JMIR Publications
The Leading eHealth Publisher

Human-AI Collaboration Supporting GPT-4o Achieving 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: Journal of Medical Internet Research
on: August 21, 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

Preprint
JMIR Publications

Human-AI Collaboration Supporting GPT-4o Achieving Human-Level User Feedback in Emotional Support Conversations: Integrative Modeling and Prompt Engineering Approaches

Yinghui Huang[1, 2] PhD; Lie Li[3] Msc; Wanghao Dong[4, 5] PhD; Yuhang Dong[4, 5] MA; Yingdan Huang[6] PhD; Hui Liu[5, 7] PhD

[1]Research Institute of Digital Governance and Management Decision Innovation Wuhan University of Technology 122 Luoshi Road Wuhan CN
[2]School of Management Wuhan University of Technology 122 Luoshi Road Wuhan CN
[3]School of Electrical and Electronic Engineering Nanyang Technological University Singapore SG
[4]Key Laboratory of Adolescent Cyberpsychology and Behavior (Ministry of Education) 152 Luoyu Road?Hongshan District Wuhan CN
[5]School of Psychology Central China Normal University 152 Luoyu Road, Hongshan District Wuhan CN
[6]Department of Lymphoma Medicine Hubei Cancer Hospital, Tongji Medical College Huazhong University of Science and Technology Wuhan CN
[7]Key Laboratory of Adolescent Cyberpsychology and Behavior (Ministry of Education) Wuhan CN

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

Preprint JMIR Publications

Abstract

Background: Emotional support plays a crucial role in enhancing social interactions, facilitating psychological interventions, and improving customer service outcomes by addressing individuals' emotional needs. The emergence of large language models (LLMs) holds promise for delivering emotional support on a large scale, but their effectiveness compared to human counselors is still not 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 large language models (LLMs), specifically GPT-4o, and to introduce an integrative automatic evaluation framework centered on user-perceived feedback. The framework is designed to enhance LLM performance in emotional support conversations (ESCs) by identifying psycholinguistic clues as intrinsic evaluation metrics and leveraging a customized Chain-of-Thought (CoT) prompting framework.

Methods: The study used a dataset of emotional support conversations from human counselors. An explanatory predictive model was developed using explainable artificial intelligence methods, following an integrative modeling paradigm rooted in computational social science. The model evaluated and interpreted user-perceived feedback scores for GPT-4o. Additionally, the study integrated Hill’s three-stage model of helping into a manually customized chain of thought prompting framework to systematically evaluate GPT-4o's performance in ESCs.

Results: GPT-4o achieved high user-perceived feedback scores, demonstrating relative stability in its performance, but it still significantly trails behind human counselors overall (Cliff's Delta = 0.087, $p < 0.001$). The evaluation framework, which identified 41 distinct linguistic clues related to emotional expression, social dynamics, cognitive processes, linguistic style, and decision-making stages, enhanced the understanding of both processes and outcomes in ESCs. Notably, GPT-4o's user-perceived feedback scores significantly improved with the use of manually customized Chain of Thought prompts ($p < 0.001$, Cohen's d: 0.378), but showing no significant difference from the average performance of human counselors overall (p-adj: 0.47, Cliff's Delta: -0.014). However, thought prompts demonstrate a significant advantage in specific emotion categories such as fear (p: 0.002, Cliff's Delta: -0.23), sadness (p: 0.012, Cliff's Delta: -0.105), and break up with partner issues (p: 0.254, Cliff's Delta: -0.06). However, GPT-4o exhibited weaknesses in emotional understanding, cognitive complexity, language fluency, and handling extreme scenarios.

Conclusions: This study provides preliminary evidence of GPT-4o's emotional support capabilities and proposes a user-

perceived feedback-centered integrative evaluation framework for ESCs. The findings suggest a cautiously optimistic outlook for the application of advanced large language models (LLMs) in emotional support services, although significant challenges remain, particularly in enhancing the depth of exploration in conversations and the personalization of language. The proposed framework encourages the integration of human expertise into LLMs, enhancing their efficacy and advancing the development of trustworthy AI-based emotional support services.

(JMIR Preprints 21/08/2024:65435)
DOI: https://doi.org/10.2196/preprints.65435

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

Preprint JMIR Publications

Original Manuscript

Preprint
JMIR Publications

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.

[*]Hui Liu and Yingdan Huang are co-corresponding authors.

Human-AI Collaboration Supporting GPT-4o Achieving Human-Level User Feedback in Emotional Support Conversations: Integrative Modeling and Prompt Engineering Approaches

Abstract

Background: Emotional support plays a crucial role in enhancing social interactions, facilitating psychological interventions, and improving customer service outcomes by addressing individuals' emotional needs. The emergence of large language models (LLMs) holds promise for delivering emotional support on a large scale, but their effectiveness compared to human counselors is still not 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 large language models (LLMs), specifically GPT-4o, and to introduce an integrative automatic evaluation framework focused on user-perceived feedback. 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 emotional support conversations 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 user-perceived feedback scores for GPT-4o. Additionally, Hill's three-stage model of helping was integrated into a manually customized Chain-of-Thought (CoT) prompting framework to systematically evaluate GPT-4o's performance in ESCs.

Preprint JMIR Publications

Results: GPT-4o achieved high user-perceived feedback 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 user-perceived feedback scores significantly improved with the use of manually customized Chain-of-Thought prompts ($p < 0.001$, Cohen's d: 0.378), showing no significant difference from the average performance of human counselors overall (p-adj: 0.47, Cliff's Delta: -0.014). However, Chain-of-Thought prompts demonstrated a significant advantage in specific emotion categories such as fear ($p = 0.002$, Cliff's Delta: -0.23), sadness ($p = 0.012$, Cliff's Delta: -0.105), and issues related to breakups with partners ($p = 0.254$, Cliff's Delta: -0.06). Despite these improvements, GPT-4o exhibited weaknesses in emotional understanding, cognitive complexity, language fluency, and handling extreme scenarios.

Conclusions: This study offers preliminary evidence of GPT-4o's emotional support capabilities and introduces a user-perceived feedback-centered integrative evaluation framework for ESCs. The findings suggest a cautiously optimistic outlook for the application of advanced large language models (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 the development of 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, evidenced by an insufficient number of professional providers and numerous geographic and economic barriers that limit access to these services. As a result, many individuals in need of assistance are unable to receive the necessary support, with underdiagnosis rates nearing 90% in certain low-income regions due to limited access to professional care[1]. Emotional support, as a key psychological intervention, helps individuals alleviate stress and challenges by providing empathy, affirmation, and encouragement, thereby promoting mental health and social adaptation[2]. Through emotional support, negative emotions can be mitigated, psychological resilience and self-confidence can be enhanced and emotional communication and social adaptation can be facilitated[2–6]. Despite its importance, the complexity and specialized nature of emotional support render expert resources scarce and services expensive. Consequently, many individuals facing life stressors are unable to access timely and effective emotional support to mitigate their 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 (ESC). While existing technologies have made progress in simulating emotional support dialogues [8–10], the advent of GPT-4 marks a significant advancement in artificial intelligence capabilities, especially in its exceptional performance in contextual understanding and language expression[11,12]. However, large language models like GPT-4 still struggle to accurately understand and respond to human emotional needs, primarily due to limitations in their emotional comprehension and expression[13,14]. Given the life-critical

nature of mental health applications, the deployment of large language models necessitates the establishment of a unified and comprehensive set of fundamental metrics that can meticulously assess the model's capabilities, identify potential errors, and implement effective feedback mechanisms. These metrics are crucial for advancing the delivery of robust, accurate, and reliable healthcare services. Existing AI-based conversational systems for evaluating emotional support dialogues lack user-centered automatic evaluation frameworks, which should include both intrinsic metrics (focused on dialogue process details) and extrinsic metrics (focused on user impact) The former is used for dynamically understanding and promoting subtle differences in complex dialogue processes, while the latter evaluates their impact on users, both of which warrant close attention.

A critical question arises when addressing real emotional support needs: Can advanced generative AI, such as GPT-4o, match or even approach the level of professional human counselors? The objective of this study is to explore and propose 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 the user experience with advanced Large Language Models (LLMs) in ESCs, with the broader aim of comprehensively exploring their potential as an effective supplement to Mental Health services. Integrative modeling refers to the approach of combining data-driven insights and human expert knowledge, such as psychological theories, to guide the design and implementation of AI systems[15]. This method not only enhances the relevance and accuracy of AI responses but also increases resonance with users' emotional states [16]. Prompt engineering, a key technology that bridges the gap between user intent and AI understanding, guides GPT-4 to generate more accurate responses through well-crafted prompts[17]. This strategy, based on the Chain of Thought (CoT) approach, requires AI to process problems through logical steps, thereby enhancing its ability to handle complex emotional dialogues[17,18]. This study will demonstrate how these techniques assess and enhance GPT-4o's ability to accurately understand and respond to users' emotional needs, thereby increasing user trust and satisfaction with AI-driven emotional support platforms and addressing 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 (gen AI and prompt engineering), the research problem (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. This process assists them in understanding and addressing the challenges they face[2]. Emotional support can be conveyed through both verbal and non-verbal behaviors. Research suggests that, from a life-span perspective, non-verbal emotional support typically precedes verbal forms, as non-verbal cues often play a crucial role in initial emotional bonding and communication[2]. In the context of emotional support, the provider must not only master various emotional support techniques and expression skills but also choose appropriate responsive strategies based on the individual's specific situation and the background of their issues[19,20]. According to Hill's three-stage model of helping, providing emotional support to a help-seeker generally involves three steps: exploration, where the help-

seeker is assisted in discovering the problem; insight, where the help-seeker gains a deeper understanding of themselves; and action, where the help-seeker is guided to make decisions on how to address the problem[21]. These stages do not necessarily occur in sequence and may repeat[21]. In this context, 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, there are distinctions 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 and incorporates emotional support by encoding context history and predicting strategy tokens to guide responses. MISC utilizes the commonsense model to infer the help-seeker's immediate mental state and employs a weighted average strategy to guide response generation[27]. GLHG leverages a graph neural network to encode the relationship between the help-seeker's global situation and local intentions for guiding responses. However, both MISC and GLHG depend on external knowledge from COMET, which may not apply to specific domains and requires considerable human effort to develop. Additionally, they are limited to the current conversation scope, overlooking abundant prior knowledge in the dataset. These studies have demonstrated that emotional support systems significantly alleviate psychological stress and improve mental health in multi-turn emotional support conversations[13,28]. They have also enhanced user satisfaction and interaction quality[29,30]. Nonetheless, emotional support conversation systems still face challenges in accurately recognizing complex emotions and selecting the most appropriate emotional support strategies[30]. Moreover, existing systems struggle to fully account for users' personalities, backgrounds, and specific needs, and their inability to provide explanations for responses results in a lack of transparency[29].

2.2 GPT-4 and Prompt Engineering

The potential of chatbots in the domain of emotional support is garnering increasing attention, with GPT-4 (Chat Generative Pretrained Transformer) being a prime example. Developed by OpenAI, GPT-4 represents a significant advancement in large language models[31,32]. It excels in contextual understanding, enabling it to accurately comprehend user inputs and generate appropriate responses, even accommodating subsequent corrections[11,33]. This ability is critical for dialogue agents in the mental health field, where interactions must be contextually relevant rather than relying on predefined content[33,34]. Furthermore, existing studies underscore GPT-4's strong reasoning abilities, which allow it to draw logical conclusions, and its effectiveness in providing clinically relevant insights. These capabilities not only enhance the model's interpretability but also build greater user trust in the system[35–38].

AI, as a powerful productivity tool, offers immense potential in enhancing various aspects of human mental health services[39]. 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 even offer treatment suggestions[40]. A meta-analysis has shown that AI conversation agents can significantly alleviate depressive symptoms in patients[41]. 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[42]. Furthermore, AI can automate the evaluation of treatment effectiveness, making it particularly valuable for training novice counselors and tracking therapeutic progress[43,44]. Building on this potential, the focus of this study is 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'[45].

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 effectively conveying user intentions to the language model. It is essential for achieving the desired functionality of large language models[12,46,47]. 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, such as 'clarity and precision,' which involves crafting specific, clear prompts to reduce output uncertainty, and 'role clarity,' where the model is assigned a clear identity, like an assistant or expert, to ensure consistent responses[47]. 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[48]. This approach encourages the model to explicitly generate a reasoning process, often leading to more accurate conclusions[18]. In emotional problem-solving, grounding CoT in a professional theoretical framework can yield more satisfactory outputs. Overall, LLMs have demonstrated some capacity for emotional understanding[37], expression[49], and empathetic responses[50]. However, these capabilities are dynamic and are best realized through human-AI collaboration in prompt engineering. Considering the critical role of CoT prompting, it's crucial to evaluate GPT-4's performance differences before and after CoT implementation.

2.3 Evaluation of Large Language Model-Driven Chatbots

Designing emotional intelligence agents, which can perceive, integrate, understand, and regulate emotions[51,], is a key research focus in dialogue systems[52]. Chatbots offer 24/7 personalized support, serving as an alternative to traditional therapy, especially where mental health professionals are scarce [53]. TThey also provide a private therapy option, potentially more acceptable to those concerned about stigma[54]. Research on emotional chatting has recently surged[17,22,23,27]. Studies indicate that users sometimes prefer chatbots over human professionals, aiding those hesitant to seek traditional therapy. Chatbots are also seen as less judgmental and biased than humans, promoting self-disclosure and greater conversational flexibility[54,55].

Despite their potential, chatbot adoption in mental health is limited by concerns over information accuracy, technological maturity, ethics, and interaction authenticity[56]. 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 own strengths and weaknesses[57]. Manual evaluations include quantitative methods, like surveys and scales measuring user satisfaction[58], and qualitative methods, such as interviews and focus groups, exploring user experiences and perceptions[59]. Manual methods are valued for their flexibility, comprehensiveness, and professionalism, as human annotators can manage complex situations

and notice details that automated methods might miss[57]. Automated methods are lauded for efficiency, objectivity, and consistency, quickly processing large data sets while reducing subjectivity, high costs, and inconsistencies of human evaluation [57]. They also mitigate ethical risks in the evaluation process[57]. However, automated methods often suffer from limited coverage, inflexibility, and a reliance on predefined benchmarks, which can cause 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[60]. These metrics, known for their computational simplicity, are categorized into general automatic and dialogue-based types. General automatic metrics, like BLEU, ROUGE, and Perplexity[57], 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, which assesses the chatbot's ability to ask relevant questions; and Average Request Turn, which tracks the average number of interactions between user and chatbot[57]. These metrics provide valuable quantitative tools for evaluating LLMs. However, they focus on surface-level similarity and language-specific aspects, making them insufficient for healthcare chatbots They fail to capture crucial elements like semantics, context, distant dependencies, and human perspectives, especially in real-world scenarios[57]. Extrinsic evaluation metrics, in contrast, assess the model's impact on users and how well it meets their expectations and needs[61]. These metrics measure language model performance by incorporating user perspectives and real-world scenarios[60]. Extrinsic metrics, collected through subjective assessments, involve human judgment and are divided into general-purpose and health-specific categories[46]. General-purpose human evaluation metrics assess LLM performance across diverse domains[57]. These metrics evaluate quality, fluency, relevance, and overall effectiveness, covering a broad range of real-world topics, tasks, and user needs[62]. Health-specific metrics focus on evaluating how healthcare-oriented LLMs and chatbots process and generate health-related information, emphasizing accuracy, effectiveness, relevance, reliability, currency, healthy behaviors, and emotional support[57]. These metrics aim to incorporate context and semantic awareness into extrinsic evaluations of LLMs. However, these 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[62]. 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[57].

Overall, in terms of evaluation methods, automated methods can effectively adapt to the dynamic evaluation scenarios required by prompt engineering and can adequately cover diverse configurations related to psychological help-seeking users, domains, and task types. Regarding evaluation metrics, intrinsic metrics emphasize computational efficiency and consistency, and can quantify a chatbot's dynamic performance in the face of complex and subtle interaction differences. Extrinsic metrics focus on user perspectives and actual application performance, but relevant research has not yet fully covered comprehensive, complex, and user-centered evaluation metrics[63]. Given that user satisfaction, therapeutic effectiveness, engagement, and reliability are widely used to measure the capabilities of GPT-4[58,59], user acceptance of the support, as reflected in subjective feedback, is closely related to the effectiveness of emotional support[64] and can serve as a direct indicator of emotional support quality. Based on subjective acceptance, integrating intrinsic and extrinsic metrics can provide more accurate evaluations and understanding, offering important insights for optimizing

chatbots and improving user satisfaction. Additionally, current evaluation schemes primarily focus on performance evaluation—measuring accuracy, reliability, and user interaction experience—but often lack the robustness needed to effectively evaluate LLMs in complex and nuanced interactions required in mental health applications[65].

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

Recent years have seen significant progress in using machine learning (ML) and deep learning (DL) to understand and evaluate psychotherapy dialogues. Supervised learning techniques are widely used to classify or predict labeled therapeutic processes and outcomes. For example, 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[66]. In contrast, unsupervised learning techniques have been applied to identify clusters in unlabeled patient or therapy data, helping to analyze therapeutic processes and outcomes[67]. Deep learning has excelled in personalized treatment prediction, achieving up to 80% accuracy in predicting treatment responses for depression[68]. Machine learning has effectively predicted dropout rates in outpatient psychotherapy. 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[69].

Techniques have excelled in coding 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[70]. Studies show that improvements in psychological states correspond with changes in language patterns, serving as indicators of therapy effectiveness[71–73]. 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[74,75]. Latent Dirichlet Allocation (LDA) models, especially labeled-LDA, have been used to automate therapy session coding and predict effectiveness[76].

Explainable AI (XAI) has gained attention in psychotherapy for its crucial role in improving model transparency and interpretability. XAI offers real-time explanatory feedback, helping therapists adjust their strategies and improve patient outcomes[77]. Using conversational AI, like chatbots, XAI provides personalized support and real-time analysis of emotional changes, delivering feedback that helps therapists understand patients' states better[38]. XAI parses and explains language patterns and emotional shifts in psychotherapy, offering detailed decision paths to help therapists identify therapeutic opportunities and risks[78]. XAI enhances human-AI collaboration by providing interpretable decision suggestions, helping clinicians understand and trust AI decisions, thereby improving treatment planning and quality[79,80]. 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 methods 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 in mental health[81]. Overall, XAI and integrative modeling show great potential in improving transparency and interpretability in predictive models, fostering human-AI collaboration in psychotherapy, and advancing the adoption of these technologies in clinical practice.

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 (PF) is a key user-centered evaluation metric. This study focuses on evaluating 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? To address this, the study proposes an automated evaluation method to explore GPT-4's PF in ESCs, comparing its performance to that of 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 between the two. The use of psychological 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? To answer this, the research uses psychological linguistic clues as intrinsic metrics and employs explainable AI-based methods to evaluate GPT-4's responses in ESCs. This method ensures explainability while effectively mitigating the risk of privacy breaches, improves efficiency, interpretability, and detailed analysis, thereby providing a more granular and insightful evaluation for the effective and ethical application of LLMs in mental health applications.

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 this, the research will use explainable ML-based methods to reassess and understand GPT-4's responses with a customized CoT prompt.

The research approach involves several steps: First, using an ESC dataset and an integrative modeling paradigm, natural language processing and machine learning methods are employed to build a predictive model for perceived feedback scores, with GPT-4o as the subject, to quantitatively assess its user-perceived feedback in ESCs. Second, explanatory modeling is used to identify and analyze key linguistic clues that affect ESC perceived feedback scores, serving as intrinsic evaluation metrics. Finally, Hill's three-stage model of helping is used as a framework, combined with XAI analysis, to construct a custom CoT prompt. GPT-4's responses are then reanalyzed to evaluate its performance. Additionally, statistical methods are used to validate the PF ratings across different emotion and problem types, along with related intrinsic metrics. A comprehensive comparison is then made between GPT-4's performance before and after custom prompting, and that of human counselors in ESCs.

3.Methods & Experiments

In this study, the study use a dataset of human counselor emotional support conversations and apply an integrative modeling process from computational social science to evaluate and understand the user feedback of generative AI in ESC. This process involves data collection, preprocessing, feature engineering, constructing the ESC perceived feedback evaluation model, and integrating human expertise through CoT prompt engineering. These steps aim to evaluate GPT-4o's emotional support capabilities in ESCs. The study propose an integrative evaluation framework and conduct a comprehensive comparison with human counselors. The following sections outline the datasets, features, algorithms, and prompt engineering methods used to develop, interpret, and evaluate these 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 emotional support conversation dataset developed by Liu et al. (2021). 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, a task that can be challenging and risky for less experienced support providers. During dataset collection, researchers provided detailed ESC framework training to support providers, selecting the top 7.8% of applicants who passed the examination, resulting in 1,053 high-quality emotional support conversations. The dataset consists of dialogue texts from real counseling scenarios, created by professional counselors and help-seekers. The dataset 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 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 dataset are presented in Tables 1 and 2. Each conversation averages 29.8 turns. The most frequently mentioned issues were persistent 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

First, we used NLP, machine learning, and deep learning to construct models that predict and explain user-perceived feedback on GPT-4o's emotional support capabilities. Next, we designed a Chain of Thought (CoT) prompting framework based on Hill's helping skills theory to enhance GPT-4o's performance in delivering emotional support.

3.2.1 Automated Evaluation of user-perceived Feedback in ESCs

To thoroughly validate GPT-4o's emotional support capabilities, we first developed an

evaluation model for PF ratings in ESCs. This model aims to predict PF ratings based on dialogue content between seekers and supporters during emotional support interactions, using machine learning and deep learning models. The seeker's PF rating for the supporter serves as the model's predictive target. The feature set is constructed using linguistic metrics like Linguistic Inquiry and Word Count (LIWC) and Language Style Matching (LSM). Additionally, the feature set includes annotations from ESCs, such as emotion categories and the types of problems faced by the seeker. This information is used to build the predictive model, ensuring an accurate assessment of GPT-4o's performance in ESCs.

The study utilized several regression algorithms, including Ridge Regression, Random Forests (RF), Extreme Gradient Boosting (XGBoost), and Support Vector Regression (SVR)[82]. 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[82]. Additionally, we employed advanced neural network architectures and pre-trained language models, including BERT-BiLSTM-Attention, BERT-BiLSTM, RoBERTa, and XLNet. Bidirectional Encoder Representations from Transformers (BERT) captures rich semantic information using a bidirectional encoder, and when combined with Bidirectional Long Short-Term Memory (Bi-LSTM), it enhances sequence modeling[83]. The attention mechanism allows the model to focus on key information, improving predictive performance. Robustly Optimized BERT Pretraining Approach (RoBERTa) is an optimized BERT version, enhancing performance with increased training data and time[84]. Extreme Language Net (XLNet) combines BERT's bidirectionality with autoregressive model advantages, using a Permutation Language Model to handle sequence dependencies flexibly, often outperforming other models in NLP tasks[85]. These models, with their distinct architectures and optimization strategies, offer strong semantic understanding and predictive capabilities, making them suitable for PF evaluation in ESCs. used Root Mean Square Error (RMSE) and Mean Absolute Percentage Error (MAPE) as evaluation metrics.

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

Integrative modeling combines explanatory and predictive approaches to predict outcomes and estimate effects in new situations, identifying causal factors and predicting their impact. This approach involves evaluating and validating the accuracy of model predictions. In this study, we evaluate GPT-4o's user-perceived feedback in ESCs, identify and estimate the impact of relevant psycholinguistic clues (as intrinsic evaluation metrics) guided by human experience, and further enhance GPT-4o's performance through customized prompt engineering.

Specifically, the study used explainable machine learning to develop customized prompts for GPT-4o’s responses in ESC, with a focus on expressing empathetic language. The core of this prompt engineering, driven by explanatory modeling, is to use the help-seeker’s perceived feedback rating 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 linguistic features influence the help-seeker's PF rating, providing a scientific basis for constructing effective prompts.

For different stages of the ESC, the study first used Recursive Feature Elimination with 10-fold Cross-Validation combined with XGBoost to filter out the most significant variables predicting PF ratings, thereby simplifying the feature set. Next, the study 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[86,87]. Based on SHAP analysis, the study conducted 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 features, improving model performance (measured by MAPE and RMSE) with each step. This process continued until all top-n 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 clues that influence PF scores in ESC, utilizing SHAP values to understand how the supporter's language impacts PF predictions. This analysis provided critical guidance for developing GPT-4o's response strategies. Drawing from Hill's three-stage helping skills model, the study combined these explainability analyses with human experience in ESCs to create a manual CoT prompt strategy framework, offering key insights to enhance GPT-4o's performance. Additionally, the study introduced the RTF (Request, Task, and Format) based CoT prompt framework as a control against the manual CoT prompts mentioned earlier[88]. 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 help-seekers' perceived feedback scores. SSecond, the study performed prompt engineering to enhance GPT-4o's emotional support capabilities. This involved identifying key linguistic features influencing help-seeker feedback through explainable machine learning and model refinement, and designing manual CoT prompts for ESC tasks. Finally, through statistical analysis, the study quantitatively assessed GPT-4o's PF rating in ESCs compared to human counselors. The main research process and methods are illustrated in Figure 1.

Preprint JMIR Publications

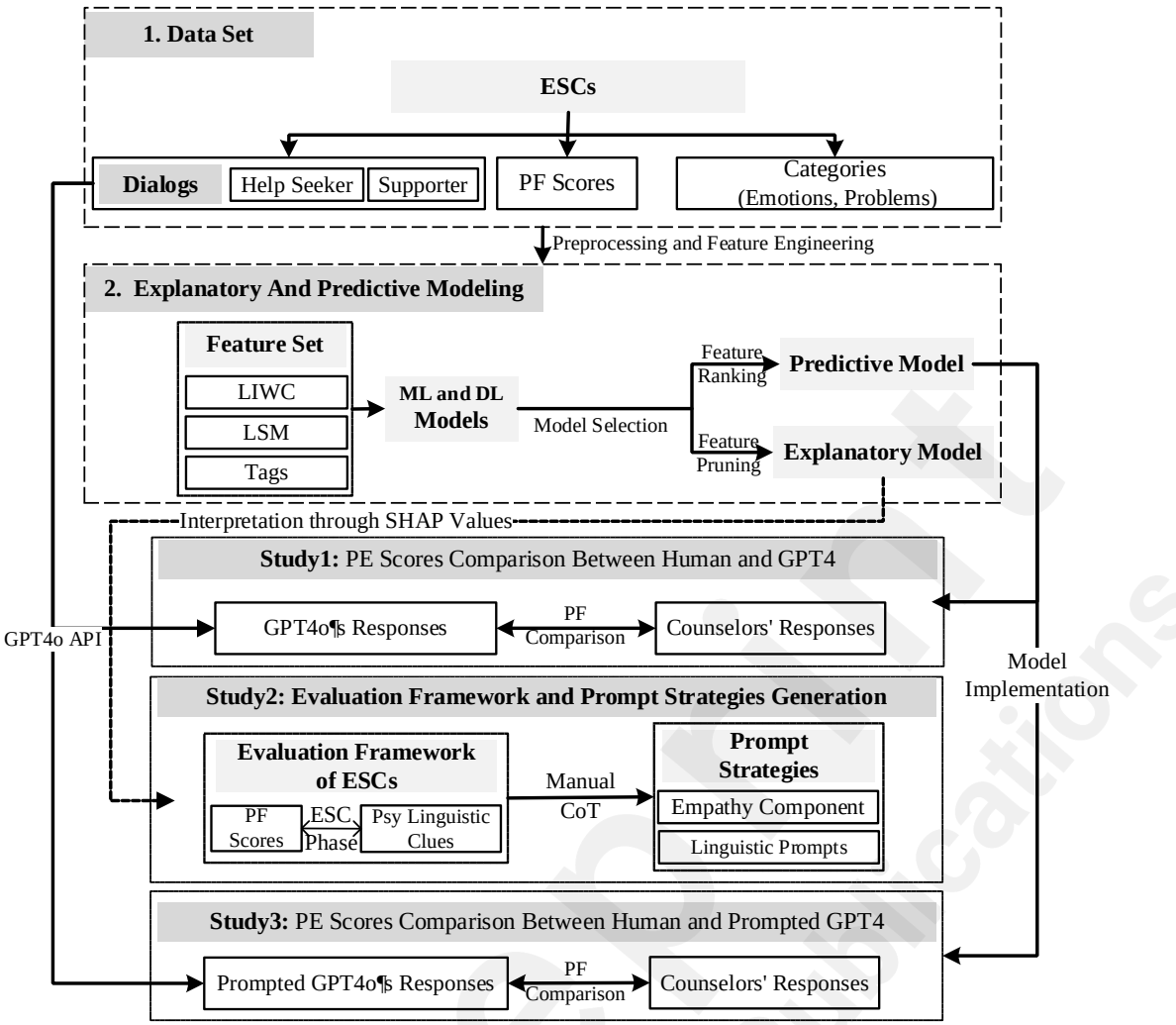


Figure 1. Research Methodology and Process

4. Results

4.1 Analysis of GPT-4o's Perceived Feedback Scores in ESCs Using Predictive Modeling

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

4.1.1 Performance Analysis of the PF Evaluation Model

The study developed various predictive models using a range of regression algorithms. The performance metrics for each model are presented in Table 3. Among the deep learning models, the BERT-BiLSTM-Attention model achieved an RMSE of 0.579 and a MAPF of 21.36%. Among the machine learning models, XGBoost, while not a deep learning model, delivered the best performance with an RMSE of 0.486 and a MAPF 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 PF Between GPT-4o and Human Counselors

Using XGBoost, the optimal machine learning model for PF prediction, we predicted the PF values for GPT-4o's responses. Since the PF ratings 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. 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 except for Anxiety, Depression, and Fear. In terms of different problem types, GPT-4o generally underperformed compared to human counselors. For themes such as Friendship Issues, Work Crisis, Persistent Depression, and Academic Pressure, the effect size of the differences was small but statistically significant, except for "Breakup" and "Friendship Issues," where 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 Category	Anxiety	517	4.0□4.0□5.0□	4.51 (4.46, 4.57)	118836	< 0.001	0.114
	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[89].

4.2 Development and Validation of the Integrative Evaluation Framework for ESCs

To further understand and evaluate GPT-4o's performance regarding PF scores, the study developed explanatory models and 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, the study conducted a sensitivity analysis to assess the impact of each feature within the best-performing model. Next, the study performed feature pruning, retaining only those features that significantly affected predictive accuracy. The model was further refined using a set of key features, based on Davis's empathy component theory and Hill's three-stage helping skills model.dictive accuracy. 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 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, with a focus on MAPF to optimize performance."

In the Exploration, Comforting, and Action stages, the explanatory model achieved local optimal performance with 14, 17, and 14 features, respectively.

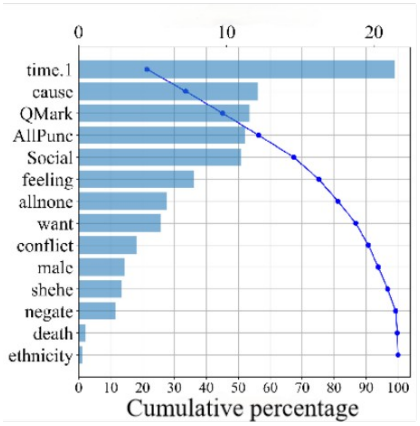


Figure A1: Cumulative contribution of the first N features to the prediction model in the exploration phase

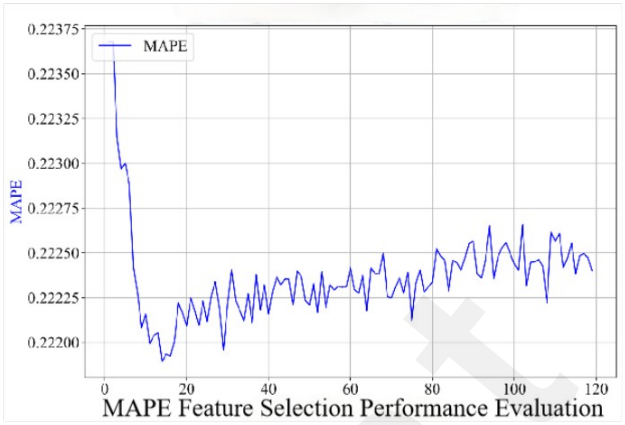


Figure A2: Trend in performance of the prediction model based on the first N features in the exploration phase

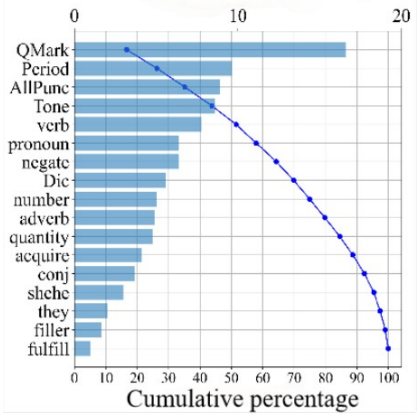


Figure B1: Cumulative contribution of the first N features to the prediction model in the comforting phase

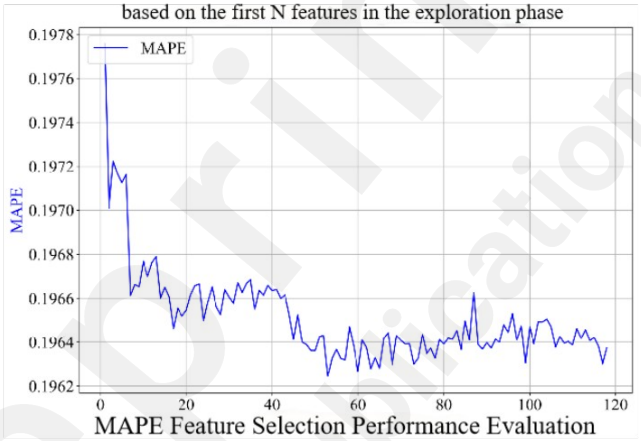


Figure B2: Trend in performance of the prediction model based on the first N features in the comforting phase

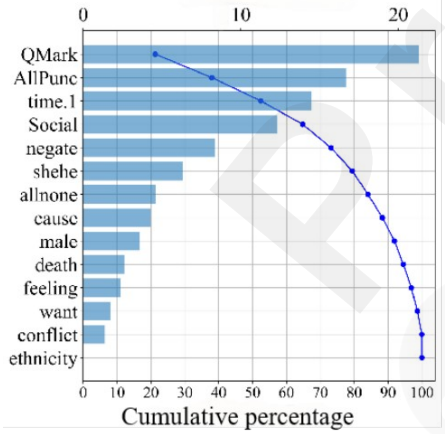


Figure C1: Cumulative contribution of the first N features to the prediction model in the action phase

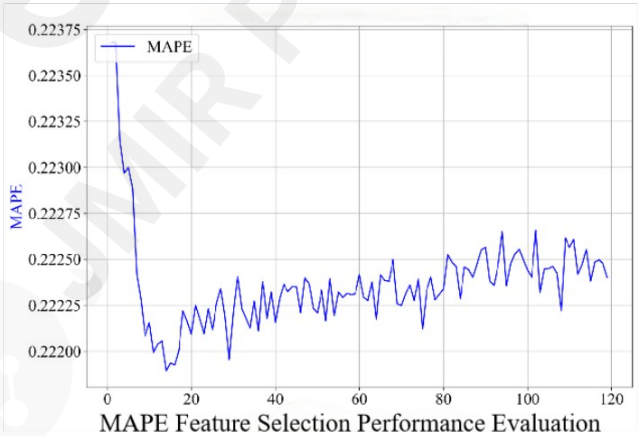


Figure C2: Trend in performance of the prediction model based on the first N features in the action phase

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, the study conducted a global interpretability analysis using SHAP values to assess how various features influence PF ratings. The study 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 identify the range and direction of each feature's impact on PF, the study consolidated individual data samples into a comprehensive SHAP interpretability plot, as shown in Figures

A2, B2, C2 in Figure 3.

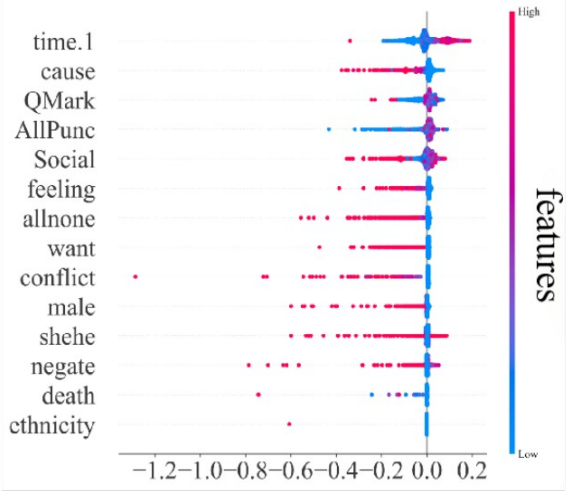


Figure a. SHAP graphs for exploring phase PF for global interpretation

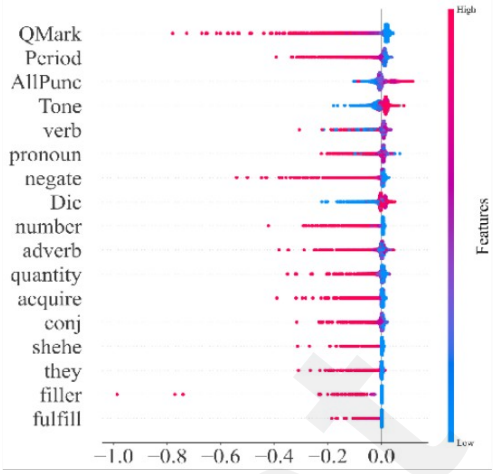


Figure b. SHAP graphs for comforting phase PF for global interpretation

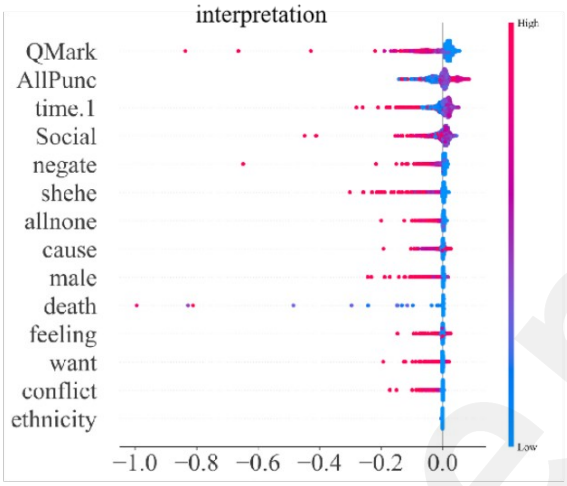


Figure c. SHAP graphs for action phase PF for global interpretation

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

This analysis revealed two primary ways in which features influence the help-seeker's perceived feedback across different ESC stages. Specifically, as shown in **Table** 5, in the three stages of Hill's helping skills theory, intrinsic metrics (*i.e.*, key linguistic cues) were identified and showed diverse relationships with the extrinsic metric (*i.e.*, *Perceived Feedback* score). In the *Exploration Stage*, intrinsic metrics positively associated with PF included punctuation-related features such as *Question Mark (QMark)* and *All Punctuation (AllPunc)*. Negatively associated intrinsic metrics included emotion-related *Feeling* in *Emotion* category, *She/He (shehe)* in *Linguistic* dimension, *Conflict, Male* and *Social* in *Social Process*, *Ethnicity* in *Cultural* category, *Want* in *State* category, *Time* in *Time Orientation*, *Cause* and *Negation (negate)* in *Cognitive Processes*, *All/None (allnone)* in *Cognitive Processes*, and *Death* in *Body-related* category. In the *Comforting Stage*, positively associated intrinsic metrics included *Verb* in *Psychological Process*, *Tone* in *Emotion*, and *Dictionary (Dic)* in *Linguistic* dimension. Negatively associated intrinsic metrics involved *Pronoun* in *Linguistic* dimension, *Number* in *Linguistic* dimension, *Adverb*, *Quantity*, *Conjunction (conj)*, *They*, *Fulfill* and *She/He (shehe)* in *Linguistic* dimension, *Filler* in *Social Process*, *Acquire* in *State* category, *Question Mark (QMark)* in *Punctuation* and *Period* in *Punctuation*, as well as *Negation (negate)* in *Cognitive Processes*. In the *Action Stage*, positively associated intrinsic metrics mainly included *Conflict* in *Social Process* and *All Punctuation (AllPunc)* in *Punctuation*. Negatively associated intrinsic metrics

included *Feeling* in Emotion category, *Male*, *Social* in *Social Process*, *Time, Question Mark (QMark)* in *Punctuation*, *She/He (shehe)* in *Linguistic* dimension, *Negation (negate)* in *Cognitive Processes*, *Death* in *Body-related* category, *Ethnicity* in *Cultural* category, and *Want* in *State* category.

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

ESC Stages	Categories of Intrinsic Metric	Intrinsic Metric (Linguistic Clues)	Relation Between Intrinsic Metric and 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 counseling stages associated with specific linguistic clues, encompassing the three stages: *Exploration*, *Comforting*, and *Action*.

[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).

Preprints JMIR Publications

4.2.2 Validation of the Integrative Evaluation Framework of ESCs

To verify the effectiveness of intrinsic evaluation metrics related to PF in ESCs for human counselors and GPT-4o, the study conducted paired sample t-tests to compare linguistic clue features across different ESC stages. Differences in intrinsic evaluation metrics between the two were analyzed using t-tests, with results presented in Table 6.

Table 6. Analysis of the Performance of GPT-4o and human counselor in the intrinsic evaluation metric

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	shehe	Negative	-5.394***

Dimensions	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 in parentheses indicating the level of statistical significance. Positive values indicate greater usage by GPT-4o, while negative values indicate greater usage by human counselors. The meaning of the asterisks is as follows: * indicates $p < 0.05$, ** indicates $p < 0.01$, *** indicates $p < 0.001$.

[b]**Category:** Refers to the category each linguistic clue belongs to, based on LIWC-22[90], covering various aspects from linguistic dimensions to psychological processes, culture, lifestyle, and physical states, as shown in Appendix Table S1 for the details.

Specifically, in the *Exploration Stage*, GPT-4o exhibited significant differences in several intrinsic metrics. In terms of *Social Processes*, GPT-4o used significantly more social language (*Social*) than human counselors, indicating its superiority in social interaction; however, it also exhibited significantly more conflict-related (*Conflict*) and male-related (*Male*) expressions, which are negatively correlated with PF, suggesting that GPT-4o is less effective than human counselors in reducing conflictual and gender-related expressions. Regarding *State* category, GPT-4o expressed significantly fewer needs (*Want*) than human counselors, showing more restraint in this area, which negatively impacts PF. In *Time Orientation*, GPT-4o used significantly more time-related expressions (*Time*) than human counselors, indicating its strength in time orientation. In *Cognitive Processes*, GPT-4o used significantly more causal expressions (*Cause*) than human counselors, indicating its less effective use of causal reasoning; it used significantly fewer all-or-none logic expressions (*Allnone*), indicating better handling in this aspect; and it showed more caution in reducing negation (*Negate*), achieving better results. In *Punctuation* usage, GPT-4o constructed significantly fewer questions (*QMark*) than human counselors, indicating a deficiency in question construction, which could impact the exploration and guidance of the conversation; however, it used significantly more overall punctuation (*AllPunc*) than human counselors, showing greater diversity and complexity in language expression. Additionally, there were no significant differences between GPT-4o and human counselors in *Culture* category (*Ethnicity*), *Emotion* category (*Feeling*), and *Physical* topics (*Death*), indicating comparable performance in these areas.

In the *Comforting Stage*, regarding *Linguistic Dimensions*, GPT-4o used significantly fewer technical terms (*Dic*) than human counselors; since this metric is positively correlated with PF, it indicates that GPT-4o is less effective in this area. However, GPT-4o used significantly more *Pronouns* (*Pronoun*), numerical expressions (*Number*), and conjunctions (*Conj*) than human counselors; these metrics are negatively correlated with PF, suggesting that GPT-4o's performance in these areas is inferior to that of human counselors. In *Emotion* category, the emotional tone (*Tone*) was significantly present, positively correlating with PF, indicating stronger performance than human counselors. In terms of *Social Processes*, GPT-4o used significantly fewer fillers (*Filler*) than human counselors; since this metric is negatively correlated with PF, this suggests superior performance. In *Punctuation* usage, GPT-4o used significantly fewer question marks (*QMark*) than human counselors; since this metric is positively correlated with PF, this indicates weaker performance. However, GPT-4o used significantly more periods (*Period*) and overall punctuation (*AllPunc*) than human counselors; since the former is negatively correlated with PF and the latter is positively correlated, it shows that GPT-4o is weaker in period usage but stronger in overall punctuation usage. In *Cognitive Processes*, GPT-4o used significantly fewer negations (*Negate*) than human counselors; since this metric is negatively correlated with PF, this suggests that GPT-4o's more cautious use of negation positively impacts PF. Additionally, in terms of *State* category, the expression of acquisition (*Acquire*) was not significant, indicating comparable performance to human counselors.

In the *Action Stage*, GPT-4o exhibited significant differences in several aspects. In *Emotion* category and *Cognitive Processes*, GPT-4o used significantly fewer all-or-none expressions

(*Allnone*) than human counselors; since this metric is negatively correlated with PF, this indicates that GPT-4o's more restrained use of all-or-none expressions positively impacts PF. In *Social Processes*, GPT-4o used significantly fewer gender-related language (*Male*) and conflict expressions (*Conflict*) than human counselors; since these metrics are negatively correlated with PF, this indicates that GPT-4o's more cautious handling of gender-related language and conflict expressions helps improve its PF. In *State* category, GPT-4o used significantly fewer expressions of need (*Want*) than human counselors; since this metric is negatively correlated with PF, this indicates that GPT-4o is less restrained than human counselors in expressing needs, negatively impacting PF. However, GPT-4o used significantly more social language (*Social*) than human counselors; since this metric is negatively correlated with PF, it indicates inferior performance in social expression. In *Punctuation* usage, GPT-4o used significantly more overall punctuation (*AllPunc*) than human counselors; since this metric is positively correlated with PF, it shows that GPT-4o has a relative advantage in using more diverse and varied punctuation. However, GPT-4o used significantly fewer question marks (*QMark*) than human counselors; since this metric is positively correlated with PF, it indicates that GPT-4o's less frequent use of question marks negatively impacts its PF rating. Additionally, GPT-4o used significantly fewer negation words (*Negate*) and gender pronouns (*SheHe*) than human counselors, suggesting that GPT-4o's more cautious use of gender pronouns and negation words positively impacts PF. Furthermore, there were no significant differences between GPT-4o and human counselors in *Emotion* category (*Feeling*), *Cognitive Processes* (*Cause*), *Culture* category (*Ethnicity*), *Time Orientation* (*Time*), and *Physical* topics (*Death*), indicating comparable performance in these areas. These findings address *RQ2*.

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

This section describes the development process of manually customized CoT prompts and presents a comparative analysis of PF 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 user-perceived Feedback

Effective prompt engineering involves considering multiple factors comprehensively. 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 user-perceived Feedback 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	Social	Increase
			conflict	Decrease

		interaction and avoiding	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 the appropriate use of 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	allnone	Decrease
			cause	Decrease

Preprint Publications

	caution.		
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 PF Between GPT-4o and Human Counselors

The study developed customized CoT prompts using the described framework, re-generated GPT-4o responses, and performed an automatic evaluation of users' PF scores in ESCs for human counselors, GPT-4o with standard CoT prompts, and GPT-4o with manually customized CoT prompts. Since GPT-4o's PF ratings were approximately normally distributed, paired sample t-tests were used to evaluate differences in PF ratings among the three groups. To assess the practical effect of the optimized prompt engineering on GPT-4o's performance, the study 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 exhibited significant improvements over those with non-customized prompts, demonstrating a substantial impact on GPT-4o's PF score. 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 disgust but showed at least small effect size improvements ($p < 0.05$) in all other subdomains. Notably, the effect size 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 PF Between GPT-4o and Human Counselors

Since the PF scores from human counselors are non-normally distributed, the study 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 shown in **Table 9**.

Regarding experience types, GPT-4o's performance was not significantly inferior to that of human counselors, suggesting GPT-4o handles different experience types comparably. For problem types, GPT-4o only showed a small effect size advantage (d = -0.13) in handling breakups with a partner. There were no significant differences from 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, GPT-4o, after manual CoT optimization, did not show a statistically significant difference in PF compared to human counselors. This indicates that GPT-4o has achieved a comparable level of emotional support to human counselors in these specific emotional contexts.

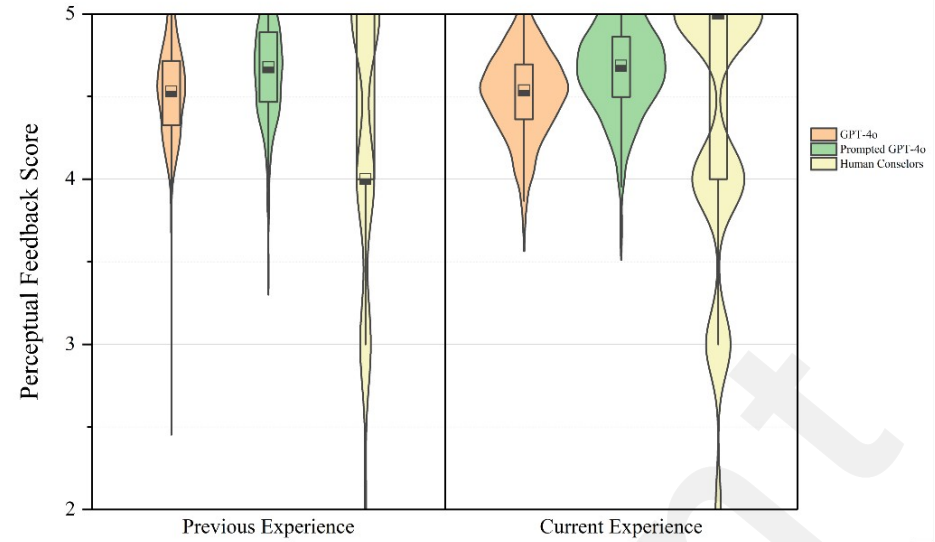
Table 9: Comparison of Perceived Feedback Scores Between Optimized GPT-4o and Human Counselors

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	Issues with Friends	251	5.0□4.0□5.0□	4.69□4.504□4.877	29672	0.254	-0.06

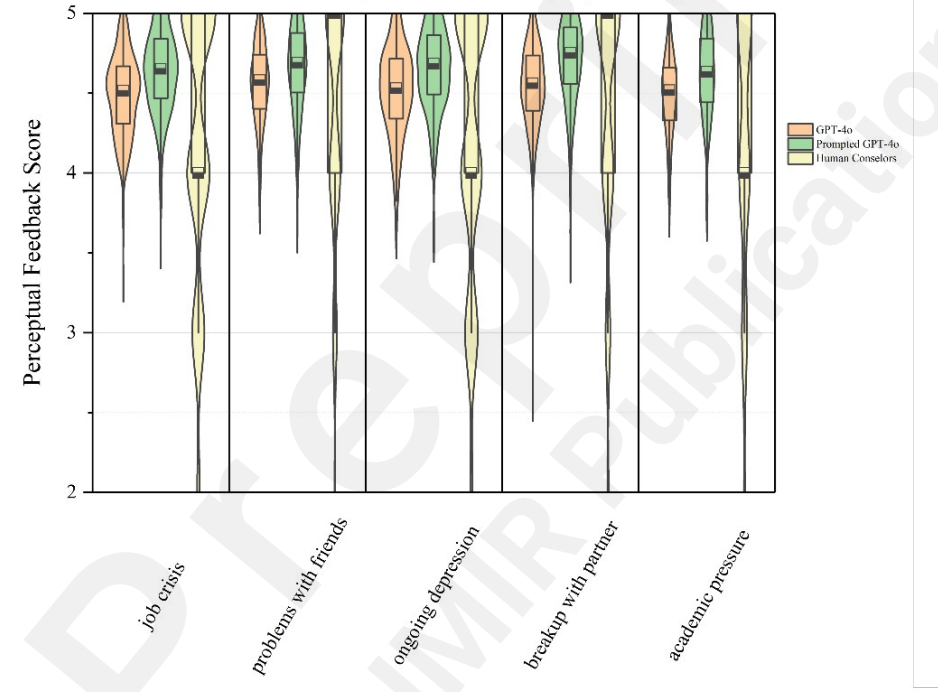
Type	Job Crisis	465	4.0□4.0□5.0□	4.652(4.469,4.84)	113515	0.183	0.05
	Ongoing Depression	451	4.0□4.0□5.0□	4.684(4.493,4.863)	107121	0.162	0.05
	Breakup with Partner	328	5.0□4.0□5.0□	4.752□4.558□4.91□	46829	0.004	-0.13
	Academic Pressure	200	4.0□4.0□5.0□	4.633□4.45,4.84□	21328	0.18	0.077

To evaluate the relative performance of human counselors and GPT-4o with manual CoT prompts in terms of Perceived Feedback (PF), the study analyzed the distribution of PF ratings across different ESC types. Figure 4 shows that human responses have a wider rating range, reflecting the diversity and variability in handling dialogue tasks. In contrast, GPT-4o's ratings are more concentrated, indicating greater consistency and predictability in response quality. Second, GPT-4o's PF ratings are concentrated between 4.3 and 4.6, showing a more clustered and smoother distribution. This clustering trend also reflects the overall improvement in GPT-4o's PF performance after applying manual CoT prompts. Third, GPT-4o lacks scores in the highest satisfaction range (5 points) compared to human counselors, indicating shortcomings in demonstrating high-level empathetic abilities.

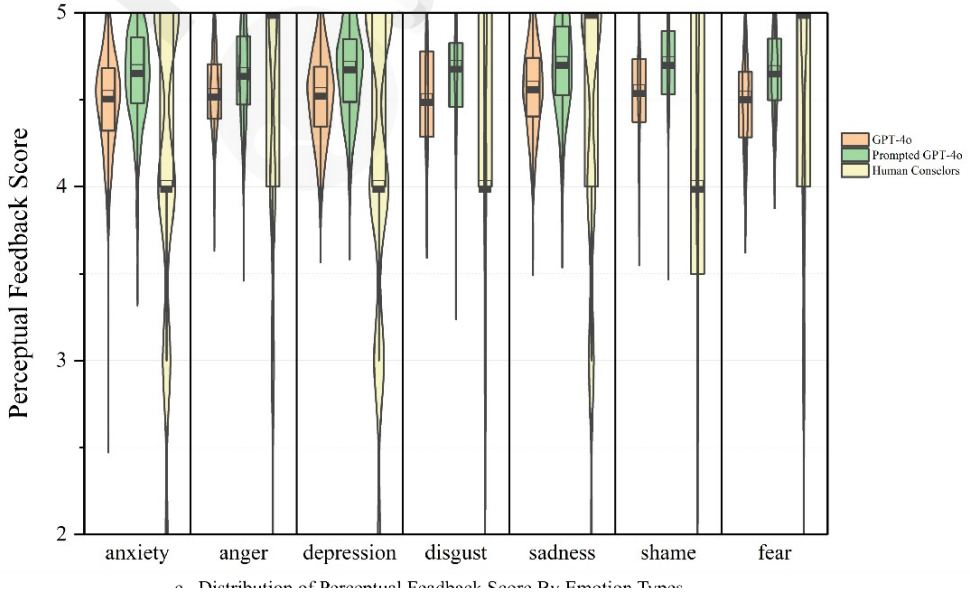
Preprint JMIR Publications



a. Distribution of Perceptual Feedback Score By Experience Types



b. Distribution of Perceptual Feadback Score By Problem Types



c. Distribution of Perceptual Feadback Score By Emotion Types

Preprint JMIR Publications

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

4.3.4 Comparison Analysis of the Key Linguistic Cues 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 PF in ESCs, the study conducted paired sample t-tests to compare linguistic clue features among human counselors, GPT-4o prompted by standard 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 real ESC case analyzed using SHAP local interpretability.

Table 10: Comparison of Average Values for Features Affecting PF 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)c
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, the results indicate that *Prompted GPT-4o* uses the term *feeling* significantly more often than both humans counselor and *GPT-4o* in the *Affect* category, suggesting a potentially excessive emphasis on emotional content, which may be less effective than human counselors and *GPT-4o*. In the *Linguistic Dimensions* category, the use of third-person singular pronouns (*she/he*) is significantly less than that of humans, indicating a relative advantage over human counselors in this regard. In the *Social Processes* category, *Prompted GPT-4o* outperforms both human counselors and *GPT-4o* in using social and conflict-related language, and it uses gender-related language (*male*) significantly less than humans, demonstrating superiority in these social process metrics. In the *Culture* category, the usage of ethnicity-related language is less than that of humans counselor and *GPT-4o*, further highlighting its cultural sensitivity. In the *States* category, the language related to *want* is used more effectively by *Prompted GPT-4o* compared to *GPT-4o* but less effectively than by human counselors. In the *Time Orientation* category, *Prompted GPT-4o* performs better than both human counselors and *GPT-4o* in using time-related language. In the *Cognitive Processes* category, language related to *cause* is used more frequently than by humans counselor but less than by *GPT-4o*, indicating better performance in cognitive processes compared to *GPT-4o* but weaker compared to human counselors. In the *Punctuation* category, the use of question marks (*QMark*) is less than *GPT-4o* but more than humans, suggesting better performance compared to human counselors but not as good as *GPT-4o*. The overall use of all punctuation (*AllPunc*) is

less than by humans counselor but more than *GPT-4o*, indicating superior punctuation usage compared to *GPT-4o* but still not matching human counselors.

In the Comforting phase, results show that *Prompted GPT-4o* performs as follows across several intrinsic evaluation metrics: In the *Linguistic Dimensions* category, the use of *pronouns* is significantly lower than *GPT-4o* but significantly higher than humans, suggesting better performance than *GPT-4o* but less effective than human counselors. The use of *dictionary coverage* (*dic*) is significantly higher than humans counselor but significantly lower than *GPT-4o*, indicating weaker performance in using standardized and professional vocabulary compared to *GPT-4o* but better than human counselors. The use of *numbers* is significantly lower than humans counselor but higher than *GPT-4o*, showing better avoidance of excessive technical or specific numerical expressions compared to human counselors, though potentially weaker than *GPT-4o*. The use of *adverbs* is significantly lower than *GPT-4o* but significantly higher than humans, indicating better performance than *GPT-4o* but weaker compared to human counselors. The linguistic *quantity* is not significantly different from humans counselor but higher than *GPT-4o*, suggesting it performs worse in this metric compared to *GPT-4o* but is on par with human counselors. The use of *conjunctions* (*conj*) is higher than *GPT-4o* but significantly lower than humans, indicating better performance in conjunction usage compared to human counselors but weaker than *GPT-4o*. The use of third-person plural pronouns (*they*) is less than both humans counselor and *GPT-4o*, indicating significant superiority over *GPT-4o* and human counselors in this metric. Expressions related to *fulfill* are significantly less than both humans counselor and *GPT-4o*, showing significant superiority over *GPT-4o* and human counselors in this regard. In the *Psychological Processes* category, the use of *verbs* is lower than *GPT-4o* but significantly higher than humans, indicating significant superiority over human counselors but less than *GPT-4o*. In the *Affect* category, positive tone expression is higher than both humans counselor and *GPT-4o*, showing significant superiority over *GPT-4o* and human counselors. In the *Social Processes* category, the use of third-person singular pronouns (*she/he*) is lower than both humans counselor and *GPT-4o*, suggesting better performance compared to *GPT-4o* and human counselors. In the *Conversational* category, the use of *fillers* is less than *GPT-4o* and not significantly different from humans, indicating better performance compared to *GPT-4o* but similar to human counselors. In the *States* category, expressions related to *acquire* are more frequent than humans counselor but not significantly different from *GPT-4o*, indicating weaker performance compared to human counselors. In the *Punctuation* category, the use of question marks (*QMark*) is less than *GPT-4o* but significantly more than humans, suggesting better performance in conjunction usage compared to *GPT-4o* but weaker than human counselors. The use of *periods* is higher than *GPT-4o* but significantly lower than humans, indicating better performance compared to human counselors but weaker than *GPT-4o*. The overall use of all punctuation (*AllPunc*) is significantly less than humans counselor but higher than *GPT-4o*, indicating superior performance compared to *GPT-4o* but still less effective than human counselors. In the *Cognitive Processes* category, the use of *negation words* (*negate*) is significantly reduced compared to *GPT-4o*, but still notably higher than that of human counselors, indicating that while it performs better than *GPT-4o* on this metric, it is weaker compared to human counselors.

In the Action phase, the results indicate that *Prompted GPT-4o* performed as follows across various intrinsic evaluation metrics categories: In *Emotional Expression* (*Affect*), the use of *feeling*-related language increased compared to *GPT-4o* and also showed a slight increase compared to human consultants. However, due to its negative correlation with user-perceived feedback (*PF*), this suggests that *Prompted GPT-4o* is weaker in this metric compared to human consultants and *GPT-4o*. In *Cognitive Processes*, the use of extreme expressions (*allnone*) such as "all" or "none" was significantly reduced compared to *GPT-4o*, but increased relative to

human consultants, indicating that *Prompted GPT-4o* performs better than *GPT-4o* but is weaker than human consultants. The use of *causal relationships* (*cause*) was less in *Prompted GPT-4o* compared to both human consultants and *GPT-4o*, suggesting that *Prompted GPT-4o* performs better than both. In *Social Processes*, the use of language related to *gender* (*male*) and *conflict* (*conflict*) was less compared to both human consultants and *GPT-4o*, indicating better restraint and adaptability in these aspects. In *Culture*, the use of language related to *ethnicity* was less compared to both human consultants and *GPT-4o*, suggesting that *Prompted GPT-4o* outperforms both in this metric. In *States*, the use of expressions related to *want* was significantly reduced compared to *GPT-4o* and slightly increased compared to human consultants, indicating that *Prompted GPT-4o* performs better than *GPT-4o* but is weaker than human consultants. In *Time Orientation*, the use of time-related expressions was reduced compared to human consultants and increased compared to *GPT-4o*, suggesting that *Prompted GPT-4o* performs better than human consultants but weaker than *GPT-4o*. In *Punctuation*, the use of all punctuation marks (*AllPunc*) increased compared to *GPT-4o* but significantly decreased compared to human consultants, indicating that *Prompted GPT-4o* is better than *GPT-4o* but weaker than human consultants. The use of *question marks* (*QMark*) decreased compared to *GPT-4o* and increased significantly compared to human consultants, indicating that *Prompted GPT-4o* performs better than *GPT-4o* but weaker than human consultants. In *Linguistic Dimensions*, the use of third-person singular pronouns (*she/he*) was less compared to both human consultants and *GPT-4o*, suggesting that *Prompted GPT-4o* performs better in this metric. In *Physical* topics, the use of topics related to *death* was less compared to both human consultants and *GPT-4o*, indicating that *Prompted GPT-4o* outperforms both in this aspect. Overall, GPT-4o with manually customized CoT prompts showed significant improvement in extrinsic evaluation metrics, achieving comparable perceived feedback scores to human counselors in certain areas. Regarding intrinsic metrics, Prompted GPT-4o outperformed both human counselors and GPT-4o in emotional expression, social language use, and positive tone across the Exploration, Comforting, and Action stages. Additionally, it showed greater restraint and adaptability in addressing gender and race-related language and avoiding conflictual expressions. These findings address *RQ3*.

5. Discussion

This study collected and analyzed responses from GPT-4o to over 1,300 real emotional support conversations. Following an integrative modeling paradigm in computational social science, the study employed machine learning, deep learning, and NLP methods to create an automatic evaluation model for assessing GPT-4o's user feedback scores in ESCs. An explainable model was then developed to identify key psychological language clues affecting perceived feedback scores, based on Hill's three-stage helping skills model. An integrated evaluation framework for ESCs was also proposed.ESCs. Additionally, customized Chain-of-Thought prompts were developed based on intrinsic and extrinsic metrics, and GPT-4o's responses were compared with those of human counselors.

5.1 Principal Findings

The results show that GPT-4o exhibits a notable level of emotional support capability. Manual CoT prompts, developed using the user feedback evaluation framework, significantly improve GPT-4o's performance in ESCs. Overall, the methods employed in this study enhance the performance of large language models in ESCs and improve the transparency and interpretability of the optimization process. By integrating the computational social science paradigm of integrative modeling, this study presents an integrative evaluation framework for

user experience in ESCs involving generative artificial intelligence. This framework includes intrinsic metrics (key linguistic cues) and extrinsic metrics (perceived feedback scores), effectively bridging and combining human expert standards with large language model capabilities. 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

This study first compared GPT-4o's performance with human counselors in emotional support conversations based on user feedback scores. The results indicate that GPT-4o generally falls short of human counselors, especially in managing emotions like anxiety, depression, and fear. GPT-4o also significantly underperforms 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 in its effectiveness when dealing with specific emotional responses and stressful situations.

Secondly, the study found that GPT-4o provides emotional support comparable to 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[37]. This discrepancy may stem from the need for deeper self-reflection and personal experience. 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 fully develop. This likely reflects limitations in the quality and diversity of training data and model structures for generative large language modelss[32], which may impact their creativity and ability to manage complex dialogues, thereby constraining their performance in higher scoring ranges.

Thirdly, the study explored GPT-4o's emotional support capabilities through prompt engineering. The experiments revealed that, with the exception of disgust, customized prompts significantly enhanced GPT-4o's performance in most scenarios. Additionally, GPT-4o with manually customized prompts achieved perceived feedback levels comparable to those of human counselors in managing specific emotions (e.g., anger, anxiety, shame, and disgust), showcasing the potential and effectiveness of generative AI in emotional support." Fourthly, the study assessed the stability of perceived feedback scores for generative AI and human counselors. It was 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. o organize and generate content in a standardized manner.[91] to organize and generate content in a standardized manner. This results in a more uniform approach to similar issues even without preset guidelines, leading to more consistent responses from GPT-4o.

5.1.2 Integrated Evaluation Framework for User Experience in Emotional Support Conversations

To address the lack of user-centered metrics for evaluating generative AI capabilities in ESCs and the inconsistencies between intrinsic and extrinsic metrics, this study employs an integrative modeling approach. It begins by identifying key linguistic clues and their impact on user-perceived feedback scores, using these clues as intrinsic evaluation metrics to assess generative AI's performance in ESCs.

In the Exploration stage of ESC, 13 metrics from 10 distinct psychological linguistic clue categories are utilized. 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 the situation, thus enhancing the dynamism and engagement of the discourse. This aligns with existing research, as LIWC effectively captures emotional and cognitive cues in language, revealing aspects such as confidence, motivation, and needs[92,93]. Additionally, metrics such as Emotional Expression, Social Dynamics, Gender Issues, Cultural Differences, and Cognitive Barriers evaluate the psychological stress and resistance faced by the recipient when addressing personal and social issues [94,95].
In the Comforting stage, 17 metrics from 7 distinct psychological linguistic clue categories are utilized. These include metrics for the use of verbs and positive tone, which measure the degree of emotional support and psychological comfort provided by the consultant, as well as the use of specialized vocabulary to assess the specificity and depth of the language. This is consistent with existing research, which indicates that LIWC analysis effectively captures emotional support and language specificity[96]. Additionally, metrics such as abstract and indirect language clues, along with communication hesitations, evaluate how responses may diminish personalization and directness in interactions, potentially creating barriers to emotional connection and understanding[92,93].
In the Action stage, 11 intrinsic metrics from 10 distinct psychological linguistic clue categories are utilized. 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[96]. 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[94,95]. These intrinsic evaluation metrics not only resolve inconsistencies between intrinsic and extrinsic metrics in existing user-centered ESC evaluation systems but also enhance understanding of how genAI influences user experience across various ESC stages. This approach supports the optimization of genAI system design and improves user interaction quality.
This study highlights the substantial influence of intrinsic and extrinsic evaluation metrics on consultation feedback across various stages of ESC. Firstly, during the Exploration stage of Hill's three-stage helping skills model, where consultants identify issues through questioning and discussion, the study 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[97] . 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[98] . 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[99] . 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[100] . Finally, in the Action stage, positive indicators like the use of conflict in Social Processes and extensive punctuation suggest active engagement and problem-solving, which are associated with effective conflict management in consultations[101] . 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[97] . Understanding these relationships enhances our knowledge of the dynamics and outcomes of

the ESC process and offers valuable insights for optimizing generative AI 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 large language models is their response uncertainty[102] , which also highlights 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, this study develops an explainable model for user-perceived feedback scores and identifies pertinent intrinsic metrics. Furthermore, by applying a framework grounded in Hill's three-stage helping skills model and an integrated evaluation metrics framework, this study introduces manually customized CoT prompts, collects responses from GPT-4o, and assesses their effectiveness.

Firstly, concerning extrinsic evaluation metrics—perceived feedback scores—the experiments reveal that customized CoT prompts substantially improve 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 fields such as mental health, medical diagnostics, intelligent customer service, and beyond[103]. This advancement supports the broader adoption and development of generative AI.

Second, this study assesses GPT-4o with Auto-CoT prompts, GPT-4o with manually customized CoT prompts, and human counselors across a range of intrinsic metrics. In the exploration stage, the GPT-4o model with customized CoT prompts was found to outperform both the Auto-CoT prompted GPT-4o and human counselors on seven metrics, including language dimensions, social processes, cultural references, temporal positioning, and punctuation. Conversely, its performance was weaker on five metrics: emotional expression, states, cognitive processes, and specific aspects of punctuation. This may indicate that AI is more adept at simulating the characteristics of exploratory and open dialogues, which helps establish an objective perspective and a foundation for problem-solving. However, its poorer performance in emotional expression and cognitive processes may highlight limitations in comprehending and articulating complex emotions and intricate cognitive relationships[104].

In the comforting stage, customized CoT prompted GPT-4o surpasses Auto-CoT prompted GPT-4o and human counselors on 11 intrinsic metrics, including language dimensions, psychological processes, emotional expression, conversation, punctuation, and cognitive processes. However, its performance was less satisfactory on five metrics: pronouns, language measurement, *conjunctions*, *states,* and punctuation. In this stage, customized CoT prompted GPT-4o excels in *language coverage*, *verb usage*, and *tone*, demonstrating its ability to provide emotional support and express empathy. These metrics reflect AI's effective use of emotionally charged language, enhancing emotional depth and empathy. Nonetheless, its weaker performance in pronouns and conjunctions may indicate deficiencies in constructing more fluid and personalized dialogues[105].

In the Action stage, the GPT-4o model with customized CoT prompts excels across nine metrics, including cognitive processes, social processes, cultural references, language dimensions, physical topics, temporal positioning, and punctuation. Conversely, the model demonstrates weaker performance on seven metrics: emotional expression, cognitive states, cognitive processes, temporal positioning, and specific aspects of punctuation. In the Action stage, the customized CoT prompted GPT-4o's effective handling of causal relationships and gender- and conflict-related language may enhance its ability to generate solutions and resolve conflicts

during consultations. Moreover, the model's adeptness at addressing physical and temporal topics, such as death and time, underscores its capability to tackle significant and sensitive issues. However, its reduced effectiveness in conveying extreme emotions and managing temporal topics may reveal limitations in handling more complex or extreme scenarios[102]. Overall, manually customized CoT prompts have shown notable strengths in social processes, temporal positioning, cognitive processes, and emotional expression. However, they also exhibit weaknesses in emotional understanding, cognitive complexity, linguistic fluency, and managing extreme scenarios.narios These differences and limitations offer valuable insights and guidance for the future design and optimization of AI systems.

5.2 Limitations and Future Research

This study assessed GPT-4o's ability to provide emotional support, achieving a relatively low mean absolute percentage error (MAPE = 17.13%) with predictive models, which reflects strong prediction performance. Nonetheless, the model's limitations indicate that it has not fully captured authentic feedback from help-seekers. Future research should integrate manual evaluation methods to thoroughly assess genAI's emotional support capabilities and address potential ethical risks.

Additionally, the study highlights that help-seekers' perceptions of emotional support are crucial for evaluating quality. Although the study assesses the effectiveness of emotional support based on help-seekers' feedback, it acknowledges that the dimensions of emotional support are much broader and more complex than what current data can capture. Therefore, future research should aim to collect more comprehensive data to provide robust evidence for the application of LLMs in mental health services.

Third, the study did not address the instability of LLM outputs, which may be influenced by the specificity of the prompts[106,107]. Considering GPT-4o's engagement with a broad spectrum of emotions and issues in emotional support, future research should focus on optimizing prompts and minimizing reliance on specific ones to enhance the model's stability and effectiveness, thereby accelerating its practical adoption in the mental health field.

Fourth, despite primarily relying on machine learning models for evaluation and prediction, the results from the deep learning model BERT were less satisfactory. Moving forward, future research should investigate more tailored transformer models to improve and enhance assessment accuracy, aiming for more precise evaluations with advanced transformer architectures.

Lastly, and crucially, before deploying LLMs for ESC, it is essential to address several ethical considerations. These include concerns about data privacy and protection, algorithmic bias and discrimination, transparency and interpretability, and the lack of adequate ethical and legal frameworks[108]. Our research used data from sources such as Liu (2021)[9], which did not directly address the ethical challenges of data collection. We improved transparency and interpretability of LLMs in ESC through integrative modeling approaches. However, several critical ethical issues must be resolved before these findings can be practically applied. Firstly, although risks from direct application were mitigated, it is vital to ensure that real-world data complies with privacy laws and is properly protected during both collection and processing[109]. Secondly, inherent biases within the data must be carefully managed to prevent algorithms from amplifying these biases during analysis[57,110,111]. Additionally, to move from research to practical application, enhancing transparency and interpretability is necessary to build trust and understanding among stakeholders regarding the research process and outcomes[110]. Furthermore, as genAI technology evolves rapidly, there is an urgent need to continuously update and refine the relevant ethical and legal frameworks[108]. Comprehensive resolution of these ethical 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 ESC datasets, to propose an integrative evaluation framework for user-perceived feedback in ESCs. The performance of GPT-4o in emotional support tasks was assessed and compared with that of human counselors. The results indicate that, although GPT-4o's emotional support capabilities improved with CoT prompting, it still lags behind experienced human counselors in certain emotional domains. By employing an integrative modeling paradigm and the Hill's three-stage helping skills model within the computational social science, the study identified intrinsic evaluation metrics relevant to user experience in ESC and developed customized CoT prompt engineering, which significantly improved GPT-4o's user-perceived feedback scores. This suggests that while AI has made strides in emotional support, incorporating human experience through CoT prompt can significantly refine and enhance the capabilities of LLM. The findings of this study provide a crucial empirical foundation and direction for improving genAI applications in the field of emotional support service.

Acknowledgments

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. Study design was conducted by Yinghui Huang. Literature review was performed by Hui Liu, Wanghao Dong, and Yinghui Huang. Data analysis was carried out by Lie Li. Writing was done by Yinghui Huang, Hui Liu, Yuhang Dong, and Lie Li. Manuscript revision was handled by Hui Liu, Yingdan Huang. Funding acquisition was managed by Yinghui Huang and Yingdan Huang. 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://www.jmir.org/api/download?filename=2865b5a9c40458e64d7149b0c26d26e6.docx&alt_name=65435-968672-1-SP.docx 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. 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

28. 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

29. 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

30. 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

31. 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.

32. Van Dis EA, Bollen J, Zuidema W, Van Rooij R, Bockting CL. ChatGPT: five priorities for research. *Nature*. 2023;614(7947):224-226.

33. 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.

34. 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

35. 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

36. 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

37. 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.

38. 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

39. 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

40. 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

41. 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.

42. 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.

43. 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

44. 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

45. Sezgin E. Artificial intelligence in healthcare: Complementing, not replacing,

doctors and healthcare providers. *Digit Health*. 2023;9:20552076231186520. doi:10.1177/20552076231186520

46. 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

47. 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

48. 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

49. 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

50. 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

51. Picard RW. Affective computing: challenges. *Int J Hum-Comput Stud*. 2003;59(1-2):55-64. doi:10.1016/S1071-5819(03)00052-1

52. Salovey P, Mayer JD. Emotional Intelligence. *Imagin Cogn Personal*. 1990;9(3):185-211. doi:10.2190/DUGG-P24E-52WK-6CDG

53. 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.

54. 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

55. 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

56. 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

57. 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

58. 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

59. 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

60. 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.

61. 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

62. 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

63. 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

64. 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

65. 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.

66. 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

67. 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

68. 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

69. 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

70. 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

71. Gross JJ. Emotion regulation: Conceptual and empirical foundations. *Handb Emot*

Regul. 2014;2:3-20.

72. 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

73. 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

74. 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

75. 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

76. 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

77. 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

78. 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

79. 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

80. 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/

81. 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

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

83. 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.

84. 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

85. 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

86. 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

87. 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

88. 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

89. 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.

90. 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

91. 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.

92. 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

93. 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

94. 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

95. 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

96. 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

97. 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

98. 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

99. 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

100. 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

101. 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

102. Shiffrin R, Mitchell M. Probing the psychology of AI models. *Proc Natl Acad Sci*. 2023;120(10):e2300963120. doi:10.1073/pnas.2300963120

103. 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

104. 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.

105. Xu FF, Alon U, Neubig G, Hellendoorn VJ. A systematic evaluation of large language models of code. In: *Proceedings of the 6th ACM SIGPLAN International Symposium on Machine Programming*. ACM; 2022:1-10. doi:10.1145/3520312.3534862

106. 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

107. 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.

108. 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

109. 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

110. 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

111. 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

Preprint JMIR Publications

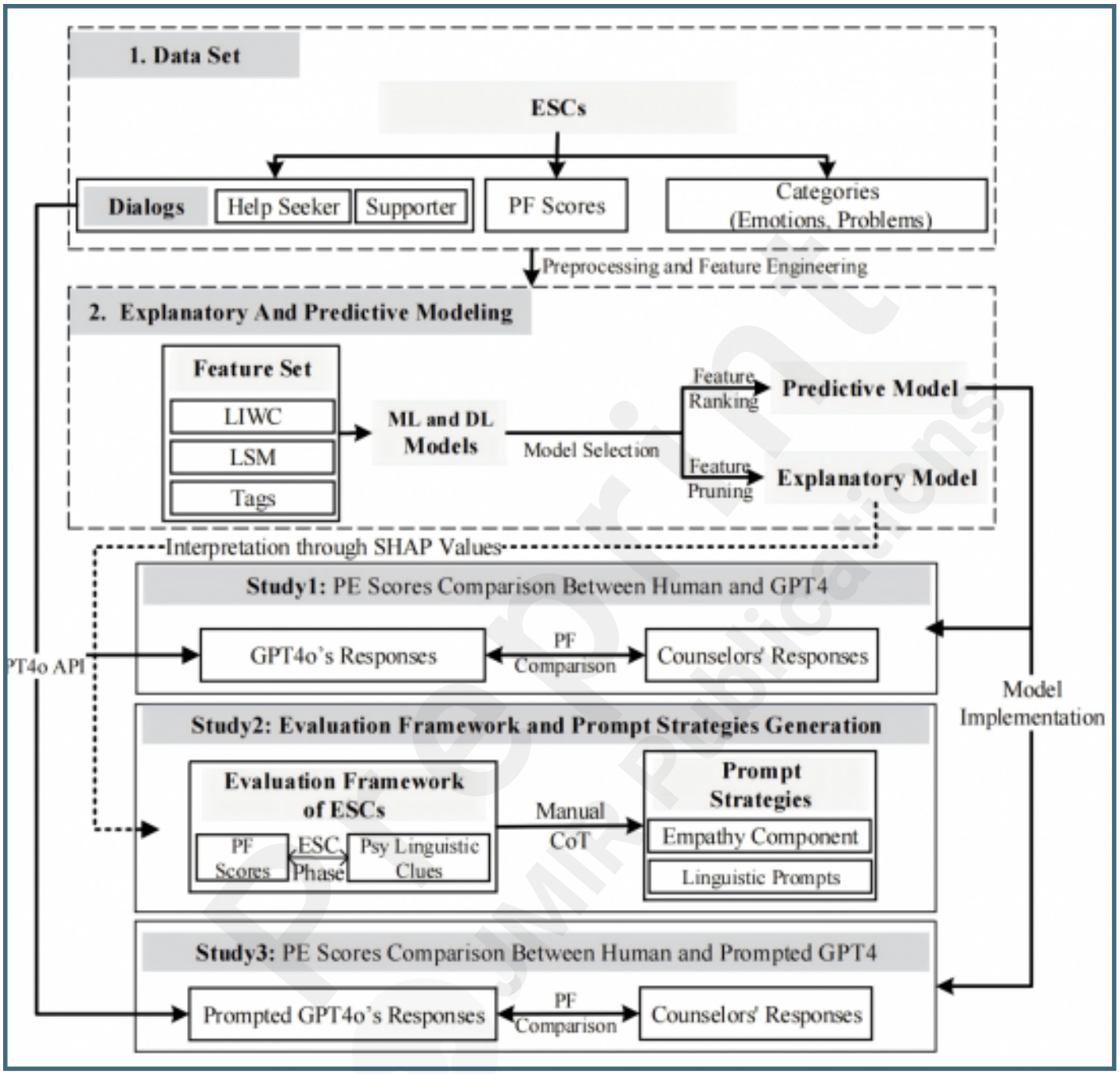
Supplementary Files

Preprint JMIR Publications

Figures

Preprint JMIR Publications

Research Methodology and Process.



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

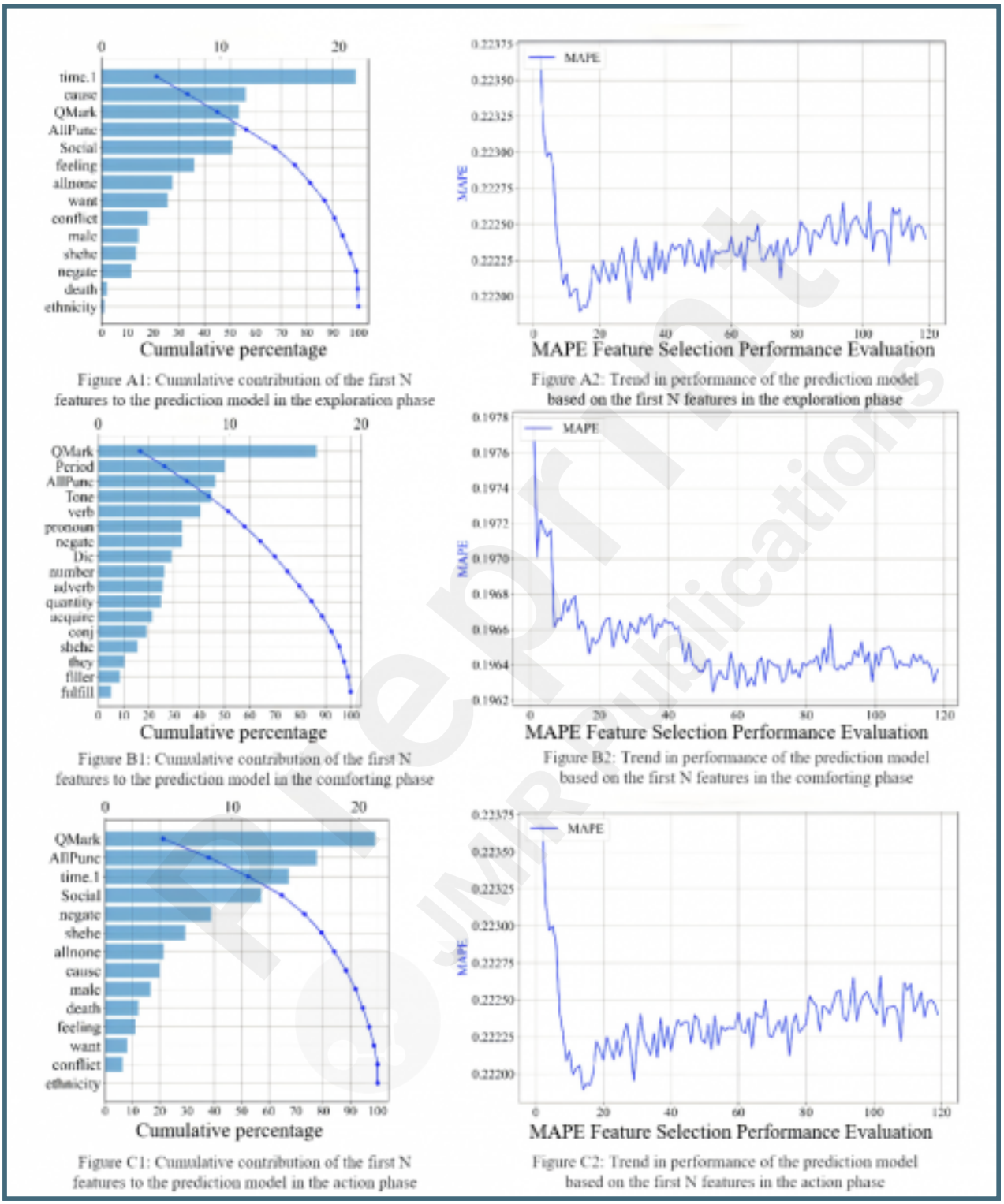


Figure A1: Cumulative contribution of the first N features to the prediction model in the exploration phase

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 Comprehensive Interpretability for Empathy Scores in ESCs.

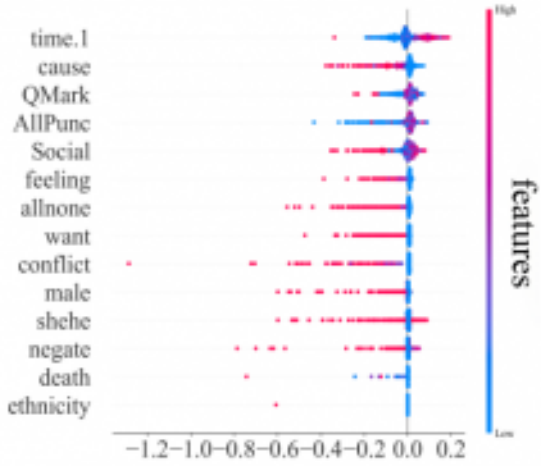


Figure a. SHAP graphs for exploration phase PF for global interpretation

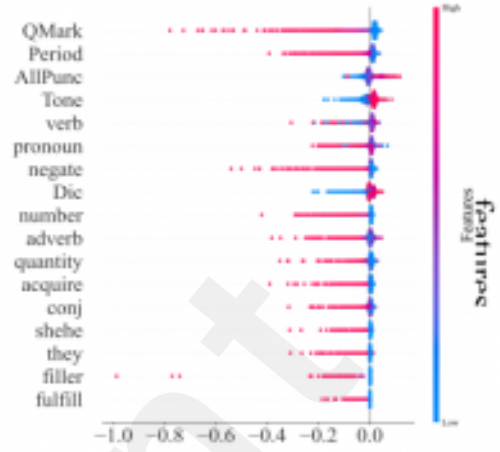


Figure b. SHAP graphs for comforting phase PF for global interpretation

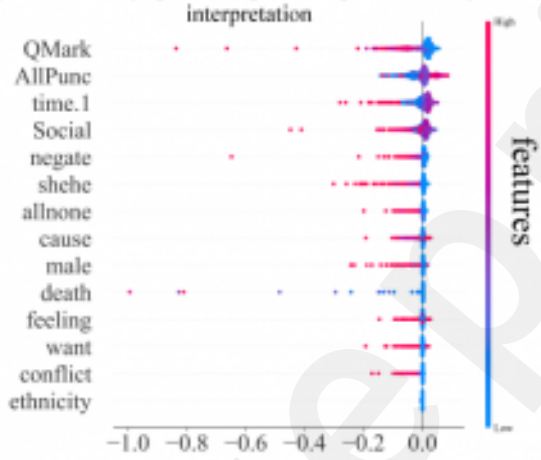
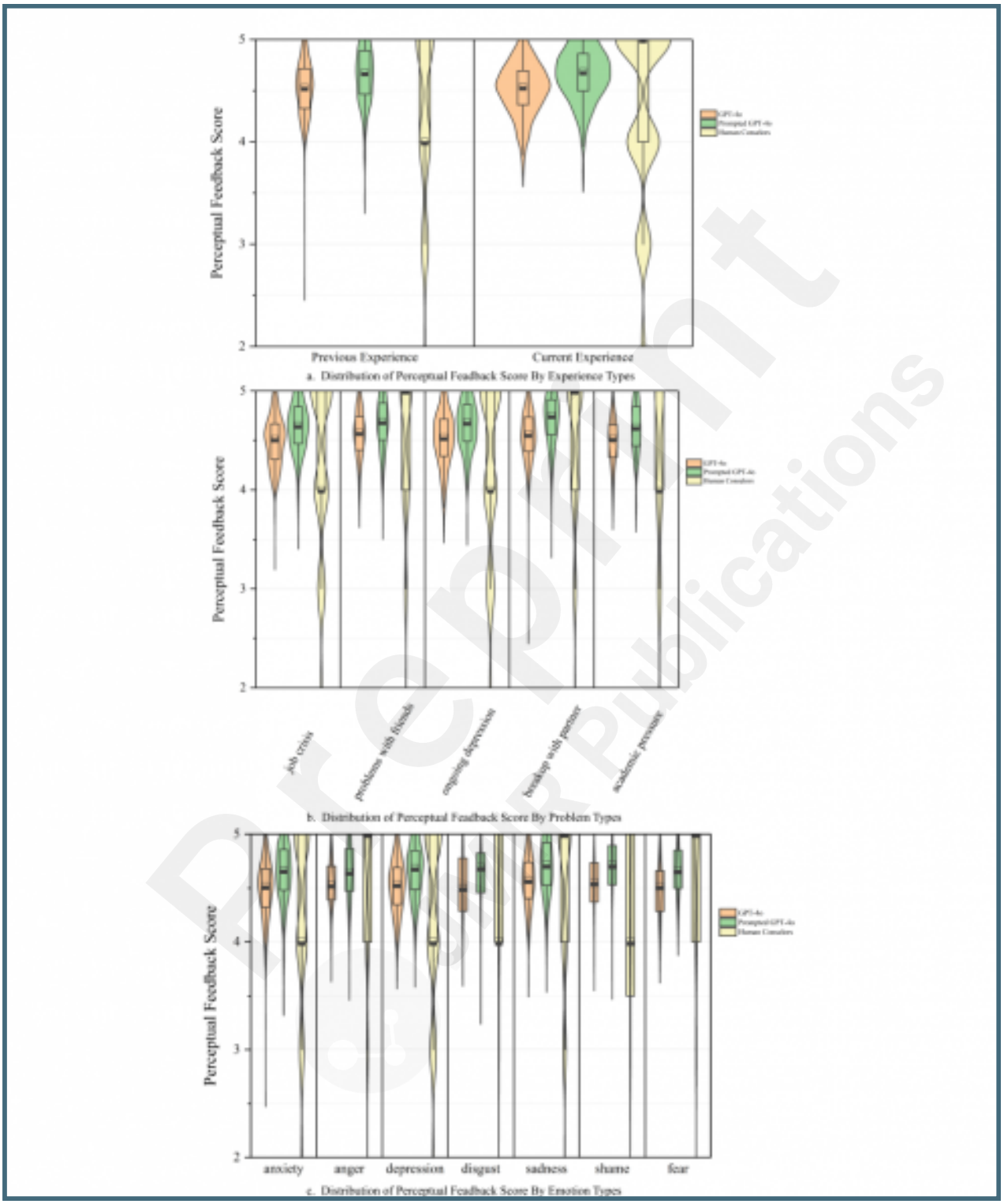


Figure c. SHAP graphs for action phase PF for global interpretation

Preprint JMIR Publications

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



Preprint JMIR Publications