE) of 21.36%. Among the machine learning models, XGBoost, though not a deep learning model, delivered the best performance with an RMSE of 0.486 and a MAPE of 17.13%.

Table 3: Performance of Different Regression Models

Model	RMSE	MAPF(%)
Ridge	0.5910	25.26
RF	0.4902	18.88
XGBoost	0.486	17.13
SVR	0.6019	22.83
Bert-Bilstm	0.581	26.09
Bert-Bilstm-Attention	0.579	21.36
RoBERTa	0.8443	21.5593
XLNet	0.8492	20.698

4.1.2 Comparison of UPF Between GPT-4o and Human Counselors

Using XGBoost, the optimal machine learning model for predicting UPF scores, we predicted the UPF values for GPT-4o's responses. Since the UPF scores of human counselors did not follow a normal distribution, We applied the Mann-Whitney U test—a non-parametric method for independent samples—for statistical analysis[94]. Additionally, we used Cliff's Delta to measure the effect size between human counselors and GPT-4o. The results are presented in Table 4.

We observed that, overall, GPT-4o's performance was lower than the average level of human counselors, with the differences showing a small effect size ($p < 0.001$). Among different emotions, GPT-4o’s performance was not significantly different from that of human counselors in cases of anxiety, depression, and fear, where it underperformed. Regarding different problem types, GPT-4o generally underperformed compared to human counselors. For themes such as friendship issues, work crises, persistent depression, and academic pressure, the effect sizes of the differences were small but statistically significant. However, for "breakup" and "friendship issues," the differences were nearly negligible. These findings address RQ1.

Table 4: Comparison of Perceived Feedback Scores Between GPT-4o and Human Counselors

ESC Category		N	Median and Interquartile Range (IQR)		U	P	Cliff's Delta
			Human Counselor	GPT-4o			
Experience Category	All Data	1851	5.0□4.0□5.0□	4.538 (4.356, 4.705)	1861449	< 0.001	0.087
	Previous Experience	435	4.0□4.0□5.0□	4.53 (4.33, 4.72)	95955	< 0.001	0.153
	Recent Experience	1416	5.0□4.0□5.0□	4.54 (4.36, 4.72)	900684	< 0.001	0.089
Emotion	Anxiety	517	4.0□4.0□5.0□	4.51 (4.46, 4.57)	11883	<	0.114

					6	0.001	
Category	Anger	165	5.0（4.0，5.0）	4.53 (4.39, 4.703)	12703	0.192	0.085
	Fear	120	5.0（4.0，5.0）	4.51 (4.46, 4.55)	6234	0.591	-0.041
	Depression	540	4.0（4.0，5.0）	4.53 (4.35, 4.69)	148337	< 0.001	0.191
	Disgust	75	4.0（4.0，5.0）	4.5(4.29, 4.73)	1397	0.516	0.074
	Sadness	376	5.0（4.0，5.0）	4.57(4.404, 4.74)	68954	0.57	0.024
	Shame	57	4.0 (3.75,5.0)	4.55 (4.37, 4.73)	1407	0.06	0.221
Issue Category	Friendship Issues	251	5.0（4.0，5.0）	4.58 (4.405, 4.74)	32765	0.431	0.04
	Work Crisis	465	4.0（4.0，5.0）	4.51 (4.31, 4.67)	124150	< 0.001	0.148
	Ongoing Depression	451	4.0（4.0，5.0）	4.53 (4.34, 4.71)	117936	< 0.001	0.16
	Breakup	328	5.0（4.0，5.0）	4.56 (4.39, 4.74)	53586	0.931	-0.003
	Academic Pressure	200	4.0（4.0，5.0）	4.52 (4.34, 4.65)	22709	<0.01	0.147

[a]**N** refers to the sample size, indicating the number of samples in each category.

[b]**Human Counselor** represents ratings provided by human counselors, expressed as the median and interquartile range (e.g., 5.0 [4.0, 5.0] indicates a median of 5.0 with an interquartile range of 4.0 to 5.0). [c]**GPT-4o** denotes ratings by GPT-4o, also presented as the median and interquartile range (e.g., 4.487 [4.437, 4.534] indicates a median of 4.487 with an interquartile range of 4.437 to 4.534).

[d]**U** refers to the Mann-Whitney U statistic, a measure used to compare the distributions of two groups.

[e]**P** indicates the p-value, which determines the statistical significance of the results, with a p-value less than 0.05 indicating significance.

[f]**Cliff's Delta** represents the effect size between two groups, where higher values indicate a stronger effect. According to Macbeth et al. (2011), a Cliff's Delta value below 0.147 is considered negligible, 0.147 to 0.333 indicates a small effect, 0.333 to 0.474 indicates a medium effect, and values above 0.474 indicate a large effect[95].

4.2 Development and Validation of the Integrative Evaluation Framework for ESCs

To further understand and evaluate GPT-4o's performance regarding UPF scores, We developed explanatory models. It integrated them with relevant theories to create a user feedback-centered integrative evaluation framework for generative AI. The framework was used to perform a detailed quantitative evaluation of the responses generated by GPT-4o and human counselors.

4.2.1 Development of the Integrative Evaluation Framework of ESCs

First, We conducted a sensitivity analysis to assess the impact of each feature within the best-performing model. Next, we performed feature pruning, retaining only those features that significantly affected predictive accuracy. The model was then refined using key features based on Davis's empathy component theory and Hill’s three-stage helping skills model. Details of the results from each process are presented in the following sections.

In this study, SHAP values were used for sensitivity analysis and feature pruning to enhance the interpretability of the predictive model. Initially, SHAP values were computed for different features to rank their importance. Features were incrementally incorporated into the model, and their impact on refinement was assessed using the MAPE.

As shown in Figure 2, during the construction of the prompt models for different ESC stages, the initial number of features was 127. After feature engineering, this number was reduced to 28 for the Exploration stage, 125 for the Comforting stage, and 115 for the Action stage. The model was further refined using SHAP values and forward stepwise selection, focusing on MAPE to optimize performance. The explanatory model achieved local optimal performance in the Exploration, Comforting, and Action stages with 14, 17, and 14 features, respectively.

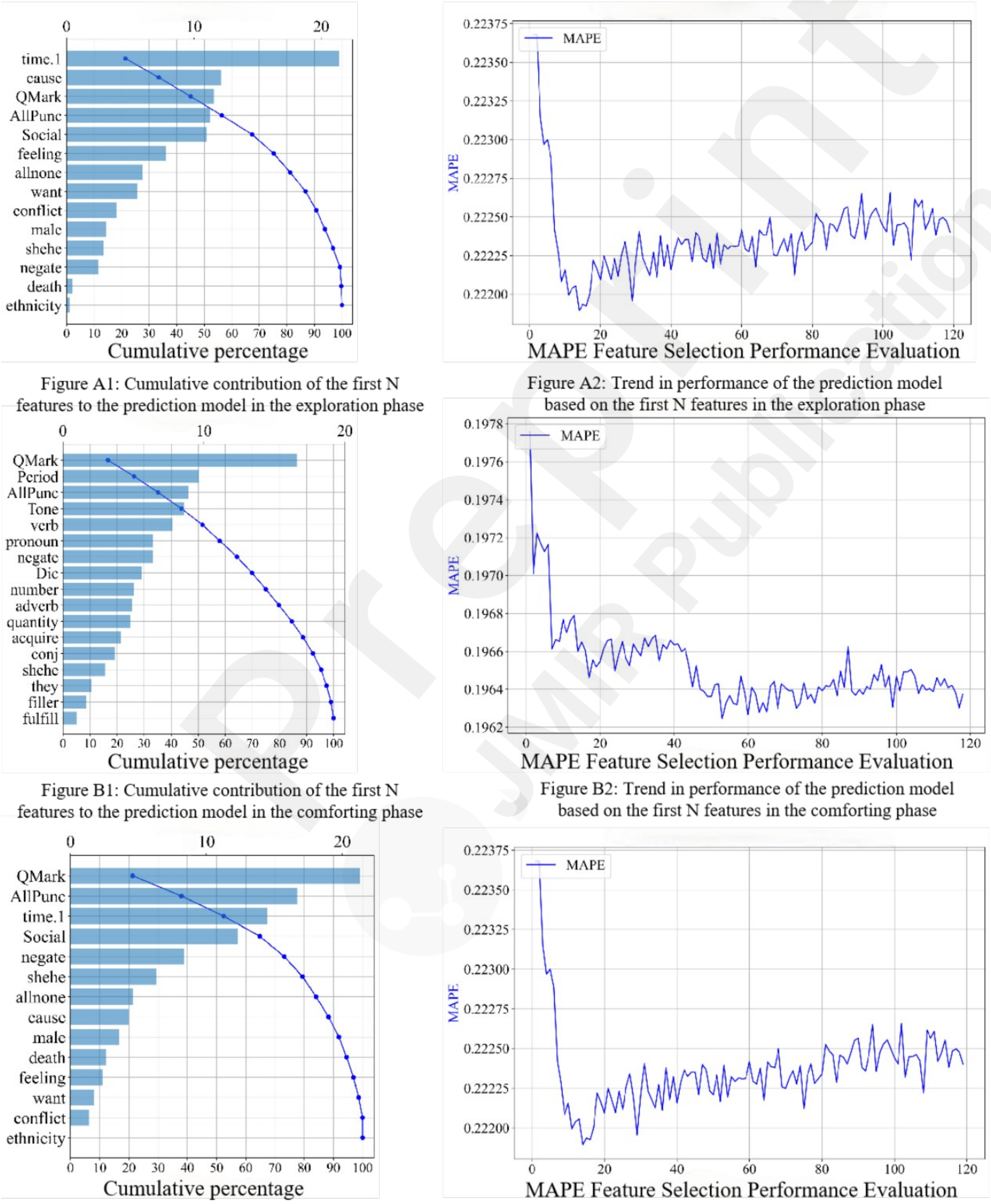


Figure A1: Cumulative contribution of the first N features to the prediction model in the exploration phase

Figure A2: Trend in performance of the prediction model based on the first N features in the exploration phase

Figure B1: Cumulative contribution of the first N features to the prediction model in the comforting phase

Figure B2: Trend in performance of the prediction model based on the first N features in the comforting phase

Figure C1: Cumulative contribution of the first N features to the prediction model in the action phase

Figure C2: Trend in performance of the prediction model based on the first N features in the action phase

Figure 2: Top-N Important Features and Performance of Feature Sets Composed of Different Numbers

of Optimal Features

To develop a customized CoT prompts framework for GPT-4o, we conducted a global interpretability analysis using SHAP values to assess how various features influence UPF scores. We identified and ranked the cumulative SHAP values of the top-N features in each ESC stage according to their impact, as shown in Figure 3 (Figures A1, B1, C1). To effectively determine the range and direction of each feature's impact on PF, we consolidated individual data samples into comprehensive SHAP interpretability plots, as shown in Figures A2, B2, and C2 in Figure 3.

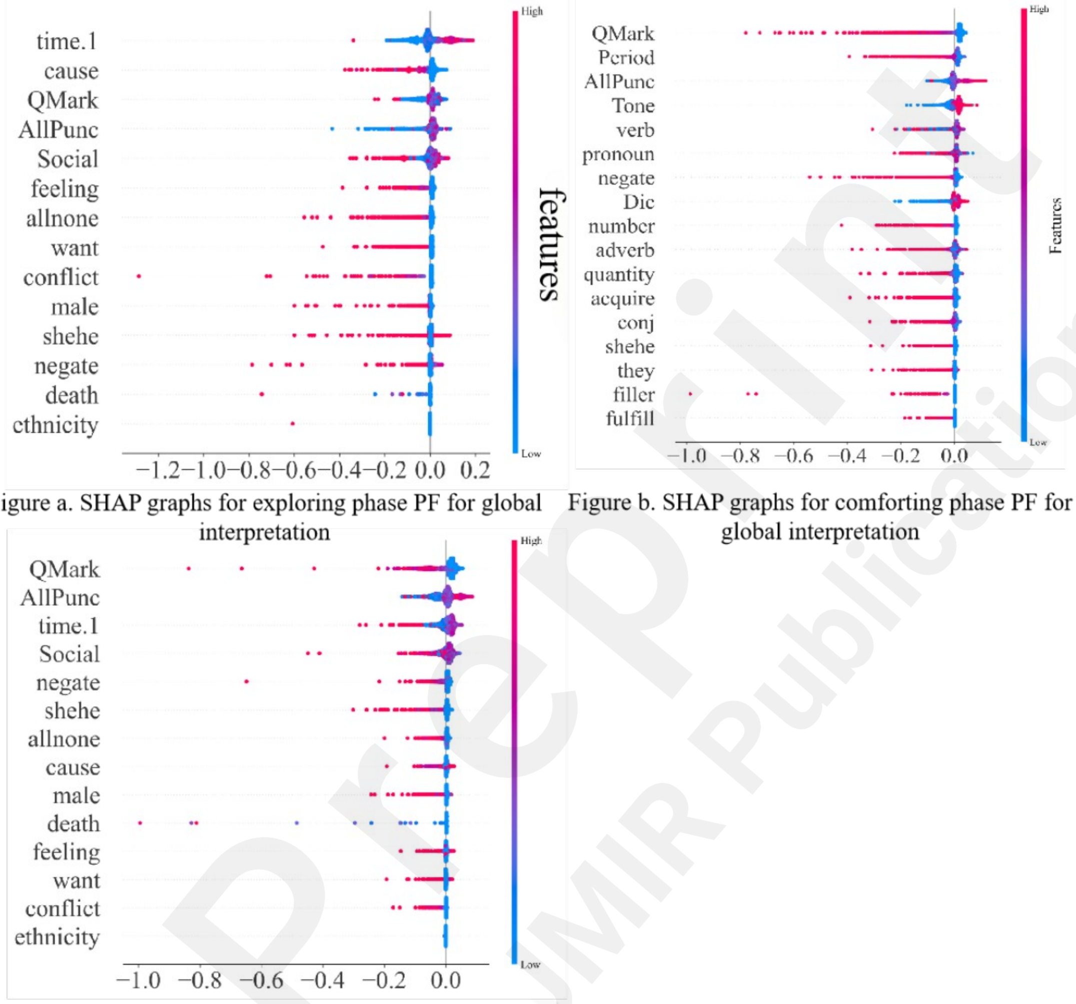


Figure a. SHAP graphs for exploring phase PF for global interpretation

Figure b. SHAP graphs for comforting phase PF for global interpretation

Figure c. SHAP graphs for action phase PF for global interpretation

Figure 3: SHAP Plot of Global Interpretability for Empathy Scores in ESCs

This analysis revealed two primary ways in which features influence UPF scores across different stages of ESCs. Specifically, as shown in Table 5, intrinsic metrics (i.e., key linguistic cues) were identified related to the three stages of Hill's helping skills theory and showed diverse relationships with the extrinsic metric (i.e., UPF score).

In the Exploration stage, punctuation-related features like Question Mark (QMark) and All Punctuation (AllPunc) were positively associated with the PF, while negatively associated features included emotional cues such as Feeling, pronouns like She/He, and various social and cognitive aspects such as Conflict, Male, Ethnicity, and Negation. The same pattern extended to other categories like Time in Time Orientation and Death in Body-related topics.

In the Comforting stage, positively associated intrinsic metrics included Verb usage in the Psychological Process, Tone in Emotion, and Dictionary Coverage (Dic) in Linguistic dimension.

Negatively associated features encompassed Pronoun and Number usage, along with Adverbs, Quantities, and Conjunctions. Other significant negative associations included the use of pronouns like They and She/He, Filler in Social Process, and Punctuation-related features like Question Mark (QMark) and Period.

In the Action stage, Conflict and All Punctuation (AllPunc) were positively linked to perceived feedback, while features such as Feeling, Male, and Social Process were negatively associated. The usage of Question Mark (QMark), She/He, and Negation in Cognitive Processes, as well as Ethnicity and Death-related topics, also showed negative associations with the PF.

Table 5: Integrative Evaluation Framework Based on user-perceived Feedback in ESCs

ESC Stages	Categories of Intrinsic Metric	Intrinsic Metric	Relation Between Intrinsic Metric and the PF
Exploration	Affect	feeling	Negative
	Linguistic Dimensions	shehe	Negative
	Social processes	social	Positive
		conflict	Negative
		male	Negative
	Culture	ethnicity	Negative
	States	want	Negative
	Time orientation	time	Positive
	Cognitive processes	cause	Negative
	Punctuation	QMark	Positive
		AllPunc	Positive
	Cognitive processes	allnone	Negative
		negate	Negative
	Physical	death	Negative
Comforting	Linguistic Dimensions	pronoun	Negative
		dic	Positive
		number	Negative
		adverb	Negative
		quantity	Negative
		conj	Negative
		they	Negative
		fulfill	Negative
	Psychological Processes	verb	Positive
	Affect	tone	Positive
	Social processes	shehe	Negative
		filler	Negative
	States	acquire	Negative
	Punctuation	QMark	Negative
		Period	Negative
		AllPunc	Positive
	Cognitive processes	negate	Negative
Action	Affect	feeling	Negative
	Cognitive processes	allnone	Negative
		cause	Negative
	Social processes	male	Negative
		conflict	Positive
	Culture	ethnicity	Negative
	State	want	Negative
	Social processes	Social	Negative
	Punctuation	AllPunc	Positive
		QMark	Negative
	Time orientation	time	Negative
	Linguistic Dimensions	shehe	Negative
		negate	Negative

Physical	death	Negative

[a]*ESC Stages* refer to the three stages of ESCs.

[b]*Linguistic Clues* highlight the specific linguistic elements utilized in each stage, including words or punctuation marks used during the conversation, such as *time* (*time*-related words) and *QMark* (question mark).

4.2.2 Validation of the Integrative Evaluation Framework of ESCs

To verify the effectiveness of intrinsic evaluation metrics related to UPF scores in ESCs between human counselors and GPT-4o, we conducted paired sample t-tests to compare linguistic clue features across different ESC stages. The differences in intrinsic evaluation metrics were analyzed using t-tests, and the results are presented in Table 6.

Table 6. Analysis of the Performance of GPT-4o and Human Counselor in the Intrinsic Evaluation Metric [a, b]

ESC Stages	Category	Intrinsic Metric (Linguistic Clues)	Relation Between Intrinsic Metric and PF	Cohen's d (GPT-4o-Human counselors)
Exploration	Affect	feeling	Negative	0.19
	Linguistic Dimensions	shehe	Negative	-6.74***
	Social processes	Social	Positive	3.914***
		conflict	Negative	-2.784**
		male	Negative	-6.014***
	Culture	ethnicity	Negative	-0.205
	States	want	Negative	-6.592***
	Time orientation	time	Positive	-2.185*
	Cognitive processes	allnone	Negative	-28.852***
		negate	Negative	-14.804***
		cause	Negative	-7.528***
	Punctuation	QMark	Positive	-26.835***
		AllPunc	Positive	37.515***
	Physical	death	Negative	-1.701
Comforting	Linguistic Dimensions	pronoun	Negative	-34.701***
		Dic	Positive	-27.733***
		number	Negative	34.190***
		adverb	Negative	-18.527***
		quantity	Negative	-2.973**
		conj	Negative	10.019***
		they	Negative	-2.947**
		fulfill	Negative	2.094*
		shehe	Negative	-7.738***
	Psychological Processes	verb	Positive	26.726***
	Affect	Tone	Positive	17.866***
	Social processes	filler	Negative	-4.356***
	States	acquire	Negative	-0.555
	Punctuation	QMark	Negative	-21.637***
		Period	Negative	9.233***
		AllPunc	Positive	49.967***
	Cognitive processes	negate	Negative	-19.962***
Action	Affect	feeling	Negative	1.609
	Cognitive processes	allnone	Negative	-12.567***
		cause	Negative	-0.067
	Social processes	male	Negative	-4.649***
		conflict	Positive	-2.469*
	Culture	ethnicity	Negative	-0.975
	States	want	Negative	-5.126***

Time orientation	Social	Negative	4.268***
Punctuation	AllPunc	Positive	38.315***
	time	Negative	-1.612
	QMark	Negative	-20.401***
Linguistic Dimensions	shehe	Negative	-5.394***
	negate	Negative	-15.482 (***)
Physical	death	Negative	-1.339

[a]**gpt-4o-Human Counselors:** Indicates the statistical differences in the use of these linguistic clues between GPT-4o and human counselors. The values represent the t-values, with asterisks indicating the level of statistical significance: p < 0.05 (*), p < 0.01 (**), p < 0.001 (***). Positive values indicate greater usage by GPT-4o, while negative values indicate greater usage by human counselors.

[b]**Category:** Refers to the category each linguistic clue belongs to, based on LIWC-22[96], covering various aspects from linguistic dimensions to psychological processes, culture, lifestyle, and physical states, as shown in Appendix Table S1 for the details.

In the exploration stage, GPT-4o demonstrated superior social interaction through increased use of social language. Compared to human counselors, it uses conflict-related and male-related expressions less frequently, indicating that it is more effective than human counselors in reducing such language, which has a positive impact on perceived feedback (UPF) scores. GPT-4o's fewer expressions of needs indicate restraint but adversely affect PF. Its increased use of time-related expressions highlights strength in time orientation. Regarding cognitive processes, GPT-4o's higher use of causal expressions suggests less effective causal reasoning, while a lower use of all-or-none logic expressions indicates better handling of binary thinking. A cautious reduction in negations contributes to better outcomes. Despite greater language complexity reflected by increased punctuation use, deficiencies in constructing questions may hinder exploration and guidance. Overall, GPT-4o performs comparably to human counselors in cultural, emotional, and physical topics.

In the comforting stage, GPT-4o is less effective than human counselors in areas where a higher Linguistic Inquiry and Word Count (LIWC) dictionary word count positively correlates with PF. Its increased use of numerical expressions and conjunctions—metrics negatively correlated with PF—indicates inferior performance in these linguistic dimensions. However, a significant presence of emotional tone, which positively correlates with PF, demonstrates that GPT-4o excels in conveying emotion. The fewer fillers used by GPT-4o suggest superior social interaction. Reduced use of question marks, despite being positively correlated with PF, indicates weaker engagement prompting. Nonetheless, increased overall punctuation use shows stronger language complexity, even though overuse of periods negatively impacts PF. GPT-4o's cautious use of negations, negatively correlated with PF, suggests a positive impact on user perceptions.

In the action stage, GPT-4o's restrained use of all-or-none expressions positively impacts PF, indicating better performance in emotional and cognitive processing. Its cautious handling of gender-related language and conflict expressions improves UPF in social interactions. The expression of needs is negatively correlated with PF, and the reduction in need expressions suggests that GPT-4o should have outperformed human counselors in this aspect. While GPT-4o shows an advantage with more diverse punctuation, its less frequent use of question marks negatively affects PF, indicating weaker ability in prompting user engagement.

Overall, GPT-4o demonstrates both strengths and weaknesses compared to human counselors across different stages. It excels in reducing negative language and, in some stages, effectively conveys emotional tone, positively impacting UPF scores. However, in the Action stage, an overemphasis on emotional content—evidenced by increased use of feeling-related language—negatively affects PF. Deficiencies in causal reasoning are apparent, as GPT-4o's higher use of causal expressions suggests less effective causal understanding. Additionally, reduced use of question marks indicates weaker engagement prompting, potentially hindering user interaction. Overuse of certain linguistic features, such as numerical expressions and conjunctions (which negatively correlate with PF), and underuse of LIWC dictionary words highlight areas where GPT-4o could improve to better match or exceed human counselor performance. These findings address RQ2.

4.3 Comparative Analysis of UPF Between Prompted GPT-4o and Human Counselors in ESCs

This section describes the development process of manually customized CoT prompts. It presents a comparative analysis of UPF scores for responses generated by human counselors, the GPT-4o model with manually customized CoT prompts, and the GPT-4o model with standard CoT prompts.

4.3.1 Development of Manually Customized CoT Prompts for users' PF

The prompt engineering process integrated the intrinsic metrics with Hill's three-stage helping skills model to develop a customized manual CoT prompt framework for the Exploration, Comforting, and Action stages. An overview of the framework is provided in Table 7, with specific prompts detailed in Appendix S2.

Table 7: Framework of Manually Customized CoT Prompts for Enhancing UPF scores in ESCs

ESC Stages	Category	Strategies of psycholinguistic.categories to Promote PF	Intrinsic Metric (Linguistic Clues)	Strategies for Linguistic Cue Usage to Promote PF
Exploration	Affect	Emphasize emotional expression to enhance empathy.	feeling	Decrease
	Linguistic Dimensions	Use inclusive and gender-neutral language to foster equitable communication.	shehe	Decrease
	Social processes	Foster engagement by enhancing social interaction and avoiding conflict.	Social	Increase
			conflict	Decrease
			male	Decrease
	Culture	Integrate multicultural understanding to reduce cultural biases.	ethnicity	Decrease
	States	Focus on the individual's current psychological state and emotional needs.	want	Decrease
	Time orientation	Emphasize time perception, paying attention to personal historical experiences.	time	Increase
	Punctuation	Enhance the exploration and clarity of individual experiences through appropriate punctuation.	QMark	Increase
			AllPunc	Increase
	Cognitive processes	Encourage thoughtful reflection and appropriate rebuttals to create a more positive dialogue atmosphere.	cause	Decrease
			allnone	Decrease
			negate	Decrease
	Physical	Minimize negative body-related language to maintain a positive dialogue.	death	Decrease
Comforting	Linguistic Dimensions	Provide empathetic responses directly and moderately.	pronoun	Decrease
			Dic	Increase
			number	Decrease

			adverb	Decrease
			quantity	Decrease
			conj	Decrease
			they	Decrease
			fulfill	Decrease
			shehe	Decrease
	Psychological Processes	Use verbs and active voice cautiously to describe emotional support and understanding.	verb	Increase
	Affect	Comfort and inspire with positive tone and emotional expression.	Tone	Increase
	Social processes	Use social support and emotional connection with care.	filler	Decrease
	States	Promote a focus on perceived personal achievements and progress with care.	acquire	Decrease
	Punctuation	Use punctuation wisely to enhance the integrity and rhythm of language.	QMark	Decrease
			Period	Decrease
			AllPunc	Increase
	Cognitive processes	Encourage reflective and self-affirming cognitive activities cautiously.	negate	Decrease
Action	Affect	Mobilize emotional resources carefully to facilitate practical action.	feeling	Decrease
	Cognitive processes	Emphasize causal reasoning and action-oriented thinking with caution.	allnone	Decrease
			cause	Decrease
	Social processes	Confront difficulties directly to promote problem-solving.	male	Decrease
			conflict	Increase
	Culture	Support diverse action strategies.	ethnicity	Decrease
	States	Focus on expressing motivation and intent to inspire goal achievement.	want	Decrease
	Time orientation	Use time management skills to encourage effective action planning.	Social	Decrease
	Punctuation	Employ rhythmic language to promote action.	AllPunc	Increase
			time	Decrease
			QMark	Decrease
	Linguistic Dimensions	Focus on the self to inspire action.	shehe	Decrease
			negate	Decrease
	Physical	Minimize negative body-related language to maintain a positive dialogue.	death	Decrease

4.3.2 Comparative Analysis of UPF Between GPT-4o and Human Counselors

Using the framework, we developed customized CoT prompts and re-generated GPT-4o responses. We then performed an automatic evaluation of UPF scores in ESCs for three groups: human counselors, GPT-4o with standard CoT prompts, and GPT-4o with manually customized CoT prompts. We used paired sample t-tests to evaluate differences in UPF scores among the three groups. To assess the practical effect of the optimized prompt engineering on GPT-4o's performance, we also calculated the effect size using Cohen's d.

Table 8 shows that, in most emotion and issue categories, GPT-4o responses with customized prompts significantly improved over those with non-customized prompts, demonstrating a substantial impact on GPT-4o's UPF scores. Specifically, the results indicated that, compared to the original version, the optimized GPT-4o did not achieve statistically significant improvements in user-perceived feedback for the emotion of disgust but showed at least small effect size improvements (p < 0.05) in all other subdomains. Notably, the effect sizes of the improvements reached medium levels in categories such as *Current Experience* (Cohen's d = 0.401), *Depression* (Cohen's d = 0.412), *Job Crisis* (Cohen's d = 0.421), and *Breakup with Partner* (Cohen's d = 0.421).

Table 8: Comparison of Perceived Feedback Scores of GPT-4o before and after manual CoT prompting

Category	Topic	N	Mean-Percentage Diff(100%)	meandiff	p	Cohen's d
	All Data	1851	11.43%	0.0302	< 0.001	0.378
Experience Type	Prior Experience	435	10.488%	0.028	< 0.001	0.245
	Current Experience	1416	11.94%	0.031	< 0.001	0.401
Emotion Type	Emotion Type	517	12.59%	0.033	<0.001	0.4
	Anger	165	7.905%	0.021	<0,001	0.282
	Fear	120	15.556%	0.041	<0.001	0.533
	Depression	540	11.99%	0.032	<0.001	0.412
	Disgust	75	7.325%	0.0202	0.104	0.2323
	Sadness	376	10.659%	0.0276	< 0.001	0.35
	Shame	57	12%	0.0315	0.008	0.394
Problem Type	Issues with Friends	251	8.07%	0.029	<0.001	0.367
	Job Crisis	465	12.39%	0.033	<0.001	0.421
	Ongoing Depression	451	12.1%	0.0321	< 0.001	0.39
	Breakup with Partner	328	13.12%	0.0337	<0.001	0.421
	Academic Pressure	200	10.56%	0.0288	<0.001	0.367

4.3.3 Comparative Analysis of UPF Scores Between GPT-4o and Human Counselors

Because the UPF scores from human counselors are non-normally distributed, we used the Mann-Whitney U test to assess whether the optimized GPT-4o provides emotional support comparable to that of human counselors. The results are presented in Table 9.

Regarding experience types, the performance of GPT-4o was not significantly different from that of human counselors, suggesting that GPT-4o handles different experience types comparably. For problem types, GPT-4o showed only a small effect size advantage (Cohen's d = -0.13) in handling breakups with a partner. There were no significant differences between GPT-4o and human counselors in other problem types.

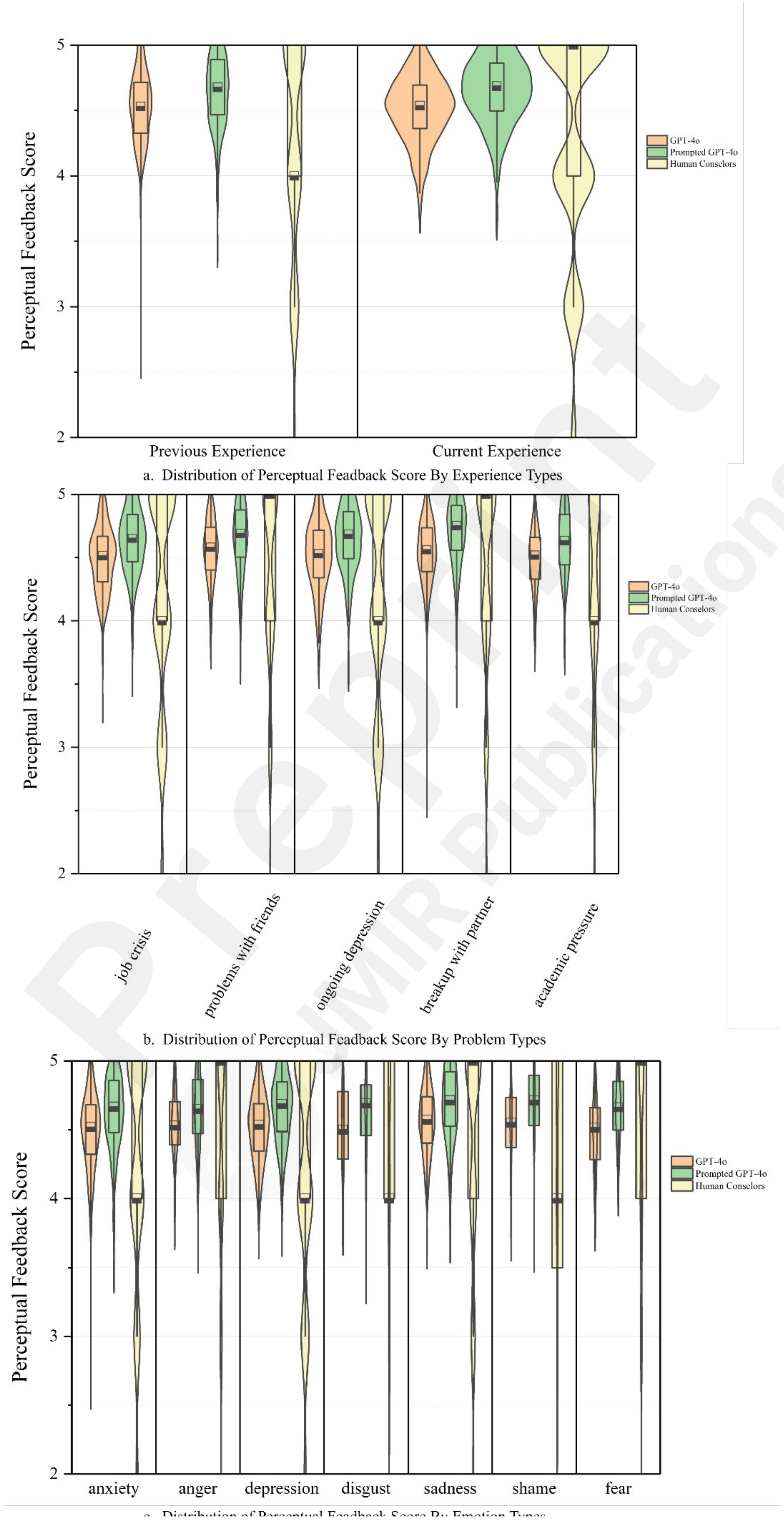
However, when analyzed by emotion types, the advantage of human counselors diminished in specific emotional contexts. For emotions such as anger, anxiety, shame, and disgust, the optimized GPT-4o did not show a statistically significant difference in UPF scores compared to human counselors. This indicates that GPT-4o has achieved emotional support performance comparable to that of human counselors in these specific emotional contexts.

Table 9: Comparison of Perceived Feedback Scores Between Optimized GPT-4o and Human Counselors

ESC Category		N	Median and Interquartile Range (IQR)		U	P-adj	Cliff's Delta
			Human Counselors	Manual CoT Prompted GPT-4o			
	All Data	1851	5.0□4.0□5.0□	4.682□4.495□4.863□	1689815	0.47	-0.014
Experience Type	Prior Experience	435	4.0□4.0□5.0□	4.676(4.497,4.861)	89298	0.07	0.07
	Recent Experience	1416	5.0□4.0□5.0□	4.686(4.467,4.888)	810536	0.38	-0.02
Emotion Type	Anxiety	517	4.0□4.0□5.0□	4.665□4.482.□4.857□	111051	0.28	0.04
	Anger	165	5.0□4.0□5.0□	4.65□4.472□4.865□	11640	0.93	-0.005
	Fear	120	5.0□4.0□5.0□	4.663(4.499,4.85)	5014	0.002	-0.23
	Depression	540	4.0□4.0□5.0□	4.688(4.487,4.847)	135411	0.016	0.09
	Disgust	75	4.0□4.0□5.0□	4.69(4.464,4.81)	1333	0.83	0.02
	Sadness	376	5.0□4.0□5.0□	4.714(4.528,4.921)	60262	0.012	-0.105
	Shame	57	4.0□3.75, 5.0□	4.712(4.535,4.893)	1308	0.251	0.135
Problem Type	Issues with Friends	251	5.0□4.0□5.0□	4.69□4.504□4.877□	29672	0.254	-0.06
	Job Crisis	465	4.0□4.0□5.0□	4.652(4.469,4.84)	113515	0.183	0.05
	Ongoing Depression	451	4.0□4.0□5.0□	4.684(4.493,4.863)	107121	0.162	0.05
	Breakup with Partner	328	5.0□4.0□5.0□	4.752□4.558□4.91□	46829	0.004	-0.13
	Academic Pressure	200	4.0□4.0□5.0□	4.633□4.45,4.84□	21328	0.18	0.077

The study analyzed the distribution of UPF scores across different ESCs to evaluate the relative performance of human counselors and GPT-4o with manually customized CoT prompts. Figure 4 shows that human responses have a wider rating range, reflecting diversity and variability in handling dialogue tasks. In contrast, GPT-4o's ratings are more concentrated, indicating greater consistency and predictability in response quality.

Firstly, GPT-4o's UPF ratings are concentrated between 4.3 and 4.6, showing a more clustered and smoother distribution. This clustering trend reflects the overall improvement in GPT-4o's UPF performance after applying manual CoT prompts. Secondly, GPT-4o lacks scores in the highest satisfaction range (5 points) compared to human counselors, indicating shortcomings in demonstrating high-level empathetic abilities.



a. Distribution of Perceptual Feedback Score By Experience Types

b. Distribution of Perceptual Feedback Score By Problem Types

c. Distribution of Perceptual Feedback Score By Emotion Types

Figure 4: Distribution of UPF Scores for GPT-4o and Human Counselors Across Different Emotion and Problem Categories

4.3.4 Comparison Analysis of Key Linguistic Clues Usage between GPT-4o After Prompt Engineering and Human Counselors

To verify the effectiveness of manually customized CoT prompting on intrinsic evaluation metrics related to UPF in ESCs, We conducted paired sample t-tests to compare linguistic clue features among human counselors, GPT-4o prompted by auto-CoT (*GPT-4o*) and GPT-4o prompted by manually customized CoT (*Prompted GPT 4o*) across different ESC stages. Results are presented in Table 10. Additionally, Appendix S3 presents a ESC case analyzed using SHAP local interpretability.

Table 10: Comparison of Average Values for Features Affecting UPF in Responses from Human Counselors and GPT-4o Before and After Prompt Engineering.

ESC Stages	Category	Intrinsic Metric (Linguistic Clues)	Relation Between Intrinsic Metric and PF	Cohen's d (Prompted GPT 4o–Humans)	Cohen's d (Prompted GPT 4o–GPT-4o)
Exploration	Affect	feeling	Negative	14.121***	19.719***
	Linguistic Dimensions	shehe	Negative	-8.164***	-1.933
	Social processes	Social	Positive	16.249***	14.448***
		conflict	Negative	-4.196***	-2.055*
		male	Negative	-6.600***	-0.742
	Culture	ethnicity	Negative	-1	-1.568
	States	want	Negative	-4.871***	2.543*
	Time orientation	time	Positive	2.477*	5.831***
	Cognitive processes	allnone	Negative	-12.787***	0.403
		negate	Negative	-15.579***	-1.462
		cause	Negative	-5.644***	2.809**
	Punctuation	QMark	Positive	-12.481***	24.801***
		AllPunc	Positive	17.376***	-24.736***
	Physical	death	Negative	-2.540*	-2.182*
Comforting	Linguistic Dimensions	pronoun	Negative	-12.066***	29.073***
		Dic	Positive	-9.959***	23.312***
		number	Negative	4.214***	-24.367***
		adverb	Negative	-7.966***	16.367***
		quantity	Negative	1.942	7.843***
		conj	Negative	1.809	-14.463***
		they	Negative	-9.156***	-10.622***
		fulfill	Negative	-3.971***	-9.096***

	Psychological Processes	verb	Positive	-11.125***	21.177***
	Affect	Tone	Positive	27.835***	12.599***
	Social processes	shehe	Negative	-10.274***	-3.699***
		filler	Negative	-4.585***	-0.49
	States	acquire	Negative	1.75	3.509***
	Punctuation	QMark	Negative	-16.563***	12.393***
		Period	Negative	6.725***	-6.238***
		AllPunc	Positive	20.330***	-38.035***
	Cognitive processes	negate	Negative	-16.001***	10.085***
Action	Affect	feeling	Negative	3.833***	2.374*
	Cognitive processes	allnone	Negative	-10.841***	4.165***
		cause	Negative	-3.519***	-4.823***
	Social processes	male	Negative	-5.834***	-1.566
		conflict	Positive	-4.008***	-2.014*
	Culture	ethnicity	Negative	-1.38	-0.803
	States	want	Negative	-4.865***	0.651
	Time orientation	Social	Negative	3.620***	-1.268
	Punctuation	AllPunc	Positive	21.445***	-15.261***
		time	Negative	14.285***	19.700***
		QMark	Negative	-12.387***	17.692***
	Linguistic Dimensions	shehe	Negative	-7.011***	-2.286*
	Physical	death	Negative	-2.323*	-1.997*

In the exploration phase, compared to GPT-4o, Prompted GPT-4o exhibits several advantages: it more effectively employs social and conflict-related language, demonstrates higher cultural sensitivity by using fewer ethnicity-related terms, and performs better in utilizing time-related language—these are strengths it shares when compared to human counselors as well. It uses more cause-related language than GPT-4o, indicating enhanced cognitive capabilities, and employs more punctuation overall. However, it significantly overuses the term "feeling," potentially overemphasizing emotional content. Compared to human counselors, prompted GPT-4o shows certain advantages: it more effectively employs social and conflict-related language, demonstrates higher cultural sensitivity through reduced use of ethnicity-related and gender-related terms, and performs better in utilizing time-related language—advantages it also has over GPT-4o. However, it significantly overuses the term "feeling," potentially overemphasizing emotional content—a shared shortcoming with its comparison to GPT-4o. Additionally, it uses fewer cause-related words than human counselors, indicating weaker performance in cognitive processes, even though it surpasses GPT-4o in this aspect. Its overall use of punctuation is higher than that of human counselors and positively correlated with UPF scores, indicating that prompted GPT-4o performs better in this aspect.

In the comforting phase, compared to GPT-4o, prompted GPT-4o demonstrates several advantages in language cue usage. It significantly reduced the use of fillers and third-person singular and plural pronouns (such as she/he/they), indicating improvements in structural and social aspects.

Additionally, it used fewer conjunctions and periods, which had a positive impact on UPF scores, indicating that the prompted GPT-4o performed better in terms of linguistic conciseness and clarity of expression, contributing to improved user understanding and satisfaction. Additionally, it exhibits higher positive tone expression and percent words captured by LIWC, suggesting enhanced affective, standardized and professional communication. However, compared to GPT-4o, the prompted GPT-4o also has some disadvantages, such as using more question marks and adverbs, which may indicate less effective use of interrogative punctuation and descriptive language. Compared to human counselors, prompted GPT-4o shows strengths by using fewer third-person singular and plural pronouns (she/he/they) and expressions related to "fulfill," demonstrating improved social processing and empathy. However, it has disadvantages such as using significantly lower percent words captured by LIWC, implying weaker use of standardized and professional vocabulary. The prompted GPT-4o also used more conjunctions, periods, and overall punctuation, while using fewer pronouns and adverbs, indicating stronger performance in grammatical structure and expression completeness, but it may lack personalization and detail in descriptions. It also shows higher dictionary coverage, which might reflect a more technical or less conversational style. Additionally, it uses more question marks, negation words, and expressions related to "acquire," indicating potential weaknesses in communication effectiveness relative to human counselors.

In the action phase, compared to GPT-4o, Prompted GPT-4o demonstrates several advantages in language cue usage. It significantly reduces gender-related (male) and conflict-related language, ethnicity-related terms, and death-related topics, indicating better restraint, adaptability, and cultural sensitivity. It also uses fewer third-person singular pronouns (she/he), suggesting improved focus on the client, and employs more overall punctuation marks, enhancing readability. However, Prompted GPT-4o exhibits disadvantages such as increased use of feeling-related language—which negatively correlates with user-perceived feedback—potentially overemphasizing emotional content. It also uses fewer causal relationship terms and expressions related to "want," which may reflect weaker cognitive processing and less emphasis on client desires. Compared to human counselors, prompted GPT-4o shows advantages by reducing gender-related (male) and conflict-related language, ethnicity-related terms, death-related topics, and third-person singular pronouns (she/he), indicating better restraint, adaptability, cultural sensitivity, and client focus. It also employs more time-related expressions, which may enhance temporal context in communication. However, it has disadvantages such as increased use of feeling-related language compared to human counselors, potentially overemphasizing emotional content and reducing nuanced communication. It uses fewer expressions related to "want," suggesting less emphasis on client desires, and employs fewer question marks, which may diminish clarity and engagement compared to human counselors.

Overall, in the exploration phase, compared to GPT-4o, prompted GPT-4o demonstrates better performance in cognitive processing and cultural sensitivity, although it tends to overemphasize emotional content. When compared to human counselors, prompted GPT-4o shows advantages in cognitive processing and cultural sensitivity but falls short in emotional regulation and cognitive depth. In the comforting phase, prompted GPT-4o displays more mature structural and emotional processing than GPT-4o but lacks in descriptive language and engagement. Compared to human counselors, it shows improvements in emotional expression and empathy, yet lags behind in cognitive processing and structural expression while also overemphasizing emotional content. In the action phase, prompted GPT-4o excels in emotional control, adaptability, and cultural sensitivity, outperforming both GPT-4o and human counselors in these areas. However, it still overemphasizes emotional content, and its cognitive depth and engagement remain weaker, falling short of ideal outcomes. These findings address RQ3.

5. Discussion

This study collected and analyzed GPT-4o's responses to over 1,300 real ESCs.. Using an integrative modeling paradigm from computational social science, we employed machine learning, deep learning, and natural language processing (NLP) methods to create an automatic evaluation model for assessing GPT-4o's UPF scores in ESCs. We then developed an explainable model, based on Hill’s three-stage helping skills model, to identify key psychological linguistic cues affecting UPF scores. We also proposed an integrated evaluation framework for ESCs. Additionally, we developed customized CoT prompts based on intrinsic and extrinsic metrics and compared GPT-4o's responses with those of human counselors.

5.1 Principal Findings

The results show that GPT-4o exhibits a notable capability in providing emotional support. Manual CoT prompts, developed using the evaluation framework, significantly improved GPT-4o's performance in ESCs. Overall, the methods employed in this study enhance the performance of LLMs in ESCs and improve the transparency and interpretability of the optimization process. By adopting an integrative modeling paradigm from computational social science, this study presents an evaluation framework for user experience in ESCs involving generative artificial intelligence. This framework includes intrinsic metrics (key linguistic cues) and extrinsic metrics (PF scores), creating an effective bridge between human expertise and the capabilities of large language models. This offers a new path and perspective for providing emotional support services using large language models.

5.1.1 Performance of GPT-4o in Emotional Support Conversations

Based on UPF scores, we first compared GPT-4o's performance with that of human counselors in emotional support conversations. The results indicate that GPT-4o generally falls short of human counselors, especially in managing emotions such as anxiety, depression, and fear. GPT-4o also significantly underperforms compared to human counselors in addressing issues such as breakups, friendship problems, work crises, chronic depression, and academic pressure, with these differences demonstrating a small effect size. These findings highlight the limitations of GPT-4o in mimicking human counselors, particularly regarding its effectiveness in dealing with specific emotional responses and stressful situations.

Secondly, we found that GPT-4o provides emotional support comparable to that of human counselors when dealing with specific emotions such as anger, shame, and disgust. However, significant differences persist when addressing deeper understanding and more complex emotional experiences, such as sadness and depression, consistent with recent research[39]. This discrepancy may stem from the need for deeper self-reflection and personal experience, which GPT-4o lacks. Human counselors, with their extensive emotional experiences and profound empathic abilities, excel in providing emotional support—a capability that current large language models have yet to develop fully. This likely reflects limitations in the quality and diversity of training data and model structures for generative large language models[33]. These limitations may impact GPT-4o's creativity and ability to manage complex dialogues, constraining its performance in higher scoring ranges.

Thirdly, we explored GPT-4o's emotional support capabilities through □ □ human-AI collaboration in prompt engineering. The experiments revealed that, except for the emotion of disgust, customized prompts significantly enhanced GPT-4o's performance in most scenarios. Additionally, GPT-4o, using manually customized prompts, achieved perceived feedback levels comparable to those of human counselors when managing specific emotions such as anger, anxiety, shame, and disgust.

Fourthly, we assessed the stability of users' UPF scores for GPT-4o and human counselors. We found that GPT-4o's perceived feedback scores were consistently clustered around the median, indicating stability, while human counselors exhibited greater variability in their scores. However, GPT-4o had fewer scores in the highest satisfaction range, indicating a need for improvement in expressing higher levels of empathy. This discrepancy may arise from the dynamic and varied nature of help-seeking issues. GPT-4o tends to organize and generate content in a standardized manner[97], resulting in a more uniform approach to similar issues even without preset guidelines. This leads to more consistent responses from GPT-4o.

5.1.2 Integrated Evaluation Framework for User Experience in Emotional Support Conversations

This study employs an integrative modeling approach to address the lack of user-centered metrics for evaluating generative AI capabilities in ESCs and the inconsistencies between intrinsic and extrinsic metrics. We began by identifying key linguistic cues and their impact on uses' UPF scores, using these clues as intrinsic evaluation metrics to assess generative AI's performance in ESCs.

In the exploration stage of ESCs, we utilized 13 metrics from 10 distinct psychological and linguistic cue categories. These metrics include indicators for the use of open-ended questions and various inquiry punctuation marks, which measure how effectively replies encourage the recipient's self-expression and exploration of their situation, thus enhancing the dynamism and engagement of the discourse. This aligns with existing research, as the LIWC tool effectively captures emotional and cognitive cues in language, revealing aspects such as confidence, motivation, and needs[98,99]. Additionally, metrics such as emotional expression, social dynamics, gender issues, cultural differences, and cognitive barriers evaluate the psychological stress and resistance the recipient faces when addressing personal and social issues [100,101].

In the comforting stage, we utilized 17 metrics from 7 distinct psychological and linguistic cue categories. These include metrics for the use of verbs and positive tone, which measure the degree of emotional support and psychological comfort provided by the counselor, as well as the use of specialized vocabulary to assess the specificity and depth of the language. This aligns with existing research indicating that LIWC analysis effectively captures emotional support and language specificity[102]. Additionally, metrics such as abstract and indirect language cues and communication hesitations evaluate how responses may diminish personalization and directness in interactions, potentially creating barriers to emotional connection and understanding [98,99].

In the action stage, we utilized 11 intrinsic metrics from 10 distinct psychological and linguistic cue categories. These metrics include those related to managing social conflict and emotional expression, which assess problem-solving and decision-making abilities, reflecting the recipient's motivation and decisiveness [102]. Additionally, metrics such as persistent emotional distress, social issues, and uncertainty regarding actions evaluate barriers to effective action implementation and the attainment of personal goals [100,101]. These intrinsic metrics enhance our understanding of how generative AI influences user experience across the stages of ESCs, supporting the optimization of GenAI system design and improving user interaction quality.

This study highlights the substantial influence of intrinsic and extrinsic evaluation metrics on consultation feedback across various stages of ESCs. Firstly, during the Exploration stage of Hill's three-stage helping skills model, where counselors identify issues through questioning and discussion, we identified positively correlated intrinsic metrics such as the use of question marks and various punctuation marks. Frequent use of question marks and punctuation marks correlates with positive consultation feedback, likely due to their role in facilitating inquiry and open-ended questions, which encourage deeper dialogue[103]. Conversely, negatively correlated intrinsic metrics include expressions of feelings, gender descriptions, social conflicts, mentions of ethnic groups, and negative emotions. This suggests that during this stage, exploring emotional and social issues in depth may lead to discomfort or barriers [104]. Secondly, in the Comforting stage, positive user-perceived feedback was associated with the use of more verbs and positive-toned language, reflecting a more supportive and encouraging communication style[105]. In contrast, frequent use of pronouns, numerals, adverbs, and expressions of estrangement or abstract language was linked to negative feedback. This indicates that personalized and specific support is more effective during this stag[106]. Finally, in the Action stage, positive indicators like conflict in social processes and extensive punctuation suggest active engagement and problem-solving, which are associated with effective conflict management in consultations[107]. Conversely, negative indicators, such as an emphasis on feelings, social issues, and negations, may hinder the implementation and progress of action plans by concentrating on negative emotions and social problems[103]. Understanding these relationships enhances our knowledge of the dynamics and outcomes of the ESCs. It offers valuable insights for optimizing GenAI applications in emotional support systems, potentially significantly improving user experience and effectiveness.

5.1.3 The Role of CoT Prompt Engineering Combined with Human Expertise in Enhancing LLM Performance in ESCs

A notable limitation of LLMs is their response uncertainty[108], which underscores their adaptability. Prompt engineering has emerged as a critical method for enhancing the performance of GenAI in specific domains. Utilizing the integrative modeling paradigm of computational social science, we developed an explainable model for user-perceived feedback scores and identified pertinent intrinsic metrics. Furthermore, by applying a framework grounded in Hill's three-stage helping skills model and an integrated evaluation metrics framework, we introduced manually customized CoT prompts, collected responses from GPT-4o, and assessed their effectiveness.

Firstly, concerning extrinsic evaluation metrics—perceived feedback scores—the experiments revealed that customized CoT prompts substantially improved GPT-4o's user feedback in ESCs. This approach enhances both the efficiency and effectiveness of prompt design. This comprehensive method demonstrates significant versatility and broad applicability, providing precise and effective solutions in mental health, medical diagnostics, intelligent customer service, and beyond[109]. This advancement supports the broader adoption and development of GenAI.

Secondly, we evaluated users' UPF scores related to responses from GPT-4o using Auto-CoT prompts, manually customized CoT prompts, and human counselors across various intrinsic metrics. In the exploration stage, the prompted GPT-4o, compared to GPT-4o using Auto-CoT prompts, demonstrated better performance in cognitive processing and cultural sensitivity, reflecting better social and cognitive performance. However, it may overemphasize emotional content. Compared to human counselors, the prompted GPT-4o showed advantages in cultural sensitivity and some aspects of cognitive processing but fell short in emotional regulation and cognitive depth. Nonetheless, it may still overly focus on emotions, and its grammatical proficiency have not yet reached human levels, highlighting limitations in comprehending and articulating complex emotions and intricate cognitive relationships[110].

In the comforting stage, the prompted GPT-4o showed enhancements over the GPT-4o using Auto-CoT prompts in emotional communication, structural organization, and cognitive processing abilities. However, it has shortcomings in questioning and the use of descriptive language. Compared to human counselors, the prompted GPT-4o improved in emotional communication and empathy but lagged behind in cognitive processing and structural expression while also overemphasizing emotional content. Its higher dictionary coverage might reflect a more technical or less conversational style, resulting in interactions that are less engaging compared to those of human counselors.

In the action stage, the prompted GPT-4o displayed better emotional control, cultural sensitivity, and client focus compared to the standard GPT-4o, with more mature cognitive processing. However, it may overemphasize emotional content and neglect in-depth understanding of client needs, leading to reduced interactivity. When compared to human counselors, the prompted GPT-4o showed improvements in emotional control, cultural sensitivity, and time management, and was more client-focused. Nevertheless, it tended to overemphasize emotional content, and its employment of fewer question marks may diminish clarity and engagement, indicating that its communication clarity and detail require enhancement.

In summary, the prompted GPT-4o demonstrated certain advantages at each stage, such as improvements in emotional expression, cultural sensitivity, and cognitive processing abilities. However, compared to the standard GPT-4o and human counselors, it still exhibits some shortcomings, such as overemphasis on emotional content, needing improvement in questioning techniques and grammatical proficiency, and requiring enhanced in-depth understanding of client needs and interactivity. These issues may reduce its effectiveness in conveying complex emotions and engaging users interactively, revealing limitations in handling more intricate scenarios[108]. These findings provide important insights for further optimizing the emotional support capabilities of the prompted GPT-4o.

5.2 Limitations and Future Research

This study evaluated the capacity of GPT-4o to provide emotional support, achieving a relatively low mean absolute percentage error (MAPE = 17.13%) with predictive models, which indicates strong predictive performance. Although the automatic evaluation model circumvents ethical risks associated with manual assessment, it fails to fully capture the perceived feedback from help-seekers. Future research should incorporate manual evaluation methods, ensuring no harm to participants, to provide a more comprehensive assessment of generative AI's emotional support capabilities and to address potential ethical risks.

Moreover, the study acknowledges GPT-4o's limitations in handling prolonged conversations and extreme emotional states, primarily due to insufficient context retention and emotional comprehension. These shortcomings underscore the need for models with enhanced contextual understanding and emotional intelligence [111,112]. Future research should explore advanced models, such as ChatGPT-o1 [113], which may enhance logical reasoning and integrate domain-specific data alongside fine-tuning techniques to improve AI-driven mental health support. This approach would not only address current deficiencies but also pave the way for more specialized interventions in mental health.

Furthermore, the study emphasizes the importance of help-seekers' perceptions in evaluating the quality of emotional support. While feedback from help-seekers was utilized to assess effectiveness, it is recognized that the dimensions of emotional support—such as empathy, therapeutic alliance, and emotional changes before and after counseling—are broader and more complex than what current data can capture [114,115]. Therefore, future research should collect more comprehensive data to provide robust evidence for the application of large language models in mental health services.

Additionally, despite relying heavily on machine learning models for evaluation and prediction, the transformer model trained on human emotional support conversation data yielded less satisfactory results [60]. Future work should investigate more tailored transformer models, pre-trained on both human and AI-generated emotional support dialogues, to enhance assessment accuracy and enable more precise evaluations using advanced transformer architectures.

While this study provides valuable insights into emotional support and empathy, it is important to acknowledge that various sociocultural contexts—including underrepresented communities and minorities—can significantly influence the dynamics of emotional support and the expression of empathy [116–118]. The current dataset may not fully capture the diversity inherent in these contexts, presenting a limitation to the generalizability of our findings. To address this limitation, future efforts should focus on globally expanding the dataset to incorporate emotional support data that includes participants from a wide range of gender identities, ethnicities, ages, and countries.

Finally, it is essential to address several ethical considerations before deploying LLMs in emotional support contexts. These include concerns about data privacy, algorithmic bias, discrimination, transparency, and interpretability, as well as the absence of adequate ethical and legal frameworks [119]. Although our research made strides in improving transparency and interpretability through human-AI collaboration, several critical ethical challenges remain unresolved. These include ensuring that real-world data comply with privacy regulations and are properly protected during both collection and processing [120], managing inherent biases within the data to prevent algorithms from amplifying these biases during analysis [60,121,122], and fostering transparency to build trust among stakeholders regarding the research process and outcomes [121]. As generative AI technology evolves rapidly, there is a pressing need to continuously update and refine relevant ethical and legal frameworks [119]. Addressing these challenges is crucial for the safe and effective implementation of research findings in real-world applications.

6. Conclusion

This study developed and utilized an AI-based evaluation model, drawing on real ESCs datasets, to propose an integrative evaluation framework for user-perceived feedback in ESCs. We assessed and compared the performance of GPT-4o in emotional support tasks with that of human counselors. The results indicate that GPT-4o's emotional support capabilities improved with CoT prompting, but it still lagged behind experienced human counselors in certain emotional domains. By employing an integrative modeling paradigm and Hill’s three-stage helping skills model within computational social science, we identified intrinsic evaluation metrics relevant to user experience in ESCs. We developed customized CoT prompt engineering, significantly improving GPT-4o's user-perceived feedback scores. This suggests that while AI has made strides in emotional support, incorporating human expertise through CoT prompts can significantly refine and enhance the capabilities of LLMs. Our findings provide a crucial empirical foundation and direction for improving genAI applications in the field of emotional support services.

Acknowledgements

This research was supported by funding from the National Natural Science Foundation of China (Award Number: 72204095), the Humanities and Social Science Young Scientist Program sponsored by the Ministry of Education of the People's Republic of China (Award Number: 22YJC880022), the Wuhan University of Technology Autonomous Innovation Research Fund (Award Number: 2024VA059), and the China National Center for Mental Health and Prevention, China Education Development Foundation, Ministry of Education Student Service and Quality Development Center (Award Number: XS24A010). None of the funders had any involvement in carrying out this research.

Contributions

All authors contributed to the conceptualization and methodology of the study. Yinghui Huang conducted the study design. The literature review was performed by Hui Liu, Wanghao Dong, and Yinghui Huang. Lie Li and Yinghui Huang carried out data analysis. Writing was done by Yinghui Huang, Hui Liu, Yuhang Dong, and Lie Li. Manuscript revision was handled by Hui Liu, Yingdan Huang. Yinghui Huang and Yingdan Huang managed funding acquisition. All authors participated in the review and editing of the draft.

Abbreviations

AI: Artificial Intelligence

BERT: Bidirectional Encoder Representations from Transformers

BiLSTM: Bidirectional Long Short-Term Memory

CoT: Chain of Thought

ESC: Emotional Support Conversations

GenAI: Generative Artificial Intelligence

LIWC: Linguistic Inquiry and Word Count

LLM: Large Language Model

LSM: Language Style Matching

MAPE: Mean Absolute Percentage Error

PF: Perceived Feedback

RFE-CV: Recursive Feature Elimination with Cross-Validation

RQ1: The first research question

RQ2: The second research question

RQ3: The third research question

RMSE: Root Mean Square Error

RoBERTa: Robustly Optimized BERT Pretraining Approach

SHAP: Shapley Additive Explanations

SVR: Support Vector Regression

XGBoost: Extreme Gradient Boosting

XLNet: Extreme Language Net

XAI: Explainable Artificial Intelligence

Multimedia Appendix 1

See https://jmir.zendesk.com/hc/en-us/articles/115003396688 for further information.

References

1. World mental health report: Transforming mental health for all. Accessed August 20, 2024. https://www.who.int/publications/i/item/9789240049338

2. Burleson BR. The experience and effects of emotional support: What the study of cultural and

gender differences can tell us about close relationships, emotion, and interpersonal communication. *Pers Relatsh*. 2003;10(1):1-23. doi:10.1111/1475-6811.00033

3. Reblin M, Uchino BN. Social and emotional support and its implication for health. *Curr Opin Psychiatry*. 2008;21(2):201-205. doi:10.1097/YCO.0b013e3282f3ad89
4. Strine TW, Chapman DP, Balluz L, Mokdad AH. Health-related quality of life and health behaviors by social and emotional support. *Soc Psychiatry Psychiatr Epidemiol*. 2008;43(2):151-159. doi:10.1007/s00127-007-0277-x
5. Gonzales FA, Hurtado-de-Mendoza A, Santoyo-Olsson J, Nápoles AM. Do coping strategies mediate the effects of emotional support on emotional well-being among Spanish-speaking Latina breast cancer survivors? *Psychooncology*. 2016;25(11):1286-1292. doi:10.1002/pon.3953
6. Wang W, Shukla P, Shi G. Digitalized social support in the healthcare environment: Effects of the types and sources of social support on psychological well-being. *Technol Forecast Soc Change*. 2021;164:120503. doi:10.1016/j.techfore.2020.120503
7. Priem JS, Solomon DH. Emotional Support and Physiological Stress Recovery: The Role of Support Matching, Adequacy, and Invisibility. *Commun Monogr*. 2015;82(1):88-112. doi:10.1080/03637751.2014.971416
8. Baltes BB, Dickson MW, Sherman MP, Bauer CC, LaGanke JS. Computer-Mediated Communication and Group Decision Making: A Meta-Analysis. *Organ Behav Hum Decis Process*. 2002;87(1):156-179. doi:10.1006/obhd.2001.2961
9. Liu S, Zheng C, Demasi O, et al. Towards Emotional Support Dialog Systems. Published online June 2, 2021. Accessed August 20, 2024. http://arxiv.org/abs/2106.01144
10. Serban I, Sordoni A, Bengio Y, Courville A, Pineau J. Building End-To-End Dialogue Systems Using Generative Hierarchical Neural Network Models. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol 30. ; 2016. doi:10.1609/aaai.v30i1.9883
11. OpenAI, Achiam J, Adler S, et al. GPT-4 Technical Report. Published online March 4, 2024. Accessed August 20, 2024. http://arxiv.org/abs/2303.08774
12. Patil R, Gudivada V. A Review of Current Trends, Techniques, and Challenges in Large Language Models (LLMs). *Appl Sci*. 2024;14(5):2074. doi:10.3390/app14052074
13. Zhou J, Chen Z, Wang B, Huang M. Facilitating Multi-turn Emotional Support Conversation with Positive Emotion Elicitation: A Reinforcement Learning Approach. Published online July 16, 2023. Accessed August 20, 2024. http://arxiv.org/abs/2307.07994
14. Narimisaei J, Naeim M, Imannezhad S, Samian P, Sobhani M. Exploring emotional intelligence in artificial intelligence systems: a comprehensive analysis of emotion recognition and response mechanisms. *Ann Med Surg*. 2024;86(8):4657-4663. doi:10.1097/MS9.0000000000002315
15. Hofman JM, Watts DJ, Athey S, et al. Integrating explanation and prediction in computational social science. *Nature*. 2021;595(7866):181-188. doi:10.1038/s41586-021-03659-0
16. Berger J, Milkman KL. What Makes Online Content Viral? *J Mark Res*. 2012;49(2):192-205. doi:10.1509/jmr.10.0353
17. Kallivalappil N, D'souza K, Deshmukh A, Kadam C, Sharma N. Empath.ai: a Context-Aware Chatbot for Emotional Detection and Support. In: *2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT)*. IEEE; 2023:1-7. doi:10.1109/ICCCNT56998.2023.10306584
18. Kojima T, Gu SS, Reid M, Matsuo Y, Iwasawa Y. Large language models are zero-shot reasoners. *Adv Neural Inf Process Syst*. 2022;35:22199-22213.
19. Bodie GD, Burleson BR. Explaining Variations in the Effects of Supportive Messages A Dual-Process Framework. *Ann Int Commun Assoc*. 2008;32(1):355-398. doi:10.1080/23808985.2008.11679082
20. Burleson BR, Hanasono LK, Bodie GD, et al. Explaining Gender Differences in Responses to Supportive Messages: Two Tests of a Dual-Process Approach. *Sex Roles*. 2009;61(3-4):265-280. doi:10.1007/s11199-009-9623-7

21. Hill CE. *Helping Skills: Facilitating Exploration, Insight, and Action*. American Psychological Association; 2020. Accessed August 20, 2024. https://psycnet.apa.org/record/2019-44086-000

22. Song Z, Zheng X, Liu L, Xu M, Huang X. Generating Responses with a Specific Emotion in Dialog. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. ; 2019:3685-3695. doi:10.18653/v1/P19-1359

23. Zhou H, Huang M, Zhang T, Zhu X, Liu B. Emotional Chatting Machine: Emotional Conversation Generation with Internal and External Memory. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol 32. ; 2018. doi:10.1609/aaai.v32i1.11325

24. Rashkin H, Smith EM, Li M, Boureau YL. Towards Empathetic Open-domain Conversation Models: a New Benchmark and Dataset. Published online August 28, 2019. Accessed August 20, 2024. http://arxiv.org/abs/1811.00207

25. Majumder N, Hong P, Peng S, et al. MIME: MIMicking Emotions for Empathetic Response Generation. Published online October 3, 2020. Accessed August 20, 2024. http://arxiv.org/abs/2010.01454

26. Roller S, Dinan E, Goyal N, et al. Recipes for Building an Open-Domain Chatbot. In: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Association for Computational Linguistics; 2021. doi:10.18653/v1/2021.eacl-main.24

27. Tu Q, Li Y, Cui J, Wang B, Wen JR, Yan R. MISC: A Mixed Strategy-Aware Model integrating COMET for Emotional Support Conversation. In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics; 2022:308-319. doi:10.18653/v1/2022.acl-long.25

28. Peng W, Hu Y, Xing L, Xie Y, Sun Y, Li Y. Control Globally, Understand Locally: A Global-to-Local Hierarchical Graph Network for Emotional Support Conversation. *Proc Thirty-First Int Jt Conf Artif Intell*. Published online July 2022:4324-4330. doi:10.24963/ijcai.2022/600

29. Deng Y, Zhang W, Yuan Y, Lam W. Knowledge-enhanced Mixed-initiative Dialogue System for Emotional Support Conversations. Published online May 17, 2023. Accessed August 20, 2024. http://arxiv.org/abs/2305.10172

30. Tu Q, Chen C, Li J, et al. CharacterChat: Learning towards Conversational AI with Personalized Social Support. Published online August 20, 2023. Accessed August 20, 2024. http://arxiv.org/abs/2308.10278

31. Xu X, Meng X, Wang Y. PoKE: Prior Knowledge Enhanced Emotional Support Conversation with Latent Variable. Published online February 15, 2023. Accessed August 20, 2024. http://arxiv.org/abs/2210.12640

32. Eysenbach G. The role of ChatGPT, generative language models, and artificial intelligence in medical education: a conversation with ChatGPT and a call for papers. *JMIR Med Educ*. 2023;9(1):e46885.

33. Van Dis EA, Bollen J, Zuidema W, Van Rooij R, Bockting CL. ChatGPT: five priorities for research. *Nature*. 2023;614(7947):224-226. doi:10.1038/d41586-023-00288-7

34. Hello GPT-4o. Accessed September 25, 2024. https://openai.com/index/hello-gpt-4o/

35. Prabhod KJ. AI-Driven Insights from Large Language Models: Implementing Retrieval-Augmented Generation for Enhanced Data Analytics and Decision Support in Business Intelligence Systems. *J Artif Intell Res*. 2023;3(2):1-58.

36. Laranjo L, Dunn AG, Tong HL, et al. Conversational agents in healthcare: a systematic review. *J Am Med Inform Assoc*. 2018;25(9):1248-1258. doi:10.1093/jamia/ocy072

37. Ji M, Genchev GZ, Huang H, Xu T, Lu H, Yu G. Evaluation Framework for Successful Artificial Intelligence–Enabled Clinical Decision Support Systems: Mixed Methods Study. *J Med Internet Res*. 2021;23(6):e25929. doi:10.2196/25929

38. Antoniadi AM, Du Y, Guendouz Y, et al. Current Challenges and Future Opportunities for

XAI in Machine Learning-Based Clinical Decision Support Systems: A Systematic Review. *Appl Sci*. 2021;11(11):5088. doi:10.3390/app11115088
39. Wójcik S, Rulkiewicz A, Pruszczyk P, Lisik W, Poboży M, Domienik-Karłowicz J. Beyond ChatGPT: What does GPT-4 add to healthcare? The dawn of a new era. *Cardiol J*. 2023;30(6):1018-1025.
40. Knapič S, Malhi A, Saluja R, Främling K. Explainable Artificial Intelligence for Human Decision Support System in the Medical Domain. *Mach Learn Knowl Extr*. 2021;3(3):740-770. doi:10.3390/make3030037
41. Denecke K, Abd-Alrazaq A, Househ M. Artificial Intelligence for Chatbots in Mental Health: Opportunities and Challenges. In: Househ M, Borycki E, Kushniruk A, eds. *Multiple Perspectives on Artificial Intelligence in Healthcare*. Lecture Notes in Bioengineering. Springer International Publishing; 2021:115-128. doi:10.1007/978-3-030-67303-1_10
42. Koutsouleris N, Hauser TU, Skvortsova V, De Choudhury M. From promise to practice: towards the realisation of AI-informed mental health care. *Lancet Digit Health*. 2022;4(11):e829-e840. doi:10.1016/S2589-7500(22)00153-4
43. Guțu SM, Cosmoiu A, Cojocaru D, Turturescu T, Popoviciu CM, Giosan C. Bot to the rescue? Effects of a fully automated conversational agent on anxiety and depression: a randomized controlled trial. *Ann Depress Anxiety*. 2021;8(1):1107.
44. Rüsch N, Corrigan PW, Powell K, et al. A stress-coping model of mental illness stigma: II. Emotional stress responses, coping behavior and outcome. *Schizophr Res*. 2009;110(1-3):65-71.
45. Goldberg SB, Flemotomos N, Martinez VR, et al. Machine learning and natural language processing in psychotherapy research: Alliance as example use case. *J Couns Psychol*. 2020;67(4):438-448. doi:10.1037/cou0000382
46. Alanezi F. Assessing the Effectiveness of ChatGPT in Delivering Mental Health Support: A Qualitative Study. *J Multidiscip Healthc*. 2024;Volume 17:461-471. doi:10.2147/JMDH.S447368
47. Sezgin E. Artificial intelligence in healthcare: Complementing, not replacing, doctors and healthcare providers. *Digit Health*. 2023;9:20552076231186520. doi:10.1177/20552076231186520
48. Liu P, Yuan W, Fu J, Jiang Z, Hayashi H, Neubig G. Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing. *ACM Comput Surv*. 2023;55(9):1-35. doi:10.1145/3560815
49. Marvin G, Hellen N, Jjingo D, Nakatumba-Nabende J. Prompt Engineering in Large Language Models. In: Jacob IJ, Piramuthu S, Falkowski-Gilski P, eds. *Data Intelligence and Cognitive Informatics*. Algorithms for Intelligent Systems. Springer Nature Singapore; 2024:387-402. doi:10.1007/978-981-99-7962-2_30
50. Zhang Z, Zhang A, Li M, Smola A. Automatic Chain of Thought Prompting in Large Language Models. Published online October 7, 2022. Accessed August 20, 2024. http://arxiv.org/abs/2210.03493
51. Pataranutaporn P, Liu R, Finn E, Maes P. Influencing human–AI interaction by priming beliefs about AI can increase perceived trustworthiness, empathy and effectiveness. *Nat Mach Intell*. 2023;5(10):1076-1086. doi:10.1038/s42256-023-00720-7
52. Zhou L, Gao J, Li D, Shum HY. The Design and Implementation of XiaoIce, an Empathetic Social Chatbot. *Comput Linguist*. 2020;46(1):53-93. doi:10.1162/coli_a_00368
53. Picard RW. Affective computing: challenges. *Int J Hum-Comput Stud*. 2003;59(1-2):55-64. doi:10.1016/S1071-5819(03)00052-1
54. Salovey P, Mayer JD. Emotional Intelligence. *Imagin Cogn Personal*. 1990;9(3):185-211. doi:10.2190/DUGG-P24E-52WK-6CDG
55. Thomas KC, Ellis AR, Konrad TR, Holzer CE, Morrissey JP. County-level estimates of mental health professional shortage in the United States. *Psychiatr Serv*. 2009;60(10):1323-1328. doi:10.1176/ps.2009.60.10.1323
56. Kretzschmar K, Tyroll H, Pavarini G, Manzini A, Singh I, NeurOx Young People's Advisory

Group. Can Your Phone Be Your Therapist? Young People's Ethical Perspectives on the Use of Fully Automated Conversational Agents (Chatbots) in Mental Health Support. *Biomed Inform Insights*. 2019;11:117822261982908. doi:10.1177/1178222619829083

57. Zhang Z, Liao L, Huang M, Zhu X, Chua TS. Neural Multimodal Belief Tracker with Adaptive Attention for Dialogue Systems. In: *The World Wide Web Conference*. ACM; 2019:2401-2412. doi:10.1145/3308558.3313598

58. Abd-Alrazaq AA, Alajlani M, Ali N, Denecke K, Bewick BM, Househ M. Perceptions and Opinions of Patients About Mental Health Chatbots: Scoping Review. *J Med Internet Res*. 2021;23(1):e17828. doi:10.2196/17828

59. Coghlan S, Leins K, Sheldrick S, Cheong M, Gooding P, D'Alfonso S. To chat or bot to chat: Ethical issues with using chatbots in mental health. *Digit Health*. 2023;9:20552076231183542. doi:10.1177/20552076231183542

60. Abbasian M, Khatibi E, Azimi I, et al. Foundation metrics for evaluating effectiveness of healthcare conversations powered by generative AI. *Npj Digit Med*. 2024;7(1):82. doi:10.1038/s41746-024-01074-z

61. Abd-Alrazaq AA, Rababeh A, Alajlani M, Bewick BM, Househ M. Effectiveness and Safety of Using Chatbots to Improve Mental Health: Systematic Review and Meta-Analysis. *J Med Internet Res*. 2020;22(7):e16021. doi:10.2196/16021

62. Vaidyam AN, Wisniewski H, Halamka JD, Kashavan MS, Torous JB. Chatbots and Conversational Agents in Mental Health: A Review of the Psychiatric Landscape. *Can J Psychiatry*. 2019;64(7):456-464. doi:10.1177/0706743719828977

63. Resnik P, Niv M, Nossal M, et al. Using intrinsic and extrinsic metrics to evaluate accuracy and facilitation in computer-assisted coding. In: *Perspectives in Health Information Management Computer Assisted Coding Conference Proceedings*. Vol 2006. ; 2006:2006.

64. Chang Y, Wang X, Wang J, et al. A Survey on Evaluation of Large Language Models. *ACM Trans Intell Syst Technol*. 2024;15(3):1-45. doi:10.1145/3641289

65. Bommasani R, Liang P, Lee T. Holistic Evaluation of Language Models. *Ann N Y Acad Sci*. 2023;1525(1):140-146. doi:10.1111/nyas.15007

66. Aggarwal CC, Hinneburg A, Keim DA. On the Surprising Behavior of Distance Metrics in High Dimensional Space. In: Van Den Bussche J, Vianu V, eds. *Database Theory — ICDT 2001*. Vol 1973. Lecture Notes in Computer Science. Springer Berlin Heidelberg; 2001:420-434. doi:10.1007/3-540-44503-X_27

67. Rafaeli E, Gleason MEJ. Skilled Support Within Intimate Relationships. *J Fam Theory Rev*. 2009;1(1):20-37. doi:10.1111/j.1756-2589.2009.00003.x

68. Olawade DB, Wada OZ, Odetayo A, David-Olawade AC, Asaolu F, Eberhardt J. Enhancing mental health with Artificial Intelligence: Current trends and future prospects. *J Med Surg Public Health*. Published online 2024:100099.

69. Aafjes-van Doorn K, Müller-Frommeyer L. Reciprocal language style matching in psychotherapy research. *Couns Psychother Res*. 2020;20(3):449-455. doi:10.1002/capr.12298

70. Bennemann B, Schwartz B, Giesemann J, Lutz W. Predicting patients who will drop out of out-patient psychotherapy using machine learning algorithms. *Br J Psychiatry*. 2022;220(4):192-201. doi:10.1192/bjp.2022.17

71. Ewbank MP, Cummins R, Tablan V, et al. Quantifying the Association Between Psychotherapy Content and Clinical Outcomes Using Deep Learning. *JAMA Psychiatry*. 2020;77(1):35. doi:10.1001/jamapsychiatry.2019.2664

72. Squarcina L, Villa FM, Nobile M, Grisan E, Brambilla P. Deep learning for the prediction of treatment response in depression. *J Affect Disord*. 2021;281:618-622. doi:10.1016/j.jad.2020.11.104

73. Barredo Arrieta A, Díaz-Rodríguez N, Del Ser J, et al. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf Fusion*. 2020;58:82-115. doi:10.1016/j.inffus.2019.12.012

74. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. *Adv Neural Inf Process Syst*. 2017;30. Accessed August 20, 2024. https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html

75. Datta A, Sen S, Zick Y. Algorithmic Transparency via Quantitative Input Influence: Theory and Experiments with Learning Systems. In: *2016 IEEE Symposium on Security and Privacy (SP)*. IEEE; 2016:598-617. doi:10.1109/SP.2016.42

76. Henelius A, Puolamäki K, Ukkonen A. Interpreting Classifiers through Attribute Interactions in Datasets. Kim B, Malioutov D, Varshney K, Weller A, eds. *Proc 2017 ICML Workshop Hum Interpret Mach Learn WHI 2017*. Published online 2017.

77. Pennebaker JW, Boyd RL, Jordan K, Blackburn K. The development and psychometric properties of LIWC2015. Published online 2015. Accessed August 20, 2024. https://repositories.lib.utexas.edu/items/705e81ca-940d-4c46-94ec-a52ffdc3b51f

78. Gross JJ. Emotion regulation: Conceptual and empirical foundations. *Handb Emot Regul*. 2014;2:3-20.

79. Ruppel EK, Gross C, Stoll A, Peck BS, Allen M, Kim SY. Reflecting on Connecting: Meta-Analysis of Differences Between Computer-Mediated and Face-to-Face Self-Disclosure. *J Comput-Mediat Commun*. 2017;22(1):18-34. doi:10.1111/jcc4.12179

80. Lee J, Lee D, Lee J gil. Influence of Rapport and Social Presence with an AI Psychotherapy Chatbot on Users' Self-Disclosure. *Int J Human–Computer Interact*. 2024;40(7):1620-1631. doi:10.1080/10447318.2022.2146227

81. Ireland ME, Pennebaker JW. Language style matching in writing: Synchrony in essays, correspondence, and poetry. *J Pers Soc Psychol*. 2017;99(3):549-571. doi:10.1037/a0020386

82. Lord SP, Sheng E, Imel ZE, Baer J, Atkins DC. More Than Reflections: Empathy in Motivational Interviewing Includes Language Style Synchrony Between Therapist and Client. *Behav Ther*. 2015;46(3):296-303. doi:10.1016/j.beth.2014.11.002

83. Gaut G, Steyvers M, Imel ZE, et al. Content Coding of Psychotherapy Transcripts Using Labeled Topic Models. *IEEE J Biomed Health Inform*. 2017;21(2):476-487. doi:10.1109/JBHI.2015.2503985

84. Fiske A, Henningsen P, Buyx A. Your Robot Therapist Will See You Now: Ethical Implications of Embodied Artificial Intelligence in Psychiatry, Psychology, and Psychotherapy. *J Med Internet Res*. 2019;21(5):e13216. doi:10.2196/13216

85. Fulmer R, Joerin A, Gentile B, Lakerink L, Rauws M. Using Psychological Artificial Intelligence (Tess) to Relieve Symptoms of Depression and Anxiety: Randomized Controlled Trial. *JMIR Ment Health*. 2018;5(4):e64. doi:10.2196/mental.9782

86. Joyce DW, Kormilitzin A, Smith KA, Cipriani A. Explainable artificial intelligence for mental health through transparency and interpretability for understandability. *Npj Digit Med*. 2023;6(1):6. doi:10.1038/s41746-023-00751-9

87. Miner AS, Milstein A, Schueller S, Hegde R, Mangurian C, Linos E. Smartphone-Based Conversational Agents and Responses to Questions About Mental Health, Interpersonal Violence, and Physical Health. *JAMA Intern Med*. 2016;176(5):619. doi:10.1001/jamainternmed.2016.0400

88. To explain or not to explain?—Artificial intelligence explainability in clinical decision support systems - PMC. Accessed August 20, 2024. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9931364/

89. Kuhn M. Applied Predictive Modeling. Published online 2013.

90. Young T, Hazarika D, Poria S, Cambria E. Recent trends in deep learning based natural language processing. *Ieee Comput Intell Mag*. 2018;13(3):55-75. doi:10.1109/MCI.2018.2840738

91. Liu Y, Ott M, Goyal N, et al. RoBERTa: A Robustly Optimized BERT Pretraining Approach. Published online July 26, 2019. Accessed August 21, 2024. http://arxiv.org/abs/1907.11692

92. Yang Z, Dai Z, Yang Y, Carbonell J, Salakhutdinov RR, Le QV. Xlnet: Generalized autoregressive pretraining for language understanding. *Adv Neural Inf Process Syst*. 2019;32.

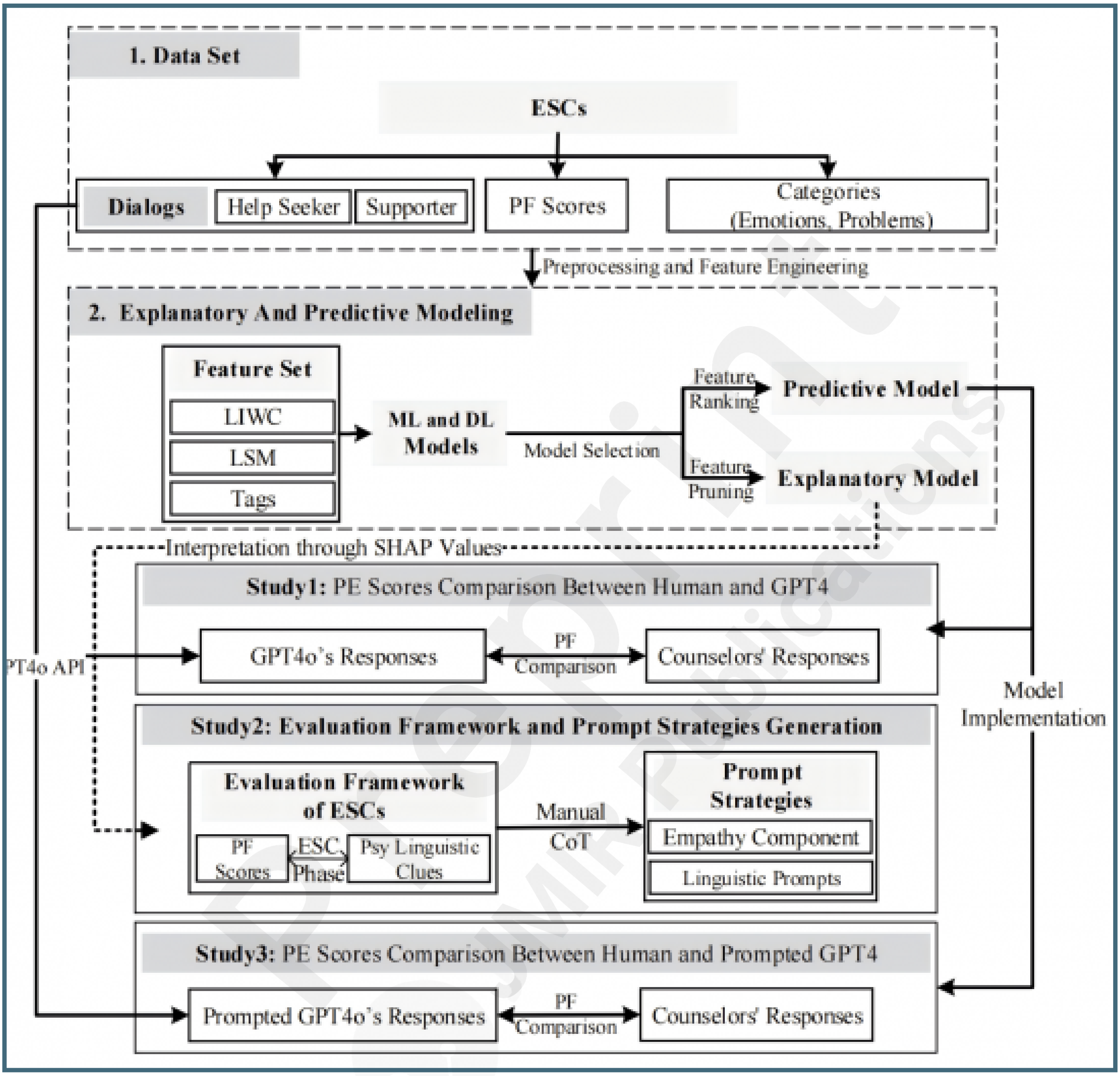
Accessed August 21, 2024. https://proceedings.neurips.cc/paper/2019/hash/dc6a7e655d7e5840e66733e9ee67cc69-Abstract.html
93. Thakur A. The Art of Prompting: Unleashing the Power of Large Language Models. Accessed August 20, 2024. https://www.researchgate.net/profile/Ayush-Thakur-9/publication/379044941_The_Art_of_Prompting_Unleashing_the_Power_of_Large_Language_Models/links/65f85d0e32321b2cff8c3104/The-Art-of-Prompting-Unleashing-the-Power-of-Large-Language-Models.pdf
94. Practical Nonparametric Statistics, 3rd Edition | Wiley. Wiley.com. Accessed September 23, 2024. https://www.wiley.com/en-us/Practical+Nonparametric+Statistics%2C+3rd+Edition-p-9780471160687
95. Macbeth G, Razumiejczyk E, Ledesma RD. Cliff's Delta Calculator: A non-parametric effect size program for two groups of observations. *Univ Psychol*. 2011;10(2):545-555. doi:10.11144/Javeriana.upsy10-2.cdcp
96. Boyd RL, Ashokkumar A, Seraj S, Pennebaker JW. The development and psychometric properties of LIWC-22. *Austin TX Univ Tex Austin*. 2022;10. Accessed August 20, 2024. https://www.researchgate.net/profile/Ryan-Boyd-8/publication/358725479_The_Development_and_Psychometric_Properties_of_LIWC-22/links/6210f62c4be28e145ca1e60b/The-Development-and-Psychometric-Properties-of-LIWC-22.pdf
97. Ouyang L, Wu J, Jiang X, et al. Training language models to follow instructions with human feedback. *Adv Neural Inf Process Syst*. 2022;35:27730-27744.
98. Syah TA, Apriyanto S, Nurhayaty A. Student's prevailing, confidence, and drives: LIWC analysis on self-description text. In: *1st International Conference on Science, Health, Economics, Education and Technology (ICoSHEET 2019)*. Atlantis Press; 2020:295-299. Accessed August 20, 2024. https://www.atlantis-press.com/proceedings/icosheet-19/125942052
99. Syah TA, Nurhayaty A, Apriyanto S. Computerized Text Analysis on Self-Description Text to Get Student's Prevailing, Confidence, and Drives. In: *Journal of Physics: Conference Series*. Vol 1764. IOP Publishing; 2021:012056. doi:10.1088/1742-6596/1764/1/012056
100. Lyu S, Ren X, Du Y, Zhao N. Detecting depression of Chinese microblog users via text analysis: Combining Linguistic Inquiry Word Count (LIWC) with culture and suicide related lexicons. *Front Psychiatry*. 2023;14:1121583. doi:10.3389/fpsyt.2023.1121583
101. Marengo D, Azucar D, Longobardi C, Settanni M. Mining Facebook data for Quality of Life assessment. *Behav Inf Technol*. 2021;40(6):597-607. doi:10.1080/0144929X.2019.1711454
102. Biggiogera J, Boateng G, Hilpert P, et al. BERT meets LIWC: Exploring State-of-the-Art Language Models for Predicting Communication Behavior in Couples' Conflict Interactions. In: *Companion Publication of the 2021 International Conference on Multimodal Interaction*. ACM; 2021:385-389. doi:10.1145/3461615.3485423
103. Oliveira DP, Klinger EF, Rodrigues GA, et al. Psychological Counseling in Contemporaneity: A Psychoanalytic Perspective. *Int Neuropsychiatr Dis J*. 2020;14(2):36-41. doi:10.9734/indj/2020/v14i230127
104. Irvine A, Drew P, Bower P, et al. Are there interactional differences between telephone and face-to-face psychological therapy? A systematic review of comparative studies. *J Affect Disord*. 2020;265:120-131. doi:10.1016/j.jad.2020.01.057
105. Dube L, Nkosi-Mafutha N, Balsom AA, Gordon JL. Infertility-related distress and clinical targets for psychotherapy: a qualitative study. *BMJ Open*. 2021;11(11):e050373. doi:10.1136/bmjopen-2021-050373
106. Singh AA, Appling B, Trepal H. Using the Multicultural and Social Justice Counseling Competencies to Decolonize Counseling Practice: The Important Roles of Theory, Power, and Action. *J Couns Dev*. 2020;98(3):261-271. doi:10.1002/jcad.12321
107. Bek H, Gülveren H. Determination of Psychological Counsellor Candidates' Competency Levels and Educational Needs in terms of Therapeutic Conditions in the Process of Individual

Counselling. *Educ Q Rev*. 2021;4(3). doi:10.31014/aior.1993.04.03.364
108. Shiffrin R, Mitchell M. Probing the psychology of AI models. *Proc Natl Acad Sci*. 2023;120(10):e2300963120. doi:10.1073/pnas.2300963120
109. Qiu C, Xie Z, Liu M, Hu H. Explainable Knowledge reasoning via thought chains for knowledge-based visual question answering. *Inf Process Manag*. 2024;61(4):103726. doi:10.1016/j.ipm.2024.103726
110. Hager P, Jungmann F, Holland R, et al. Evaluation and mitigation of the limitations of large language models in clinical decision-making. *Nat Med*. Published online 2024:1-10.
111. Verma M, Bhambri S, Kambhampati S. Theory of Mind Abilities of Large Language Models in Human-Robot Interaction: An Illusion? In: *Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*. ACM; 2024:36-45. doi:10.1145/3610978.3640767
112. Strachan JW, Albergo D, Borghini G, et al. Testing theory of mind in large language models and humans. *Nat Hum Behav*. Published online 2024:1-11.
113. Introducing OpenAI o1. 2024. Accessed September 17, 2024. https://openai.com/index/introducing-openai-o1-preview/
114. Elliott R, Bohart AC, Watson JC, Greenberg LS. Empathy. *Psychotherapy*. 2011;48(1):43-49. doi:10.1037/a0022187
115. Ardito RB, Rabellino D. Therapeutic Alliance and Outcome of Psychotherapy: Historical Excursus, Measurements, and Prospects for Research. *Front Psychol*. 2011;2:270. doi:10.3389/fpsyg.2011.00270
116. Cauce AM, Domenech-Rodríguez M, Paradise M, et al. Cultural and contextual influences in mental health help seeking: A focus on ethnic minority youth. *J Consult Clin Psychol*. 2002;70(1):44-55. doi:10.1037/0022-006X.70.1.44
117. De Choudhury M, Sharma SS, Logar T, Eekhout W, Nielsen RC. Gender and Cross-Cultural Differences in Social Media Disclosures of Mental Illness. In: *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. Vol 7. ACM; 2017:353-369. doi:10.1145/2998181.2998220
118. Satcher D. Mental health: Culture, race, and ethnicity—A supplement to mental health: A report of the surgeon general. Published online 2001. Accessed September 17, 2024. https://drum.lib.umd.edu/items/fe061df3-8c83-435b-9b2e-d56e377e6352
119. Haltaufderheide J, Ranisch R. The ethics of ChatGPT in medicine and healthcare: a systematic review on Large Language Models (LLMs). *Npj Digit Med*. 2024;7(1):183. doi:10.1038/s41746-024-01157-x
120. Parray AA, Inam ZM, Ramonfaur D, Haider SS, Mistry SK, Pandya AK. ChatGPT and global public health: applications, challenges, ethical considerations and mitigation strategies. Published online 2023. Accessed August 20, 2024. https://www.sciencedirect.com/science/article/pii/S2589791823000087
121. Diaz-Asper C, Hauglid MK, Chandler C, Cohen AS, Foltz PW, Elvevåg B. A framework for language technologies in behavioral research and clinical applications: Ethical challenges, implications, and solutions. *Am Psychol*. 2024;79(1):79-91. doi:10.1037/amp0001195
122. Weidinger L, Uesato J, Rauh M, et al. Taxonomy of Risks posed by Language Models. In: *2022 ACM Conference on Fairness, Accountability, and Transparency*. ACM; 2022:214-229. doi:10.1145/3531146.3533088

Supplementary Files

Figures

Research Methodology and Process.



Top-N Important Features and Performance of Feature Sets Composed of Different Numbers of Optimal Features.

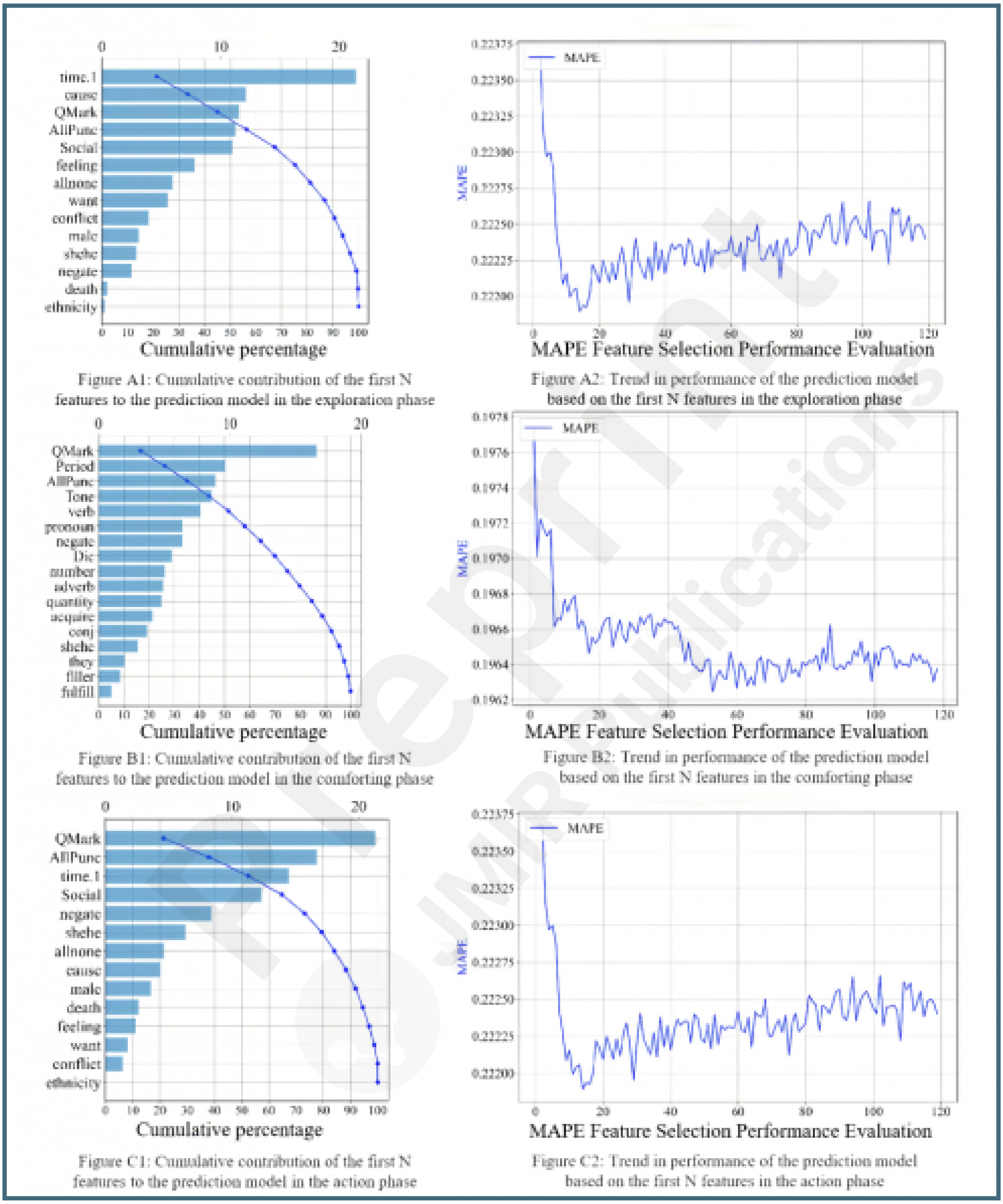


Figure A1: Cumulative contribution of the first N features to the prediction model in the exploration phase

Figure A2: Trend in performance of the prediction model based on the first N features in the exploration phase

Figure B1: Cumulative contribution of the first N features to the prediction model in the comforting phase

Figure B2: Trend in performance of the prediction model based on the first N features in the comforting phase

Figure C1: Cumulative contribution of the first N features to the prediction model in the action phase

Figure C2: Trend in performance of the prediction model based on the first N features in the action phase

SHAP Plot of Global Interpretability for Empathy Scores in ESCs.

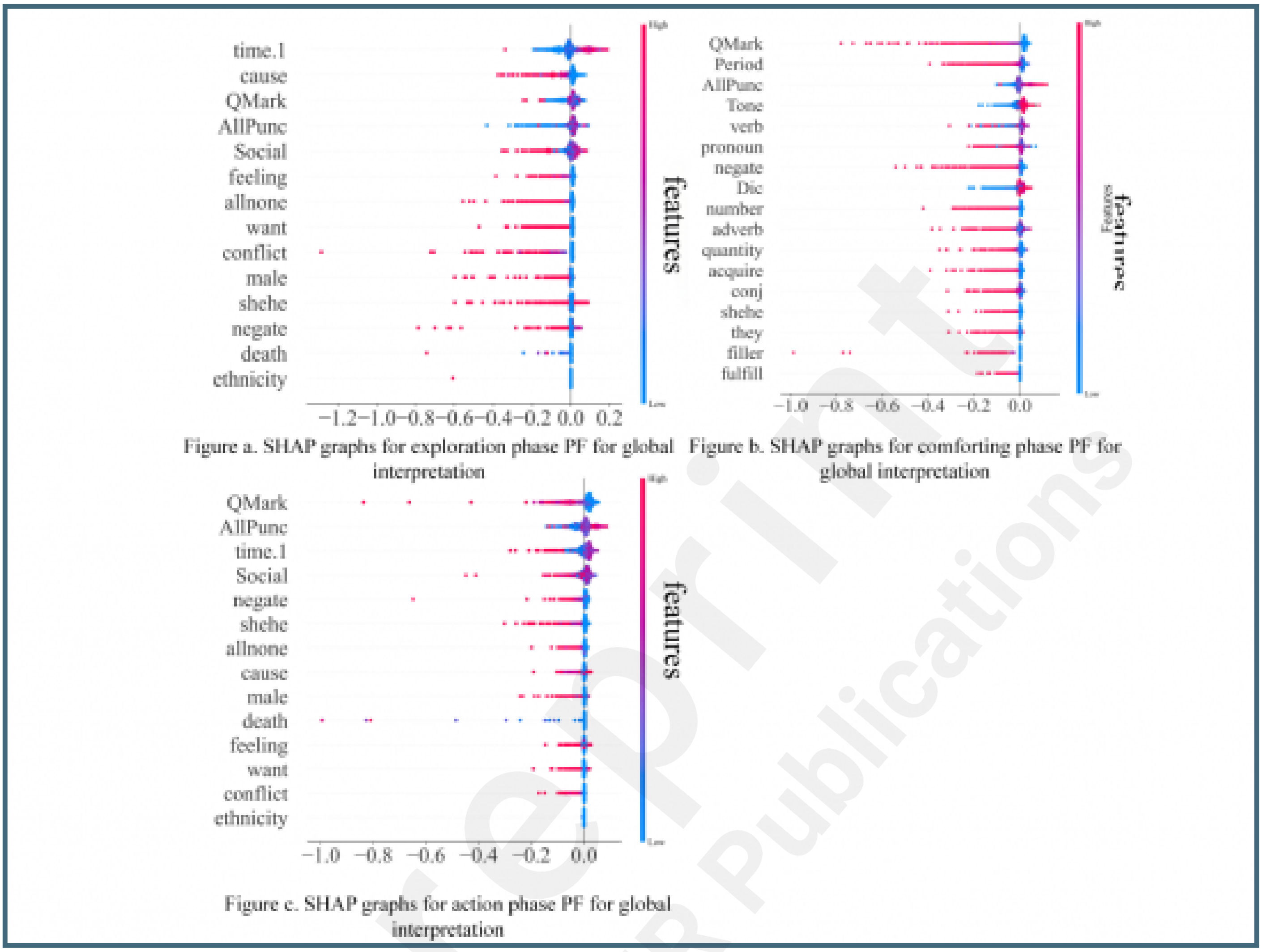
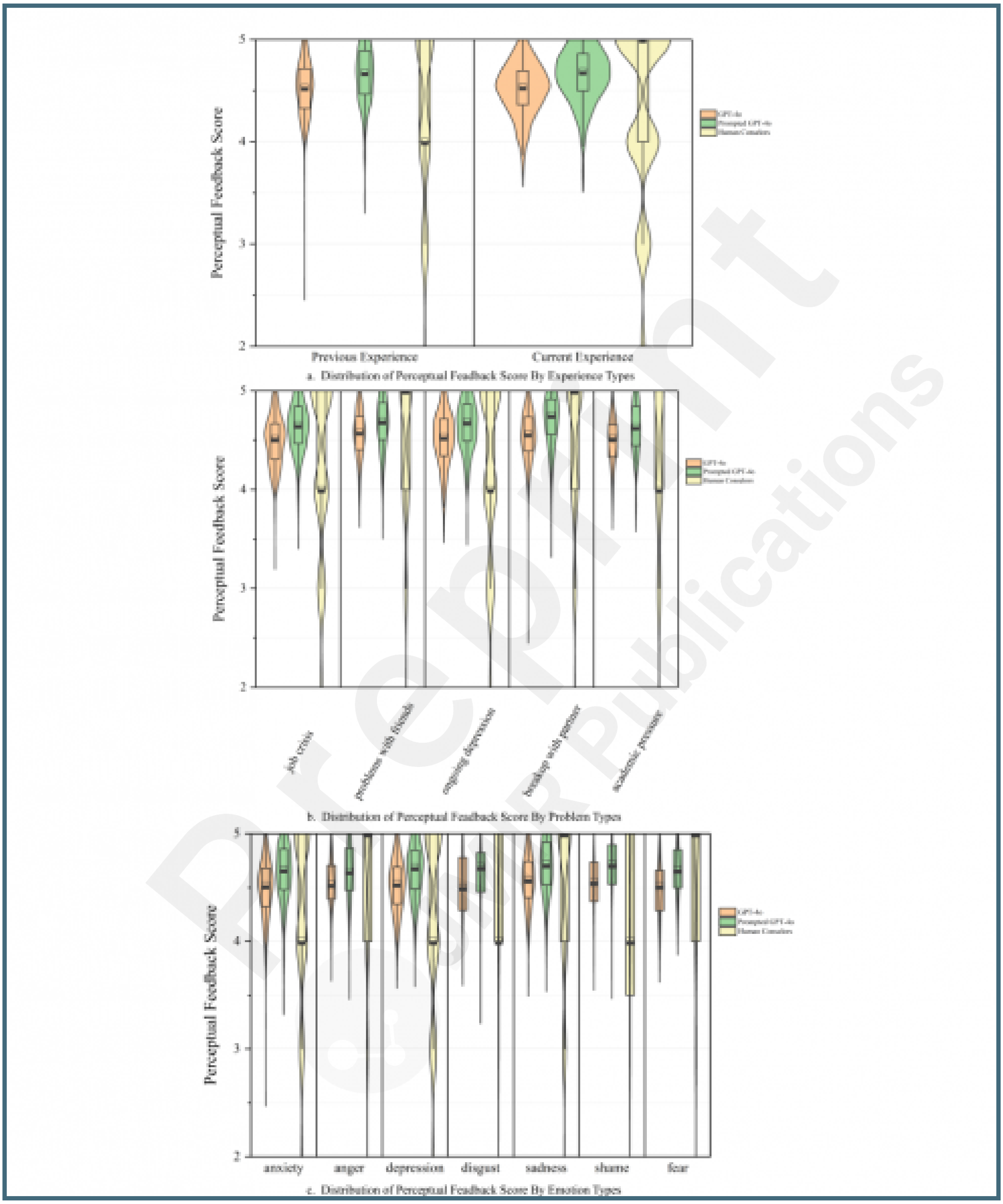


Figure a. SHAP graphs for exploration phase PF for global interpretation

Figure b. SHAP graphs for comforting phase PF for global interpretation

Figure c. SHAP graphs for action phase PF for global interpretation

Distribution of UPF Scores for GPT-4o and Human Counselors Across Different Emotion and Problem Categories.



Multimedia Appendixes

The appendix files referenced in the manuscript.
URL: http://asset.jmir.pub/assets/4f77d22d8b6fe98591030b9c4bb9d1f8.docx