Exploring Modular Prompt Design for Emotion and Mental Health Recognition

Minseo Kim*
Department of ELLT (English Linguistics & Language Technology)
Hankuk University of Foreign Studies
Seoul, Republic of Korea
er1123090@gmail.com

Taemin Kim*
Intelligent System
Hansung University
Seoul, Republic of Korea
taemin6697@hansung.ac.kr

Thu Hoang Anh Vo
School of Computing
KAIST
Daejeon, Republic of Korea
thuvo@kaist.ac.kr

Yugyeong Jung†
School of Computing
KAIST
Daejeon, Republic of Korea
yugyeong.jung@kaist.ac.kr

Uichin Lee‡
School of Computing
KAIST
Daejeon, Republic of Korea
uclee@kaist.ac.kr

Abstract

Recent advances in large language models (LLM) offered human-like capabilities for comprehending emotion and mental states. Prior studies explored diverse prompt engineering techniques for improving classification performance, but there is a lack of analysis of prompt design space and the impact of each component. To bridge this gap, we conduct a qualitative thematic analysis of existing prompts for emotion and mental health classification tasks to define the key components for prompt design space. We then evaluate the impact of major prompt components, such as persona and task instruction, on classification performance by using four LLM models and five datasets. Modular prompt design offers new insights into examining performance variability as well as promoting transparency and reproducibility in LLM-based tasks within health and well-being intervention systems.

CCS Concepts

• **Computing methodologies** → **Natural language processing**; • **Human-centered computing** → **Human computer interaction (HCI)**; • **Applied computing** → *Life and medical sciences.*

Keywords

Large language model, prompt engineering, emotion, mental health

ACM Reference Format:
Minseo Kim, Taemin Kim, Thu Hoang Anh Vo, Yugyeong Jung, and Uichin Lee. 2025. Exploring Modular Prompt Design for Emotion and Mental Health Recognition. In *CHI Conference on Human Factors in Computing Systems (CHI '25), April 26–May 01, 2025, Yokohama, Japan.* ACM, New York, NY, USA, 18 pages. https://doi.org/10.1145/3706598.3713888

*Equal contribution.
†Corresponding author
‡Corresponding author

ACM ISBN 979-8-4007-1394-1/25/04
https://doi.org/10.1145/3706598.3713888

1 Introduction

Large language models (LLMs) such as GPT-4 [46], Llama [63], and Gemini [12] have demonstrated remarkable capabilities across various tasks, including content generation [34, 36], problem-solving [80], and understanding human emotion and mental health [2, 44, 62]. *Prompts* are instructions provided to LLMs to enforce rules and ensure the quality of outputs, with *prompt engineering* playing a crucial role in maximizing the utility and accuracy of the models [9, 70]. In this work, we focus on the prompt engineering of mental health tasks, specifically the automated detection and categorization of emotional and mental health states in textual data including stress [2], anxiety [62], and depression [44]. LLM prompt design has gained attention in the field of HCI, because it serves as an enabler of novel intelligent health and wellbeing services like mental health counseling [51], mood journaling [26], and facilitating children's emotional sharing [56].

While these studies underscore the growing need for refined prompt design in the sensitive domains of emotion and mental health, a critical gap remains in designing and evaluating prompts for these tasks. Given that substantial variations in performance arise depending on prompt design [31, 38], recent work focuses on prompt strategies to enhance emotion and mental health analysis. However, open-ended prompt design in this domain makes it challenging to establish what constitutes quality prompts and whether such prompts are generalizable. This is a major departure from traditional machine-learning approaches with well-established analytical and optimization pipelines. While unique strategies for emotion and mental health classification have emerged, these are often fragmented, making systematic evaluation or reproducible prompt design difficult. Consequently, there is a lack of understanding regarding the key components of prompt design and their impact on task performance.

Furthermore, unlike traditional machine learning, where data scientists design and evaluate features and models, LLM-based service developers face the challenge of crafting and refining prompts to optimize system performance despite recent advances in LLMOps tools for prompt engineering [3, 27]. This shift in responsibility places more emphasis on the developers' ability to formulate, experiment with, and iterate on prompts. However, LLM-based service developers often struggle to develop high-performance, reliable prompts due to the complexity of the design space and the unpredictability of LLM behavior [15]. This challenge underscores the need for more systematic and reproducible methods for prompt design and its evaluation for emotion and mental health tasks.

Thus, our study addresses the following questions: *RQ1.* How can we define key components of LLM prompts for emotion and mental health tasks, such as automated detection and categorization of emotion, stress, and suicidal ideation status in textual data? *RQ2.* How can we apply modular prompt design to systematically evaluate LLM prompts? We conducted a comprehensive review of 30 existing studies and performed a thematic analysis to derive key components of modular prompt design for these tasks. To demonstrate its utility, as a case study, we evaluate two components, persona and task instruction style, on emotion recognition and mental health classification across five datasets. Our evaluation revealed significant performance variations. This indicates that there is no one-size-fits-all prompt design, suggesting further work on prompt optimization. We propose practical implications for prompt engineering to guide researchers in this domain.

Key contributions of this work include (1) a modular prompt design proposal for emotional and mental health tasks based on a literature review and thematic analysis of existing prompts and (2) a case study demonstrating the utility of modular prompt design through a systematic evaluation of its key components. The code for evaluation is available on GitHub.[1]

2 Background and Related Works

A substantial body of research in HCI and AI has focused on leveraging LLMs to understand emotions and mental health. These studies can be categorized into two main areas: optimizing LLMs or prompting strategies for emotional and mental health tasks [13, 53, 65] and designing LLM-based systems to promote emotional and mental health [26, 51, 56]. In the following, we review recent studies on model optimization and prompting strategies for emotional and mental health tasks.

Numerous studies have investigated optimized models for tasks related to stress [73], anxiety [68], depression [62], and suicide risk [82]. For instance, Mental-LLM optimized mental health prediction through zero-shot and few-shot prompting, as well as instruction tuning [71]. EmoLLMs focused on fine-tuning LLMs for a range of emotion analysis tasks [35], while MentalLlama emphasized both mental health prediction and the interpretability of LLM outputs [74]. In addition, researchers explored the design and evaluation of prompting strategies for emotion and mental health tasks. These strategies can be classified into two directions: 1) approaches that frame emotion recognition as a complex, multi-dimensional problem-solving task and 2) works that incorporate emotion awareness as part of the problem-solving process.

The first research direction [49, 82] adopted advanced techniques, such as in-context learning [6] and chain-of-thought prompting [69], enabling large language model to decompose the task into intermediate steps. Moreover, prior studies [32, 75] suggested an emotion-oriented in-context learning and chain of thought prompting, allowing models to perform a deeper analysis. Further, employing such prompting strategies enhances the models' capacity to follow emotionally intelligent reasoning processes [49], thereby enabling a more nuanced understanding and accurate classification of emotional states.

Meanwhile, the second research direction focused on incorporating emotional cues with prompts, improving LLM's awareness of emotions. For instance, incorporating emotional textual signals enhances LLMs' ability to interpret emotional content [65]. Similarly, emotional cues can improve models' emotional comprehension and responses [29]. Yang et al. [73] suggested the use of emotion-enhanced prompts to improve the model's ability to detect and interpret emotional cues, thereby enhancing both prediction accuracy and explainability in mental health tasks. These advances illustrate how prompt engineering can improve both emotional sensitivity and fairness in LLMs, making them more effective for applications like mental health analysis and emotion recognition.

Despite numerous studies proposing various approaches to prompt design for emotion and mental health analysis, a comprehensive understanding of the essential prompt components remains lacking. Recent reviews of prompt engineering only offer basic guidelines, e.g., a prompt pattern catalog composed of elements such as input semantics, output customizations, error identification, prompt improvement, interaction, and context control [70], as well as various prompting strategies, such as chain-of-thoughts, self-consistency, and prompt decomposition [9]. So far, regarding emotional and mental health as sensitive classification tasks, most research offers unique, fragmented strategies, leaving a gap in the systematic evaluation of what these components are and how they affect model performance. Our study modularizes key prompt components, drawing on a comprehensive prompt review by thematic analysis of previous literature and assessing their impact on performance, offering a structured framework for future research and application in these fields.

3 Modular Prompt Design

3.1 Method

3.1.1 *Paper selection process.*

We aimed to collect prompt examples specifically designed for emotion recognition and mental health analysis tasks using LLMs. To achieve this, we conducted a systematic literature review using three primary databases: ACM Digital Library, IEEE Access, and Google Scholar. A set of keywords was carefully selected, focusing on core concepts, i.e., "emotion recognition," "mental health," "LLM," and "prompt." Our initial search in ACM and IEEE yielded 348 and 203 records, respectively. After a full-text assessment, we only included the publications if they provided prompt examples, focused on text data, and evaluated the prompt for emotional and mental health analysis tasks. As a result, we have a total of 12 papers. To include wider sources and publication types, we expanded our search to Google Scholar, yielding an initial set of 9,260 papers. Among them, we only included papers that use LLMs to classify emotions or mental health status, resulting in 36 papers. We removed duplicated papers found in the ACM and IEEE databases, resulting in 31 papers. We then performed a

[1] https://github.com/Kaist-ICLab/Exploring-Modular-Prompt-Design-for-Emotion-and-Mental-Health-Recognition

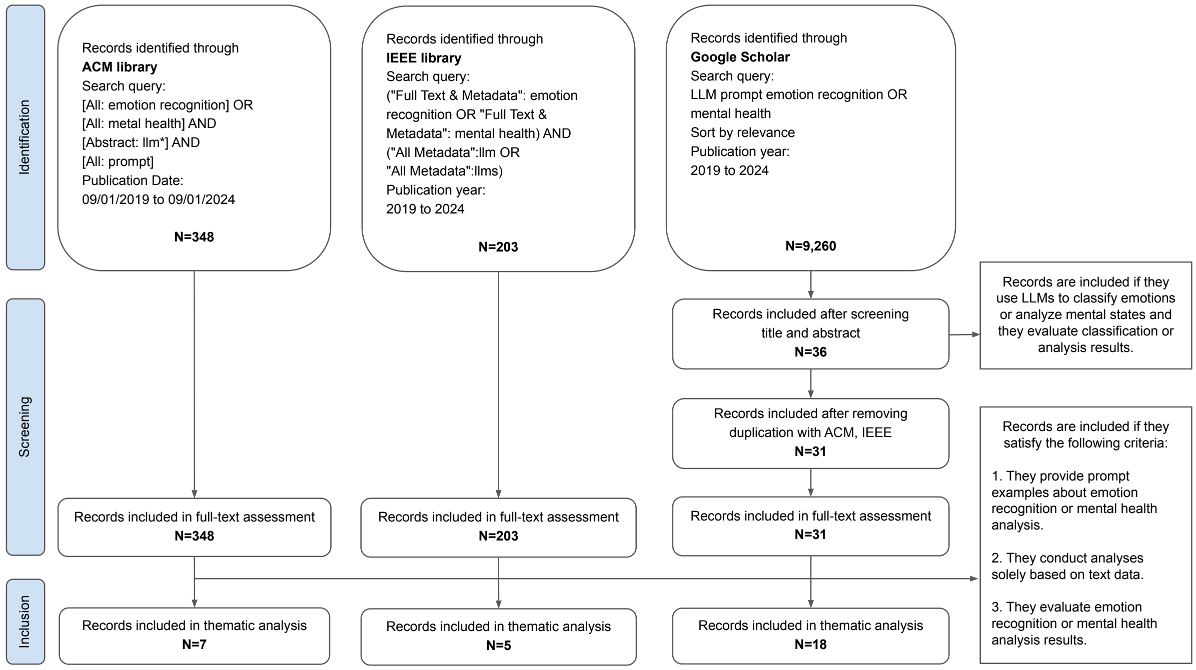

Figure 1: Flow diagram of paper search and selection process

full-text assessment of the papers using the same inclusion criteria applied to the ACM and IEEE sources. The number of papers at each stage, along with the inclusion criteria, is shown in Figure 1. Combining the results from all three databases (ACM, IEEE, and Google Scholar), a total of 30 publications were included in the thematic analysis (see Table 5 in Appendix). Of these, 14 focused on emotion recognition and 16 on mental health. We then extracted prompts designed for emotion recognition and mental health analysis tasks, resulting in a collection of 54 prompts that were used for thematic analysis.

3.1.2 Thematic Analysis. We conducted thematic analysis [11] of the prompts to find out the key components. Thematic analysis was conducted by two of the authors with HCI and computer science background. The first step of our analysis involved thoroughly reading all the data by two researchers. Some extracted prompt examples are provided in Table 7 in Appendix. In the second step, we conducted open coding. Here, we did not use pre-defined codes but instead developed and refined them gradually. Once coding was complete, we compared, discussed, and adjusted our codes. In the third step, we searched for patterns by grouping codes with similar functions. During the fourth step, we gathered prompt segments relevant to each category and re-examined whether they were truly aligned with the theme. Finally, we defined the key components and produced the final result. The code book is presented in Table 6 of Appendix.

3.2 Results

We uncovered six components that commonly construct prompt design: Persona, Task, N-shot examples, Input, Output, and Template. Among the 54 prompt samples we analyzed, Task Instruction and Input were the most frequently occurring components, each appearing in 54 samples. Output (in 23 samples) and Persona (in 22 samples) appeared less frequently but still significantly more often than N-shot Example and Template appearing in 7 and 5 samples, respectively. Figure 2 provides a comprehensive overview of the prompt structure with an example.

3.2.1 Persona. This component is used to define various aspects of a persona that could influence the model's behavior, including two elements, i.e., role and capability.

- *Role* instructs the AI to adopt a specific role or behave in a particular way. This can be used to adjust the tone, style, or depth of the information generated.
- *Capability* can describe skills, knowledge, and abilities that the persona possesses.

3.2.2 Task. This is the most important component, including (1) contextual information, (2) task knowledge, (3) task instruction, (4) step-by-step thinking, and (5) emphasis.

- *Context information* gives details about what the input is or the source of input (post, Twitter, diary, etc.).
- *Task knowledge* provides the model with domain-specific knowledge or background information that it can utilize to carry out the analysis. Prompt example: "*Generalised anxiety disorder is a mental health illness that is defined by people having feelings of excessive anxiety.*" [4].

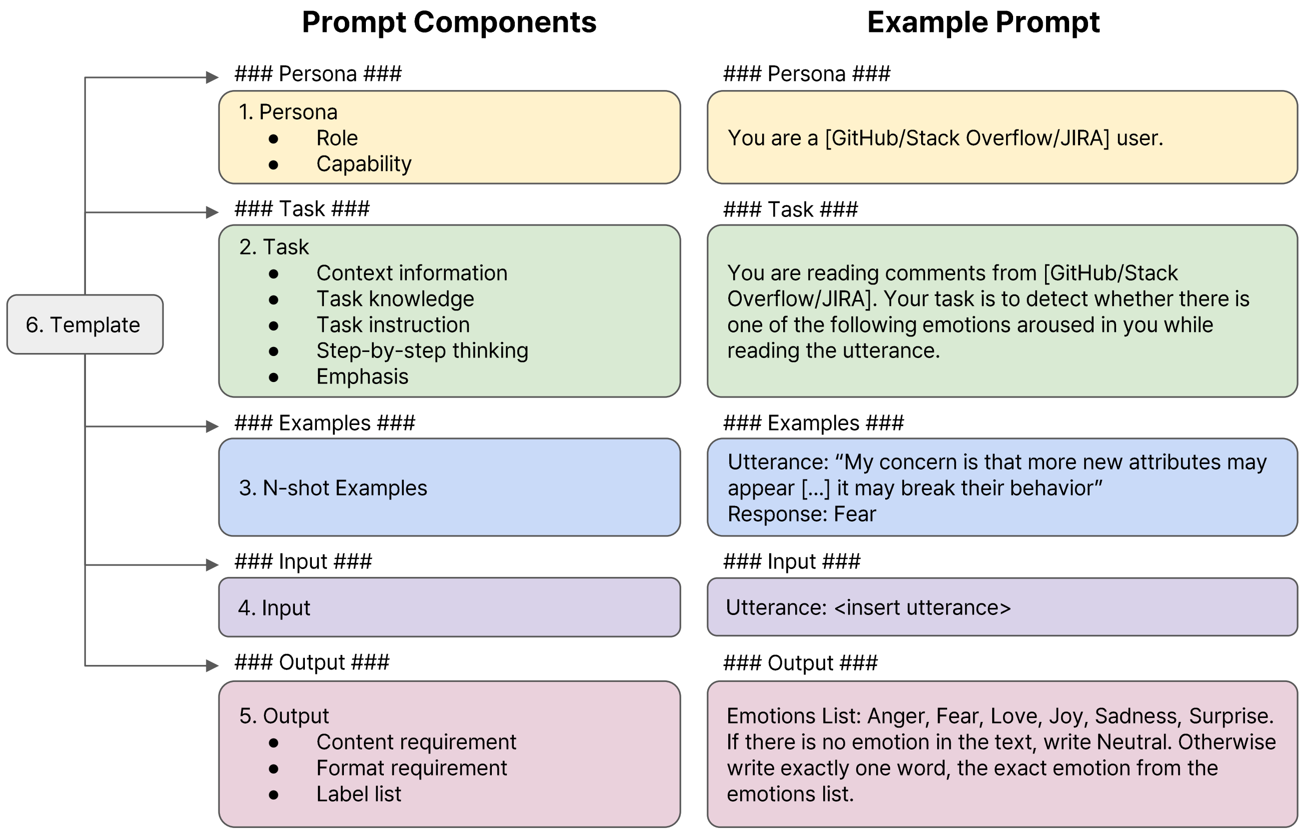

Figure 2: Key prompt components for LLM-based emotion and mental health tasks (left), along with example prompts to demonstrate how they are expressed to achieve a task (right).

- *Task instruction* is the main query that drives the task, instructions, or principles that direct how a task should be performed or approached. Prompt examples: "*Analyse the conversation to determine whether the respondent's emotional state is depression or anxiety.*" [62].
- *Step-by-step thinking* is used to break down tasks into sequential steps, allowing the model to approach problems methodically and systematically. Prompt example:
"*Let's think about it step by step:*
Step 1: Describe the content of the news.
Step 2: Think about emotional reactions...
Step 3: Think about how you need to express..." [32].
- *Emphasis* element or stimuli is used to emphasize the importance of the task. Previous studies in psychology have shown that expectancy, confidence, and social influence can beneficially impact individual performance. Prompt example: "*This is very important to my career.*" Li's study tried to apply this factor to LLM prompt and prove that it can lead to better performance [29].

3.2.3 N-shot Examples. N-short example provides examples to demonstrate how the model should handle similar tasks, helping the LLM generalize from the provided instances.

"*Example 1:*
Post: Does everyone else just hurt all the time It's not like physical pain or soreness, it's just this overwhelming feeling of exhaustion...
Response: Yes. Reasoning: The post conveys a deep sense of emotional pain, exhaustion, and numbness..." [74].

3.2.4 Input. Input is actual data or content submitted for the task, which could include sources like social media posts, diary entries, or conversational threads relevant to the analysis.

3.2.5 Output.

- *Content requirement* defines the essential information that must or must not be included in the output, ensuring that the model addresses all necessary elements of the task. For examples: "*The response should not imply negative emotions toward anyone or anything, such as disgust, resentment, discrimination, hatred, etc.*"[32].
- *Format requirement* specifies the format or structure that the output must follow to ensure consistency, clarity, and relevance in the model's response. "*Provide the answers in JSON format with the following columns: text, topic, risk level.*" [17].

- *Label list* is a predefined set of labels or categories that the AI can select from when generating outputs, ensuring standardized classification or tagging. Prompt example: "*Only from this emotion list: [Emotion List]. Only return the assigned word.*" [67].

3.2.6 Template. A predefined framework is used to structure the prompt, dividing it into sections or headings to ensure the model receives well-organized and clear instructions.
Prompt example: "*[System] ... [Context] ... [Prompt] ... [Response] ... [Criteria] ....* [32].

4 Evaluation of Modular Prompt Design in Model Performance: A Case Study

Based on the thematic analysis, we found that LLM prompts in the emotion and mental health domain can be modularized into six major components. This modular design implies that a systematic evaluation of each component's importance and performance is possible. In other words, it allows for independent analysis of how specific modules affect model performance and enables us to know if the removal or modification of the modules positively or negatively impacts the model performance. For components that significantly influence performance, it is possible to understand how variations in these modules affect the performance.

4.1 Evaluation Scope

We investigate how the presence, absence, or variation of each component in modular prompts affects model performance. While many candidate variations can be derived from the modules, there were constraints in time and resources to experiment with all of them. Therefore, as a case study, we selected two key components deemed most essential for this research: *persona* and *task instructions*. In the following, we elaborate on the rationale for choosing these two components and present experiments that explore their impact on performance in emotional and mental health tasks.

Persona refers to a set of characteristics, such as personality, style, and profession, that shape how the model generates responses to simulate a consistent behavior or identity [30]. Recent studies [22] claimed that persona prompting provides statistically significant improvements in LLM predictions, though the extent of improvements varies. As highlighted in Table 6, the high frequency of persona and its relevance to the domain suggest that persona is a critical factor in enhancing performance in mental health-related tasks. Our experience in mental and emotion recognition underscores the critical role of the LLM's persona module. For instance, psychiatrists contribute to labeling sensitive datasets, such as those for suicidal ideation [16], emphasizing the necessity of domain-specific knowledge for accurate and reliable recognition. As detailed in Table 8 of Appendix C, we developed an 'expert system' persona tailored to specific emotion and mental health domains to optimize performance.

Task instruction refers to the specific directions provided to the model to define and execute a given task. This module focuses on how tasks are described and executed, allowing for varied instruction styles based on the needs of the persona or context [76]. This module was selected due to the high frequency observed in the thematic analysis (see Table 6). We generated three variations within this module, considering previous findings in prompting and its possible relevance with the mental health domain. *Clear and Direct* is designed to provide straightforward, clear instructions based on prior LLM research [55]. This observation is consistent with principles from communication research [47] and effective interpersonal communication theories [58], which emphasize the critical role of clarity and directness in communication. We generated this variation by instructing the LLM (GPT-4o): "Provide simple, easy-to-follow instructions with concise language." The second variation, *Emotionally Descriptive*, enhances the emotional richness of the instructions, as emotions are intertwined with cognitive functions like attention and decision-making [23, 61]. Also, prior studies [42, 73] emphasized that emotion infusion in prompts can inspire LLM to concentrate on emotional clues, enhancing performance in an emotion-intensive setting. We instructed the LLM (GPT-4o) to "Incorporate vivid language and emotional depth, focusing on enhancing emotional aspects." The third variation, *Technical & Analytical*, uses expert terminology to align with professional communication standards to align with the domain-expert persona settings. As in prior work [45], we generated instructions focused on precise language relevant to psychology and mental health: "Use technical jargon and expert language suitable for professionals, with emphasis on analysis." The detailed variations of prompts are in Table 8 in Appendix.

4.2 Datasets

We used open datasets to evaluate the prompt components for emotion recognition and mental health analysis. For emotion recognition, we focused on complex emotion understanding and fine-grained classification, while the mental health dataset addressed stress, depression, and suicidal ideation. The datasets were selected based on task difficulty to assess prompt performance across varying complexity. We randomly selected 200 samples from each dataset for consistent evaluation, with dataset details and statistics provided in Table 10 in Appendix.

4.3 LLM Models

With advancements in LLM performance, there is increasing interest in evaluating relatively smaller models. We assess emotion and mental health capabilities by comparing both large and small models, examining how prompt variations affect performance across different model sizes. For large models, we use Google's API-based Gemini, available in versions such as Ultra, Pro, and Nano, optimized for various use cases. Specifically, we selected the Gemini-1.5-Pro-001 [52]. Additionally, we evaluated GPT-4o, regarded as the largest and highest-performing LLM at the time of writing, using the gpt-4o-2024-05-13 [1] version. For small open-source models, we used Alibaba's Qwen2-7B-Instruct [72], a model with 7B parameters, suitable for comparison against other open-source models. The base model for Qwen2-7B-Instruct is Qwen2-7B, which is pre-trained on a large-scale corpus and then instruction-tuned. We also selected Mistral-7B-Instruct-v0.3 [8], another instruction-tuned model based on the Mistral-7B architecture. We excluded Llama models from this study due to their limitations in safety restrictions, especially when dealing with suicide-related content.

4.4 Evaluation Metrics

The primary evaluation metrics used in this study are Accuracy and F1-macro, as in prior studies [17, 73]. After applying the same prompt technique to various models, we experimentally validated how these models perform and how prompt techniques behave in different models.

5 Results

This section summarizes the results of various prompt components, particularly the Persona and Task Instruction components. We first analyze the results to understand the individual effects of prompt components on model performance (Section 5.1). We then extend this analysis by examining how components interact with each other (Section 5.2). We set the baseline as prompt without Persona and Task Instruction, while all other components are included and fixed. We differentiate Persona and Task Instruction settings for a systematic evaluation. Table 9 in Appendix shows the component settings for each experiment and settings for the baseline prompt.

5.1 Impact of Individual Components

5.1.1 Impact of Persona on Model Performance. Figure 3 and Table 2 assess the impact of Persona across datasets. Applying Persona improved performance across four datasets, with varying degrees. For GoEmotions, GPT-4o improved by 0.82%, Gemini by 4.03%, and Mistral by 6.22%. In EmoBench, the improvement was minimal. In Dreaddit, GPT-4o and Gemini showed no significant changes, but Qwen2 and Mistral decreased by 10.8% and 15%, respectively. In SDCNL, all models except Gemini saw modest gains, while in the CSSRS-Suicide dataset, GPT-4o improved by 8.55%, Gemini by 1.76%, Qwen2 by 8.74%, and Mistral by 3.89%. These results suggest that Persona is particularly beneficial in tasks involving subtle labels, like suicide severity. Overall, both GPT-4o and Gemini showed consistent improvements with Persona, while smaller models like Qwen2 and Mistral displayed inconsistent results. While the impact of Persona is more pronounced in larger models, it remains a valuable component for enhancing performance.

5.1.2 Impact of Task Instruction on Model Performance. The results from Figure 4 and Table 3 clearly demonstrate the influence of Task Instruction on model performance. In the EmoBench, GoEmotions, and CSSRS-Suicide datasets, incorporating Task Instruction generally improved performance. However, in the Dreaddit dataset, adding Task Instruction tended to reduce performance. Given that Dreaddit is a binary classification task, it is likely that the complexity and length added by the Task Instruction negatively impacted performance. This suggests that overly complex prompts may hinder performance in simpler tasks such as binary classification. In EmoBench, the "Technical & Analytical" instruction was especially effective. When applied to GPT-4o, it resulted in a 5.92% performance improvement, while Gemini saw a 5.23% improvement. This highlights the ability of the "Technical & Analytical" prompt to significantly boost performance in large models, particularly in scenarios requiring complex emotional reasoning. The "Emotionally Descriptive" prompt had a strong positive impact on the CSSRS-Suicide dataset. GPT-4o and Gemini showed performance improvements of 18.40% and 10.58%, indicating enhanced label differentiation in suicide severity classification. Although Qwen2's performance declined on the SDCNL dataset, the accuracy drop was 3.59%, and the F1 drop was 3.56% less with the "Emotionally Descriptive" prompt compared to the "Clear & Direct" prompt. Moreover, the "Emotionally Descriptive" prompt helped mitigate the performance decline in Mistral, indicating that emotional prompts can alleviate performance degradation.

5.2 Impact of Component Interactions on Model Performance

Through Figure 5, we analyzed the impact of Task Instructions on LLM performance with and without the application of Persona, using Z-Scores. The Clear & Direct Task showed above-average performance in most datasets, with the most notable improvement seen in EmoBench. This suggests that providing clear, straightforward task instructions results in consistent improvement, especially when coupled with an expert persona.

Contrary to the expectation that combining two components would enhance performance, in some cases, we observed a decline in performance. In Table 4, the combination of Emotionally Descriptive task instruction with Persona-Expert shows a decrease in performance. For instance, in the case of GPT-4o, accuracy dropped by 2.21% and macro F1 by 2.09%, while for Gemini, accuracy decreased by 6.00% and macro F1 by 3.71% in the CSSRS-Suicide dataset. Several other datasets also showed a similar trend where the performance did not improve, instead the benefits of each component were offset. This indicates that when Persona and Task Instruction components are used together, their combined strengths are not always maximized. Performance may even fall below that of the Baseline + Task combination.

Additionally, Table 4 shows that the GPT-4o model consistently outperformed other models. Notably, in the absence of Persona, GPT-4o achieved the highest performance in three out of five datasets—EmoBench, Dreaddit, and CSSRS-Suicide. This indicates that even without the assistance of Persona, GPT-4o excels in handling complex emotional and mental health tasks. When Persona was applied, GPT-4o continued to achieve top performance across most datasets. It recorded the highest F1 scores in all datasets, except for GoEmotions. This implies that GPT-4o not only performs well without Persona but also benefits significantly from its application, improving its performance more than other models under the same conditions.

6 Discussion

We discuss how the modular prompt design can be applied in the HCI domain, providing detailed guidelines for researchers and practitioners. We also reflect on ethical and privacy considerations in modular prompt engineering.

6.1 Modular Prompt Design for Emotion and Mental Health Research in HCI

We proposed a modular prompt design for the emotion and mental health domain, grounded in a comprehensive thematic analysis of existing prompts. While previous studies have explored prompt strategies for tasks such as emotion recognition [5, 42], anxiety or

Figure 3: Comparison of F1-scores for 4 LLMs across 5 datasets (EmoBench, GoEmotions, Dreaddit, SDCNL, CSSRS-Suicide): the baseline vs. the combination of a persona component.

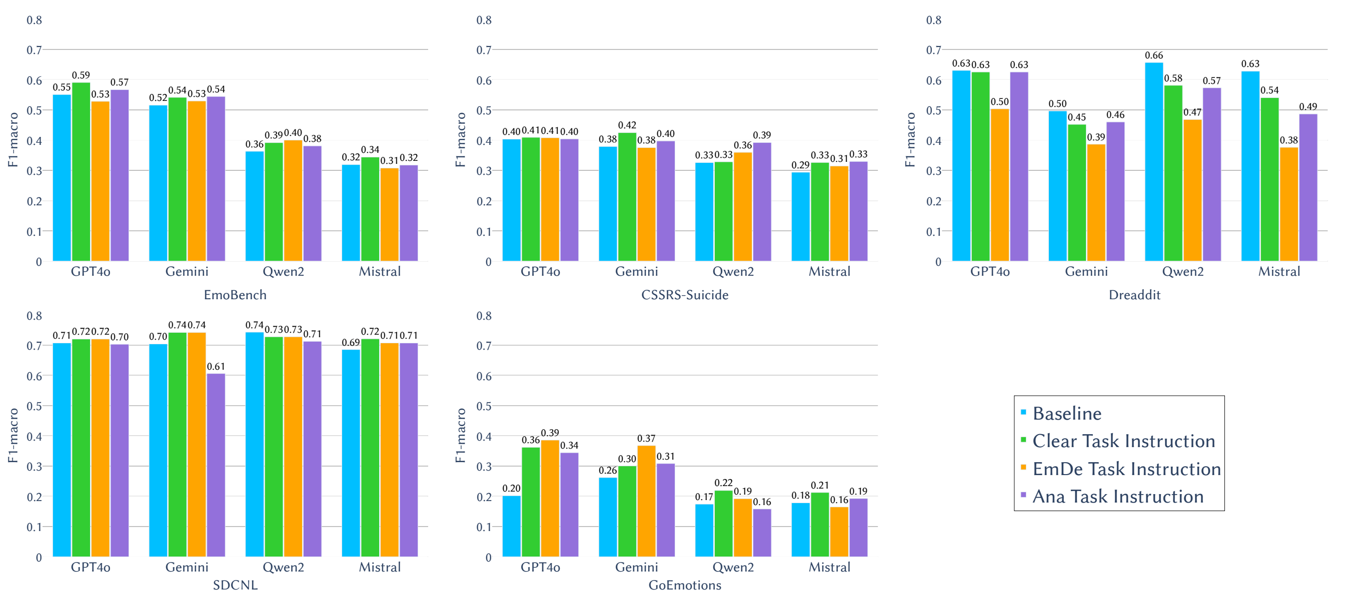

Figure 4: Comparison of F1-scores for 4 LLMs across 5 datasets (EmoBench, GoEmotions, Dreaddit, SDCNL, CSSRS-Suicide). Bars represent the baseline and the application of Task Instruction variations.

Persona	Model	Emobench		Goemotion		Dreaddit		SDCNL		CSSRS	
		Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
Baseline	Mistral	0.3367	0.3189	0.3196	0.2939	0.6533	0.6279	0.7035	0.6852	0.2700	0.1784
	Qwen2	0.3807	0.3631	0.3298	0.3258	**0.6800**	**0.6568**	**0.7437**	**0.7428**	0.2400	0.1741
	Gemini	0.5404	0.5155	0.3759	0.3789	0.5808	0.3966	0.7050	0.7037	**0.3116**	**0.2620**
	GPT4o	**0.5729**	**0.5508**	**0.4082**	**0.4033**	0.6650	0.6306	0.7071	0.7070	0.2950	0.2019
Baseline + Persona	Mistral	0.3586	0.3311	0.3636	0.3561	0.5700	0.4779	0.7172	0.7136	0.2800	0.2173
	Qwen2	0.3827	0.3731	0.3073	0.3014	0.6150	0.5480	**0.7626**	**0.7604**	0.3050	0.2615
	Gemini	0.5477	0.5217	0.4121	**0.4192**	0.5850	0.5036	0.6750	0.6720	0.3000	0.2796
	GPT4o	**0.5528**	**0.5326**	**0.4167**	0.4115	**0.6750**	**0.6392**	0.7186	0.7180	**0.3385**	**0.2874**

Table 2: Evaluation results of each model using Persona. The bold texts indicate the highest performance in terms of Accuracy and F1-score for each dataset.

Model	Task Instruction	Emobench		Goemotion		Dreaddit		SDCNL		CSSRS	
		Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
Mistral	Clear & Direct	+0.0237	+0.0251	+0.0289	+0.0322	-0.0483	-0.0875	+0.0201	+0.0361	0.0000	+0.0343
	Emotionally Descriptive	-0.0117	-0.0112	+0.0205	+0.0206	-0.0841	-0.1597	+0.0015	+0.0165	-0.0300	-0.0136
	Technical & Analytical	+0.0034	-0.0013	+0.0375	+0.0356	-0.0783	-0.1413	+0.0050	+0.0161	-0.0150	+0.0143
Qwen2	Clear & Direct	+0.0254	+0.0285	+0.0070	+0.0026	-0.0492	-0.0755	-0.0514	-0.0509	+0.0626	+0.0452
	Emotionally Descriptive	+0.0326	+0.0371	+0.0367	+0.0339	-0.1600	-0.2805	-0.0155	-0.0153	+0.0113	+0.0180
	Technical & Analytical	+0.0102	+0.0178	+0.0670	+0.0663	-0.0595	-0.0837	-0.0309	-0.0304	-0.0092	-0.0160
Gemini	Clear & Direct	+0.0274	+0.0260	+0.0460	+0.0458	-0.0252	0.0556	+0.0318	+0.0330	+0.0484	+0.0384
	Emotionally Descriptive	+0.0146	+0.0140	-0.0069	-0.0035	-0.0558	-0.0100	+0.0368	+0.0381	+0.1084	+0.1058
	Technical & Analytical	+0.0011	+0.0523	+0.0284	+0.0184	-0.0208	+0.0636	-0.0820	-0.0977	+0.0530	+0.0464
GPT4o	Clear & Direct	+0.0371	+0.0400	+0.0110	+0.0060	0.0000	-0.0052	+0.0029	+0.0029	+0.1050	+0.1601
	Emotionally Descriptive	-0.0279	-0.0224	+0.0080	+0.0046	-0.0800	-0.1270	+0.0129	+0.0130	+0.1321	+0.1840
	Technical & Analytical	+0.0179	+0.0592	+0.0093	+0.0007	0.0000	-0.0052	-0.0036	-0.0042	+0.0989	+0.1423

Table 3: Performance changes in Accuracy and F1-macro metrics based on Task Instruction compared to the Baseline prompt. Persona is not applied. Blue indicates a performance improvement, while red indicates a decline.

depression detection [4, 62], and suicidal risk detection [17], there has been a lack of systematic understanding of the components that constitute these prompts and their effects on performance. To address this gap, we identified the core components of prompts used and analyzed how each component influences performance. Our modular prompt design offers a systematic evaluation framework tailored for HCI researchers and software developers working on LLM-based mental health systems. The modular prompt design could be adopted to evaluate and refine the prompts used in their intervention systems, such as recognizing children's emotions [56], diagnosing stress or depression [26, 51], and detecting suicidal ideation risks [59]. This approach enables researchers to iteratively test and optimize prompt configurations, enhancing the precision of mental health detection and increasing the efficacy of intervention outcomes. A promising future direction involves automating this process by incorporating modular components into LLM-assisted prompt engineering [27, 50, 81].

6.2 Guidelines for Modular Prompt Design and Systematic Evaluation

Informed by our findings, we propose guidelines to design, refine, and evaluate LLM prompts for emotion and mental health tasks. These guidelines are intended to assist researchers and practitioners in creating high-performing, reproducible, and reusable prompts.

Step 1. Decompose existing prompts into six modules and check clarity

First, decompose existing prompts into key modules, which can then be used for systematic evaluation: Persona, Task Instruction, N-shot, Template, Input, and Output. Each module should be aligned

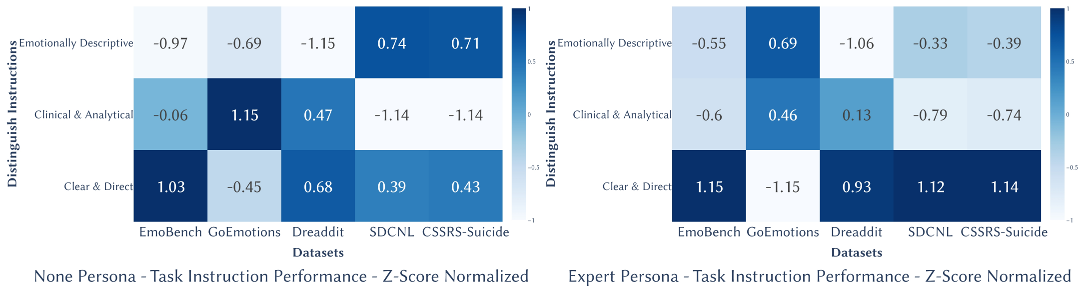

Figure 5: Analysis of the impact of Task Instructions on LLM performance, with and without the application of Persona. We averaged the F1-macro scores of each model across and then normalized the values using z-scores to visualize the relative performance differences. Positive values indicate above-average performance, Negative values indicate below-average performance.

Task Instruction	Model	Emobench		Goemotion		Dreaddit		SDCNL		CSSRS	
		Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
Persona-None											
Clear & Direct	**Mistral**	0.3604	0.3440	0.3485	0.3261	0.6050	0.5404	0.7236	0.7213	0.2700	0.2127
	Qwen2	0.4061	0.3916	0.3368	0.3284	0.6308	0.5813	0.6923	0.6919	0.3026	0.2193
	Gemini	0.5678	0.5415	**0.4219**	**0.4247**	0.5556	0.4522	0.7368	0.7367	0.3600	0.3004
	GPT4o	**0.6100**	0.5908	0.4192	0.4093	0.6650	0.6254	0.7100	0.7099	0.4000	0.3620
Emotionally Descriptive	**Mistral**	0.3250	0.3077	0.3401	0.3145	0.5692	0.4682	0.7050	0.7017	0.2400	0.1648
	Qwen2	0.4133	0.4002	0.3665	0.3597	0.5200	0.3763	0.7282	0.7275	0.2513	0.1921
	Gemini	0.5550	0.5295	0.3690	0.3754	0.5250	0.3866	**0.7418**	**0.7418**	0.4200	0.3678
	GPT4o	0.5450	0.5284	0.4162	0.4079	0.5850	0.5036	0.7200	0.7200	**0.4271**	**0.3859**
Technical & Analytical	**Mistral**	0.3401	0.3176	0.3571	0.3295	0.5750	0.4866	0.7085	0.7013	0.2550	0.1927
	Qwen2	0.3909	0.3809	0.3968	0.3921	0.6205	0.5731	0.7128	0.7124	0.2308	0.1581
	Gemini	0.5415	0.5678	0.4043	0.3973	0.5600	0.4602	0.6230	0.6060	0.3646	0.3084
	GPT4o	0.5908	**0.6100**	0.4175	0.4040	**0.6650**	**0.6254**	0.7035	0.7028	0.3939	0.3442
Persona-Expert											
Clear & Direct	**Mistral**	0.3452	0.3213	0.3469	0.3188	0.5700	0.4725	0.7250	0.7213	0.2550	0.1939
	Qwen2	0.4082	0.3915	0.2893	0.2734	0.6256	0.5627	0.7077	0.7072	0.2615	0.2058
	Gemini	0.5578	0.5297	0.3782	0.3813	0.5850	0.5036	0.6528	0.6403	**0.4150**	**0.3824**
	GPT4o	**0.6080**	**0.5953**	0.4154	0.4073	**0.6700**	**0.6349**	**0.7250**	**0.7238**	0.3650	0.3347
Emotionally Descriptive	**Mistral**	0.3434	0.3216	0.3401	0.3192	0.5897	0.5042	0.7150	0.7124	0.2550	0.1816
	Qwen2	0.3949	0.3828	0.3402	0.3704	0.5400	0.4165	0.6974	0.6954	0.2205	0.1546
	Gemini	0.5276	0.5063	0.3866	0.3955	0.5550	0.4451	0.6515	0.6307	0.3600	0.3307
	GPT4o	0.5500	0.5340	**0.4278**	**0.4172**	0.6600	0.6184	0.7236	0.7229	0.4050	0.3650
Technical & Analytical	**Mistral**	0.3316	0.3096	0.3214	0.2995	0.5500	0.4420	0.7136	0.7078	0.2600	0.1962
	Qwen2	0.3980	0.3853	0.3757	0.3665	0.6205	0.5546	0.7026	0.7025	0.2103	0.1479
	Gemini	0.5377	0.5092	0.4062	0.4003	0.5800	0.4900	0.6564	0.6231	0.3900	0.3387
	GPT4o	0.5707	0.5553	0.4227	0.4118	0.6550	0.6114	0.7200	0.7182	0.3737	0.3297

Table 4: Evaluation results of each model using two Persona and three Task Instruction combinations. The bold texts indicate the highest performance in terms of Accuracy and F1-score for each dataset.

with its intended purpose and adjusted as necessary. After decomposition, ensure each module is clear, concise, and easy to understand, revising ambiguous elements to improve comprehension.

Note that our findings indicate that modular prompts enable the creation of flexible and reusable prompts. Modular prompts allow flexible removal or addition of specific modules to achieve the desired outcome and the reuse of modules that have proven effective in certain mental health tasks for similar tasks.

Step 2. Identify and evaluate variation and interactions

Identify variations in each module and evaluate interactions between modules to understand their impact on performance. Note that *interactions* between modules could also be tested to analyze synergies or trade-offs.

Note that *variations* can be tested for each module, such as different tones in the 'Persona' module, including empathetic, neutral, or expert. Based on our findings, we recommend using the 'Persona-Expert' module in suicidal risk prediction tasks because this module leads to more accurate responses. For the 'Task Instruction' module, we recommend using 'Clear and Direct' instructions for simple tasks like binary classification because overly complicated instructions can hinder performance. In contrast, for tasks like suicidal risk detection, 'Emotionally Descriptive' instructions are recommended because they can mitigate performance degradation. Furthermore, our findings indicate that combining modules does not always lead to improved performance. For instance, combining 'Emotionally Descriptive' task instructions with the 'Persona-Expert' module resulted in performance degradation in suicidal risk prediction. Therefore, each module should be tested both individually and in combination to identify configurations that improve performance.

6.3 Ethical and Privacy Safeguards for Sensitive Mental Health Applications

Large language models show promise in emotion and mental health analysis, serving as complementary tools that can assist experts instead of replacing humans. However, their deployment requires careful oversight, particularly in addressing ethical and privacy concerns. One such ethical concern is a model bias where certain groups may be overrepresented or underrepresented in training data, leading to biased results [7, 39]. An ethical safeguard is the iterative design of bias-contributing modules by testing variations of each module and refining them through repeated iterations. For instance, if the persona-expert module introduces bias, the prompt could be adjusted to: "*You are an 'unbiased' expert, specializing in emotional classification that is not biased toward gender.*" Bias-aware design of each module allows for targeted improvements without overhauling the entire prompt. Additionally, privacy concerns about *leakage of sensitive personal information* must be carefully considered when an LLM diagnoses a user's mental health based on third-party data or past information [41, 78]. To safeguard privacy, the 'Input' and 'Output' modules should filter personally identifiable information and sensitive data, with local processing recommended to prevent leakage. This would help reduce the risk of privacy breaches and prevent the unintended exposure of confidential information.

6.4 Limitation and Future Work

Although we evaluated the major components of persona and task instruction, the limitation of not being able to test all combinations of prompt components still exists. To fully understand the impact of each prompt component on model performance, additional research that considers the interaction between various components is necessary. In particular, a detailed analysis of how each component, individually or in combination, affects performance is crucial, as such analysis could maximize the flexibility and reusability of prompt design. Furthermore, the analysis of the results from the Dreaddit and SDCNL datasets, as shown in Table 2, Table 4, Fig. 3, and Fig. 4, revealed that open-source models (smaller models) performed better than closed-source (large models) models. While smaller models generally tend to underperform compared to larger models [79], it is notable that the open-source models outperformed in these specific datasets. This may indicate the possibility that these datasets were included in the training data of the open-source models. If these datasets were indeed used for training, the evaluation results could be skewed, and the actual effect of certain prompt components may not be accurately reflected. Therefore, when interpreting the study's findings, it is essential to consider the potential for data leakage, and future research should further investigate this issue. By eliminating the possibility of dataset duplication and the resulting performance distortion, more reliable research outcomes can be achieved.

7 Conclusion

We proposed a modular prompt design approach for emotion and mental health tasks. Our findings underscore the value of modularity in prompt engineering for software developers, offering flexibility and reusability for optimizing prompts in emotionally sensitive contexts. As a case study, we explored a systematic evaluation of persona and task instruction variations. While persona and task instruction can enhance performance individually, contrary to our expectations, their combination did not always yield better results, emphasizing the need for task-specific prompt design. Our modular framework provides systematic ways of designing effective prompts, with future work focusing on expanding the analysis of prompt components and designing new tools for prompt design.

Acknowledgments

This research was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Korean government (MSIT) (2022R1A2C2011536).

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* (2023).

[2] Zeyad Alghamdi, Tharindu Kumarage, Garima Agrawal, Huan Liu, and H Russell Bernard. 2024. Less is More: Stress Detection through Condensed Social Media Contents. In *European Conference on Social Media*, Vol. 11. 13–22.

[3] Ian Arawjo, Chelse Swoopes, Priyan Vaithilingam, Martin Wattenberg, and Elena L. Glassman. 2024. ChainForge: A Visual Toolkit for Prompt Engineering and LLM Hypothesis Testing. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 304, 18 pages.

[4] Mihael Arcan, Paul-David Niland, and Fionn Delahunty. 2024. An assessment on comprehending mental health through large language models. *arXiv preprint arXiv:2401.04592* (2024).

[5] Ankita Bhaumik and Tomek Strzalkowski. 2024. Towards a Generative Approach for Emotion Detection and Reasoning. *arXiv preprint arXiv:2408.04906* (2024).

[6] Tom B Brown. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165* (2020).

[7] Karikarn Chansiri, Xinyu Wei, and Ka Ho Brian Chor. 2024. Addressing Gender Bias: A Fundamental Approach to AI in Mental Health. In *2024 5th International Conference on Big Data Analytics and Practices (IBDAP)*. IEEE, 107–112.

[8] Devendra Singh Chaplot. 2023. Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, William El Sayed. *arXiv preprint arXiv:2310.06825* (2023).

[9] Banghao Chen, Zhaofeng Zhang, Nicolas Langrené, and Shengxin Zhu. 2023. Unleashing the potential of prompt engineering in Large Language Models: a comprehensive review. *arXiv preprint arXiv:2310.14735* (2023).

[10] Jiyu Chen, Vincent Nguyen, Xiang Dai, Diego Molla, Cecile Paris, and Sarvnaz Karimi. 2024. Exploring Instructive Prompts for Large Language Models in the Extraction of Evidence for Supporting Assigned Suicidal Risk Levels. In *Proceedings of the 9th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2024)*. 197–202.

[11] Victoria Clarke and Virginia Braun. 2017. Thematic analysis. *The journal of positive psychology* 12, 3 (2017), 297–298.

[12] Google DeepMind. 2023. Gemini: Multimodal and Language Model. https://www.deepmind.com/research/gemini Accessed: 2023-09-05.

[13] Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. GoEmotions: A Dataset of Fine-Grained Emotions. arXiv:2005.00547 [cs.CL] https://arxiv.org/abs/2005.00547

[14] Claudia Diamantini, Alex Mircoli, Domenico Potena, Simone Vagnoni, et al. 2023. An Experimental Comparison of Large Language Models for Emotion Recognition in Italian Tweets.. In *itaDATA*.

[15] Angela Fan, Beliz Gokkaya, Mark Harman, Mitya Lyubarskiy, Shubho Sengupta, Shin Yoo, and Jie M. Zhang. 2023. Large Language Models for Software Engineering: Survey and Open Problems. In *2023 IEEE/ACM International Conference on Software Engineering: Future of Software Engineering (ICSE-FoSE)*. 31–53.

[16] Manas Gaur, Amanuel Alambo, Joy Prakash Sain, Ugur Kursuncu, Krishnaprasad Thirunarayan, Ramakanth Kavuluru, Amit Sheth, Randy Welton, and Jyotishman Pathak. 2019. Knowledge-aware assessment of severity of suicide risk for early intervention. In *The world wide web conference*. 514–525.

[17] Hamideh Ghanadian, Isar Nejadgholi, and Hussein Al Osman. 2024. Socially Aware Synthetic Data Generation for Suicidal Ideation Detection Using Large Language Models. *IEEE Access* (2024).

[18] Daniel Goleman. 1996. Emotional intelligence. Why it can matter more than IQ. *Learning* 24, 6 (1996), 49–50.

[19] Ziyin Gu and Qingmeng Zhu. 2023. MentalBlend: Enhancing Online Mental Health Support through the Integration of LLMs with Psychological Counseling Theories. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, Vol. 46.

[20] Kenta Hama, Atsushi Otsuka, and Ryo Ishii. 2024. Emotion Recognition in Conversation with Multi-step Prompting Using Large Language Model. In *International Conference on Human-Computer Interaction*. Springer, 338–346.

[21] Ayaan Haque, Viraaj Reddi, and Tyler Giallanza. 2021. Deep learning for suicide and depression identification with unsupervised label correction. In *Artificial Neural Networks and Machine Learning–ICANN 2021: 30th International Conference on Artificial Neural Networks, Bratislava, Slovakia, September 14–17, 2021, Proceedings, Part V 30*. Springer, 436–447.

[22] Tiancheng Hu and Nigel Collier. 2024. Quantifying the Persona Effect in LLM Simulations. arXiv:2402.10811 [cs.CL] https://arxiv.org/abs/2402.10811

[23] Mary Helen Immordino-Yang and Antonio R. Damasio. 2007. We feel, therefore we learn: The relevance of affective and social neuroscience to education. *Mind, Brain, and Education* 1, 1 (2007), 3–10.

[24] Mia Mohammad Imran, Preetha Chatterjee, and Kostadin Damevski. 2024. Uncovering the causes of emotions in software developer communication using zero-shot llms. In *Proceedings of the IEEE/ACM 46th International Conference on Software Engineering*. 1–13.

[25] Hyolim Jeon, Dongje Yoo, Daeun Lee, Sejung Son, Seungbae Kim, and Jinyoung Han. 2024. A Dual-Prompting for Interpretable Mental Health Language Models. *arXiv preprint arXiv:2402.14854* (2024).

[26] Taewan Kim, Seolyeong Bae, Hyun Ah Kim, Su-woo Lee, Hwajung Hong, Chanmo Yang, and Young-Ho Kim. 2024. MindfulDiary: Harnessing Large Language Model to Support Psychiatric Patients' Journaling. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–20.

[27] Tae Soo Kim, Yoonjoo Lee, Jamin Shin, Young-Ho Kim, and Juho Kim. 2024. EvalLM: Interactive Evaluation of Large Language Model Prompts on User-Defined Criteria. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 306, 21 pages.

[28] Xiaochong Lan, Yiming Cheng, Li Sheng, Chen Gao, and Yong Li. 2024. Depression Detection on Social Media with Large Language Models. *arXiv preprint arXiv:2403.10750* (2024).

[29] Cheng Li, Jindong Wang, Yixuan Zhang, Kaijie Zhu, Wenxin Hou, Jianxun Lian, Fang Luo, Qiang Yang, and Xing Xie. 2023. Large language models understand and can be enhanced by emotional stimuli. arXiv. *arXiv preprint arXiv:2307.11760* (2023).

[30] Junyi Li, Ninareh Mehrabi, Charith Peris, Palash Goyal, Kai-Wei Chang, Aram Galstyan, Richard Zemel, and Rahul Gupta. 2024. On the steerability of large language models toward data-driven personas. arXiv:2311.04978 [cs.CL] https://arxiv.org/abs/2311.04978

[31] Lei Li, Yongfeng Zhang, and Li Chen. 2023. Prompt distillation for efficient llm-based recommendation. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*. 1348–1357.

[32] Zaijing Li, Gongwei Chen, Rui Shao, Dongmei Jiang, and Liqiang Nie. 2024. Enhancing the emotional generation capability of large language models via emotional chain-of-thought. *arXiv preprint arXiv:2401.06836* (2024).

[33] Chenxiao Liu, Zheyong Xie, Sirui Zhao, Jin Zhou, Tong Xu, Minglei Li, and Enhong Chen. [n. d.]. Speak From Heart: An Emotion-Guided LLM-Based Multimodal Method for Emotional Dialogue Generation. In *Proceedings of the 2024 International Conference on Multimedia Retrieval*.

[34] Chenxiao Liu, Zheyong Xie, Sirui Zhao, Jin Zhou, Tong Xu, Minglei Li, and Enhong Chen. 2024. Speak From Heart: An Emotion-Guided LLM-Based Multimodal Method for Emotional Dialogue Generation. In *Proceedings of the 2024 International Conference on Multimedia Retrieval* (Phuket, Thailand) *(ICMR '24)*. Association for Computing Machinery, New York, NY, USA, 533–542. doi:10.1145/3652583.3658104

[35] Zhiwei Liu, Kailai Yang, Qianqian Xie, Tianlin Zhang, and Sophia Ananiadou. 2024. Emollms: A series of emotional large language models and annotation tools for comprehensive affective analysis. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 5487–5496.

[36] Sebastian Lubos, Thi Ngoc Trang Tran, Alexander Felfernig, Seda Polat Erdeniz, and Viet-Man Le. 2024. LLM-generated Explanations for Recommender Systems. In *Adjunct Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization*. 276–285.

[37] Meng Luo, Han Zhang, Shengqiong Wu, Bobo Li, Hong Han, and Hao Fei. 2024. NUS-Emo at SemEval-2024 Task 3: Instruction-Tuning LLM for Multimodal Emotion-Cause Analysis in Conversations. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*. 1599–1606.

[38] Ruotian Ma, Xiaolei Wang, Xin Zhou, Jian Li, Nan Du, Tao Gui, Qi Zhang, and Xuanjing Huang. 2024. Are Large Language Models Good Prompt Optimizers? *arXiv preprint arXiv:2402.02101* (2024).

[39] Zilin Ma, Yiyang Mei, Yinru Long, Zhaoyuan Su, and Krzysztof Z Gajos. 2024. Evaluating the Experience of LGBTQ+ People Using Large Language Model Based Chatbots for Mental Health Support. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–15.

[40] Usman Malik, Simon Bernard, Alexandre Pauchet, Clément Chatelain, Romain Picot-Clemente, and Jérôme Cortinovis. 2024. Pseudo-Labeling With Large Language Models for Multi-Label Emotion Classification of French Tweets. *IEEE Access* (2024).

[41] Niloofar Mireshghallah, Hyunwoo Kim, Xuhui Zhou, Yulia Tsvetkov, Maarten Sap, Reza Shokri, and Yejin Choi. 2023. Can llms keep a secret? testing privacy implications of language models via contextual integrity theory. *arXiv preprint arXiv:2310.17884* (2023).

[42] El Habib Nfaoui and Hanane Elfaik. 2024. Evaluating Arabic Emotion Recognition Task Using ChatGPT Models: A Comparative Analysis between Emotional Stimuli Prompt, Fine-Tuning, and In-Context Learning. *Journal of Theoretical and Applied Electronic Commerce Research* 19, 2 (2024), 1118–1141.

[43] Arkadiusz Nowacki, Wojciech Sitek, and Henryk Rybiński. 2024. LLMental: Classification of Mental Disorders with Large Language Models. In *International Symposium on Methodologies for Intelligent Systems*. Springer, 35–44.

[44] Julia Ohse, Bakir Hadžić, Parvez Mohammed, Nicolina Peperkorn, Michael Danner, Akihiro Yorita, Naoyuki Kubota, Matthias Rätsch, and Youssef Shiban. 2024. Zero-Shot Strike: Testing the generalisation capabilities of out-of-the-box LLM models for depression detection. *Computer Speech & Language* 88 (2024), 101663.

[45] Ephraim Okoro, Melvin C. Washington, and Otis Thomas. 2017. The Impact of Interpersonal Communication Skills on Organizational Effectiveness and Social Self-Efficacy: A Synthesis. *International Journal of Language and Linguistics* 4, 3 (2017), 29–31.

[46] OpenAI. 2023. GPT-4 Technical Report. https://openai.com/research/gpt-4 Accessed: 2023-09-05.

[47] Craig L. Pearce and Edwin A. Locke. 2023. *Principles of Organizational Behavior: The Handbook of Evidence-Based Management* (3rd ed.). John Wiley & Sons P&T.

[48] Liyizhe Peng, Zixing Zhang, Tao Pang, Jing Han, Huan Zhao, Hao Chen, and Björn W Schuller. 2024. Customising General Large Language Models for Specialised Emotion Recognition Tasks. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 11326–11330.

[49] YHPP Priyadarshana, Ashala Senanayake, Zilu Liang, and Ian Piumarta. 2024. Prompt engineering for digital mental health: a short review. *Frontiers in Digital Health* 6 (2024).
[50] Reid Pryzant, Dan Iter, Jerry Li, Yin Tat Lee, Chenguang Zhu, and Michael Zeng. 2023. Automatic prompt optimization with "gradient descent" and beam search. *arXiv preprint arXiv:2305.03495* (2023).
[51] Huachuan Qiu and Zhenzhong Lan. 2024. Interactive Agents: Simulating Counselor-Client Psychological Counseling via Role-Playing LLM-to-LLM Interactions. *arXiv preprint arXiv:2408.15787* (2024).
[52] Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530* (2024).
[53] Sahand Sabour, Siyang Liu, Zheyuan Zhang, June M. Liu, Jinfeng Zhou, Alvionna S. Sunaryo, Juanzi Li, Tatia M. C. Lee, Rada Mihalcea, and Minlie Huang. 2024. EmoBench: Evaluating the Emotional Intelligence of Large Language Models. arXiv:2402.12071 [cs.CL] https://arxiv.org/abs/2402.12071
[54] Misha Sadeghi, Bernhard Egger, Reza Agahi, Robert Richer, Klara Capito, Lydia Helene Rupp, Lena Schindler-Gmelch, Matthias Berking, and Bjoern M Eskofier. 2023. Exploring the capabilities of a language model-only approach for depression detection in text data. In *2023 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI)*. IEEE, 1–5.
[55] Shubhra Kanti Karmaker Santu and Dongji Feng. 2023. TELeR: A General Taxonomy of LLM Prompts for Benchmarking Complex Tasks. arXiv:2305.11430 [cs.AI] https://arxiv.org/abs/2305.11430
[56] Woosuk Seo, Chanmo Yang, and Young-Ho Kim. 2024. ChaCha: Leveraging Large Language Models to Prompt Children to Share Their Emotions about Personal Events. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–20.
[57] Loitongbam Gyanendro Singh, Junyu Mao, Rudra Mutalik, and Stuart E Middleton. 2024. Extraction and Summarization of Suicidal Ideation Evidence in Social Media Content Using Large Language Models. In *Proceedings of the Ninth Workshop on Computational Linguistics and Clinical Psychology, Association for Computational Linguistics*.
[58] Denise Solomon and Jennifer Theiss. 2022. *Interpersonal Communication: Putting Theory into Practice* (2nd ed.). Routledge.
[59] Logan Stapleton, Sunniva Liu, Cindy Liu, Irene Hong, Stevie Chancellor, Robert E Kraut, and Haiyi Zhu. 2024. " If This Person is Suicidal, What Do I Do?": Designing Computational Approaches to Help Online Volunteers Respond to Suicidality. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–21.
[60] William Stern, Seng Jhing Goh, Nasheen Nur, Patrick J Aragon, Thomas Mercer, Siddhartha Bhattacharyya, Chiradeep Sen, and Van Minh Nguyen. 2024. Natural Language Explanations for Suicide Risk Classification Using Large Language Models.. In *ML4CMH@ AAAI*. 74–83.
[61] Howard E. Sypher, Beverly Davenport Sypher, and John W. Haas. 1988. Getting Emotional: The Role of Affect in Interpersonal Communication. *American Behavioral Scientist* 31, 3 (1988), 327–340.
[62] Yongfeng Tao, Minqiang Yang, Hao Shen, Zhichao Yang, Ziru Weng, and Bin Hu. 2023. Classifying anxiety and depression through LLMs virtual interactions: A case study with ChatGPT. In *2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2259–2264.
[63] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and Efficient Foundation Language Models. arXiv:2302.13971 [cs.CL] https://arxiv.org/abs/2302.13971
[64] Elsbeth Turcan and Kathleen McKeown. 2019. Dreaddit: A Reddit Dataset for Stress Analysis in Social Media. *ArXiv* abs/1911.00133 (2019). https://api.semanticscholar.org/CorpusID:207870937
[65] Noor Ul Huda, Sanam Fayaz Sahito, Abdul Rehman Gilal, Ahsanullah Abro, Abdullah Alshanqiti, Aeshah Alsughayyir, and Abdul Sattar Palli. 2024. Impact of Contradicting Subtle Emotion Cues on Large Language Models with Various Prompting Techniques. *International Journal of Advanced Computer Science & Applications* 15, 4 (2024).
[66] Ahmet Yavuz Uluslu, Andrianos Michail, and Simon Clematide. 2024. Utilizing large language models to identify evidence of suicidality risk through analysis of emotionally charged posts. In *Proceedings of the 9th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2024)*. 264–269.
[67] Naoki Wake, Atsushi Kanehira, Kazuhiro Sasabuchi, Jun Takamatsu, and Katsushi Ikeuchi. 2023. Bias in Emotion Recognition with ChatGPT. *arXiv preprint arXiv:2310.11753* (2023).
[68] Jiashuo Wang, Yang Xiao, Yanran Li, Changhe Song, Chunpu Xu, Chenhao Tan, and Wenjie Li. 2024. Towards a Client-Centered Assessment of LLM Therapists by Client Simulation. *arXiv preprint arXiv:2406.12266* (2024).
[69] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems* 35 (2022), 24824–24837.
[70] Jules White, Quchen Fu, Sam Hays, Michael Sandborn, Carlos Olea, Henry Gilbert, Ashraf Elnashar, Jesse Spencer-Smith, and Douglas C Schmidt. 2023. A prompt pattern catalog to enhance prompt engineering with chatgpt. *arXiv preprint arXiv:2302.11382* (2023).
[71] Xuhai Xu, Bingsheng Yao, Yuanzhe Dong, Saadia Gabriel, Hong Yu, James Hendler, Marzyeh Ghassemi, Anind K Dey, and Dakuo Wang. 2024. Mental-llm: Leveraging large language models for mental health prediction via online text data. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 8, 1 (2024), 1–32.
[72] An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671* (2024).
[73] Kailai Yang, Shaoxiong Ji, Tianlin Zhang, Qianqian Xie, Ziyan Kuang, and Sophia Ananiadou. 2023. Towards interpretable mental health analysis with large language models. *arXiv preprint arXiv:2304.03347* (2023).
[74] Kailai Yang, Tianlin Zhang, Ziyan Kuang, Qianqian Xie, Jimin Huang, and Sophia Ananiadou. 2024. MentaLLaMA: interpretable mental health analysis on social media with large language models. In *Proceedings of the ACM on Web Conference 2024*. 4489–4500.
[75] Zhou Yang, Zhaochun Ren, Chenglong Ye, Yufeng Wang, Haizhou Sun, Chao Chen, Xiaofei Zhu, Yunbing Wu, and Xiangwen Liao. 2024. E-ICL: Enhancing Fine-Grained Emotion Recognition through the Lens of Prototype Theory. *arXiv preprint arXiv:2406.02642* (2024).
[76] Fan Yin, Jesse Vig, Philippe Laban, Shafiq Joty, Caiming Xiong, and Chien-Sheng Jason Wu. 2023. Did you read the instructions? rethinking the effectiveness of task definitions in instruction learning. *arXiv preprint arXiv:2306.01150* (2023).
[77] Tianlin Zhang, Kailai Yang, Shaoxiong Ji, Boyang Liu, Qianqian Xie, and Sophia Ananiadou. 2024. SuicidEmoji: Derived Emoji Dataset and Tasks for Suicide-Related Social Content. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1136–1141.
[78] Zhiping Zhang, Michelle Jia, Hao-Ping Lee, Bingsheng Yao, Sauvik Das, Ada Lerner, Dakuo Wang, and Tianshi Li. 2024. "It'sa Fair Game", or Is It? Examining How Users Navigate Disclosure Risks and Benefits When Using LLM-Based Conversational Agents. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery New York, NY, USA, 1–26.
[79] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223* (2023).
[80] Xuanyan Zhong, Haiyang Xin, Wenfeng Li, Zehui Zhan, and May-hung Cheng. 2024. The Design and application of RAG-based conversational agents for collaborative problem solving. In *Proceedings of the 2024 9th International Conference on Distance Education and Learning*. 62–68.
[81] Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. 2022. Large language models are human-level prompt engineers. *arXiv preprint arXiv:2211.01910* (2022).
[82] Jingwei Zhu, Ancheng Xu, Minghuan Tan, and Min Yang. 2024. XinHai@ CLPsych 2024 Shared Task: Prompting Healthcare-oriented LLMs for Evidence Highlighting in Posts with Suicide Risk. In *Proceedings of the 9th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2024)*. 238–246.

A Appendix

Topic	Associated studies
Emotion recognition	[5] [42] [32] [67] [14] [24] [20] [37] [35] [29] [48] [40] [33] [19]
Anxiety and depression/stress detection	[62] [4] [28] [43] [54] [73] [74] [71]
Suicide risk detection	[17] [25] [66] [10] [57] [77] [60] [82]

Table 5: Studies from which prompts were extracted for thematic analysis. The studies are categorized by their focus: emotion recognition, anxiety and depression/stress detection, and suicide risk detection.

Component	Code Label	Freq.	Definition	Examples
Persona	Role	**22**	Instructs the AI to adopt a specific role or behave in a particular way. This can be used to adjust the tone, style, or depth of the information generated.	"*You are a psychiatrist.*"[25] "*You're an expert in sentiment analysis and emotion cause identification*"[37]
	Capability	**4**	Describes the skills, knowledge, and abilities that the persona is expected to possess, indicating what the AI should be able to perform or understand.	"*You can accurately assess people's emotional states*"[32] "*capable of understanding the sentiment within a text.*"[67]
Task	Contextual information	**15**	Specifies the nature or origin of the input data (e.g., social media posts, diary entries, or transcripts), providing necessary context for the task.	"*This person wrote this paragraph on social media.*"[71] "*You will be provided with a tweet written in Arabic variants (Modern Standard Arabic and Dialectal Arabic)*" [42]
	Task knowledge	**11**	Provides the model with domain-specific knowledge or background information that it can utilize to carry out the analysis.	"*Generalised anxiety disorder is a mental health illness that is defined by people having feelings of excessive anxiety.*" [4]
	Task instruction	**54**	The primary query or set of instructions guiding the AI on how to perform the task or address the problem at hand.	"*Consider the emotions expressed from this post to answer the question: Is the poster likely to suffer from very severe [Condition]?*" [73] "*Your task is to generate a suicidal text for each of the following "topics" with different Risk levels*" [17]
	Step-by-step thinking	**10**	Breaks down tasks into logical, sequential steps, enabling the model to address complex tasks systematically and methodically.	"*Let's think about it step by step:* *Step 1: Describe the content of the news.* *Step 2: Think about emotional reactions...* *Step 3: Think about how you need to express...*"[32]
	Emphasis	**3**	Emphasis element or stimuli is used to emphasize the importance of the task.	"*This is very important to my career.*" [29] "*You'd better be sure.*" [29]
N-shot Example		**7**	Provides examples to demonstrate how the model should handle similar tasks, helping the AI generalize from the provided instances.	"*Example 1:* *Post: Does everyone else just hurt all the time It's not like physical pain or soreness, it's just this overwhelming feeling of exhaustion...* *Response: Yes. Reasoning: The post conveys a deep sense of emotional pain, exhaustion, and numbness...*"[74]
Input		**54**	Actual data or content submitted for the task, which could include sources like social media posts, diary entries, or conversational threads relevant to the analysis.	"*Tweet: @CScheiwiller can't stop smiling*"[35] "*Post: Does everyone else just hurt all the time It's not like physical pain or soreness, it's just this overwhelming feeling of exhaustion...*"[74]
Output	Content requirement	**4**	Defines the essential information that must or must not be included in the output, ensuring that the model addresses all necessary elements of the task.	"*The response should not imply negative emotions toward anyone or anything, such as disgust, resentment, discrimination, hatred, etc.*" [32] "*Just give me the final word, no further analysis.*" [62]
	Format requirement	**23**	Specifies the format or structure that the output must follow to ensure consistency, clarity, and relevance in the model's response.	"*Provide the answers in JSON format with the following columns: text, topic, risk level.*" [17] "*Formatting: Strictly provide each snippet and only the snippets delimited by a semicolon(';')*" [66]
	Label list	**10**	A predefined set of labels or categories that the AI can select from when generating outputs, ensuring standardized classification or tagging.	"*Only from this emotion list: [Emotion List]. Only return the assigned word.*" [67] "*Only return Yes or No,*" [73]
Template		**5**	A predefined framework used to structure the prompt, dividing it into sections or headings to ensure the model receives well-organized and clear instructions.	"*[System] ... [Context] ... [Prompt] ... [Response] ... [Criteria] ....*" [32]

Table 6: Code book contains main components, code labels, frequency of code labels, definitions, and illustrative examples derived from thematic analysis.

Prompt examples	Associated studies
You will be presented with a post. Consider the emotions expressed in this post to identify whether the poster suffers from [condition]. Only return Yes or No, then explain your reasoning step by step. Here are N examples: Post: [example 1] Response: [response 1] ... Post: [example N] Response: [response N] Post: [Post] Response:	[5]
Analyze the dialogue to determine whether the respondent's emotional state is depression or anxiety. Question: [question], Answer: [text],..., Question: [question], Answer: [text], tell me the respondent's emotion in the following format: "anxiety" or "depression". Just give me the final word, no further analysis.	[62]
Your task is to generate a suicidal text for each of the following "topics" with different Risk levels. 1 - Depression 2 - Anxiety 3-Hopelessness 4-Anger 5-Perfectionism 6-Family issues 7-Relationship problems 8-Unemployment 9-FinancialCrisis 10-Education 11-Being Bullied 12-Death of close one 13-Immigration 14-Racism Provide the answers in JSON format with the following columns: text, topic, risk level. Risk level criteria: These are the criteria of different suicide risk level: Risk Level=Non Suicidal: I do not see evidence that this person is at risk for suicide Risk Level = Suicidal: I believe this person is at high risk of attempting suicide in the near future.	[17]
Determine whether each item in the following list of emotions is conveyed in the text below, which is delimited with triple backticks. Give your answer as a list with labels and 0 or 1 for each label. List of emotions: Anger, Anticipation, Disgust, Fear, Joy, Love, Optimism, Pessimism, Sadness, Surprise, Trust, neutral Text : I am filled with jealous rage, I am feeling quite sad, sorry for myself but I will snap out of it soon.	[42]
Task: Categorize the tweet into an ordinal class that best characterizes the tweeter's mental state, considering various degrees of positive and negative sentiment intensity. 3: very positive mental state can be inferred. 2: moderately positive mental state can be inferred. 1: slightly positive mental state can be inferred. 0: neutral or mixed mental state can be inferred. -1: slightly negative mental state can be inferred. -2: moderately negative mental state can be inferred. -3: very negative mental state can be inferred Tweet: Beyoncé resentment gets me in my feelings every time. Intensity Class:	[35]

Table 7: Examples of extracted prompts from existing studies, illustrating the variety of prompt-based approaches used in mental health and emotion analysis tasks. Each prompt is designed to elicit model responses for specific tasks such as emotion classification, risk assessment, and sentiment intensity categorization.

Component	Sub-Components	Emotion Recognition		Mental Health Analysis		
		Emobench	**GoEmotion**	**Dreaddit**	**SDCNL**	**CSSRS-Suicide**
Persona	Role	You are an expert system specializing in emotion classification, designed to analyze text with a highly analytical and empathetic approach.		You are an expert system specializing in mental health analysis, designed to evaluate text with a highly sensitive and empathetic approach.		
	Capability	You excel at detecting and interpreting a wide range of emotions, considering nuanced language and complex emotional cues.		Your expertise lies in identifying signs of mental health concerns, including anxiety, depression, and stress, by carefully analyzing nuanced language and subtle emotional cues.		
Task	Task Instruction: 1. Clear & Direct	Review the scenario, note the emotions the subject is feeling, and choose the right answer to the question.	Read the Reddit post, identify the emotions expressed, and choose the emotion label that best matches the overall sentiment.	Read the post, focus on the writer's mental state and emotions, and answer the question with a clear "yes" or "no".	Read the post, focus on the writer's mental state and emotions, and decide if they are "suicidal" or experiencing "depression."	Read the post, focus on the writer's mental state and emotions, and choose the suicide severity scale that best matches their condition.
	Task Instruction: 2. Emotionally Descriptive	Immerse yourself in the scenario, attentively observing the waves of emotion the subject is experiencing. Let the depth of these feelings guide you as you select the answer that truly resonates with the emotional core of the situation.	Delve into the Reddit post, paying close attention to the emotional undertones and expressive language. Feel the intensity of the emotions conveyed, and select the emotion label that most accurately captures the heart of the sentiment.	Immerse yourself in the post, deeply sensing the writer's emotional state, their mental turmoil, and the underlying thoughts that guide their feelings. Let this emotional insight inform your response, answering the question with a definitive "yes" or "no".	Carefully examine the post, tuning into the writer's emotional depth, mental struggles, and the underlying despair in their thoughts. Use this emotional insight to determine whether the writer is "suicidal" or suffering from "depression."	Immerse yourself in the post, paying close attention to the writer's emotional turmoil, mental state, and the underlying thoughts that reveal their struggles. Let this emotional understanding guide you in selecting the suicide severity scale that most accurately reflects their mental condition.
	Task Instruction: 3. Technical & Analytical	Conduct a thorough analysis of the scenario, with a particular focus on the subject's affective states and emotional responses. Apply your understanding of psychological principles to identify the most accurate answer, ensuring that your choice reflects a nuanced interpretation of the subject's emotional and cognitive processes.	Analyze the Reddit post with a focus on identifying and categorizing the emotional expressions. Utilize psychological frameworks to determine the most appropriate emotion label that encapsulates the overarching sentiment of the post, considering both explicit and nuanced emotional cues.	Conduct a thorough assessment of the post, analyzing the writer's mental state, emotional expressions, and cognitive processes. Using clinical reasoning and psychological insight, determine the most appropriate answer to the question, responding with a precise "yes" or "no".	Perform a detailed analysis of the post, evaluating the writer's mental state, emotional expressions, and cognitive patterns. Utilize your psychological expertise to accurately diagnose whether the writer's condition is indicative of "depression" or "suicidal" ideation, and provide your answer accordingly.	Conduct a comprehensive assessment of the post, focusing on the writer's mental state, affective expressions, and cognitive processes. Utilize established psychological frameworks to determine the most appropriate suicide severity scale, ensuring it accurately reflects the writer's current mental condition and risk level.
N-shot Examples		0-shot				
Input		Input : "input content for each dataset sample"				
Output		[Requirements] Provide your response in text. Only select the Label from "{label_list}". Do not generate labels that are not in the list. Your response must include 'Label: ' followed by the selected label and 'Confidence Score: ' followed by a score from 0 to 1 indicating your confidence. Respond according to the [Format]. [Format] <Label>: [Your Selected Label Here] <Confidence Score>: [Your Confidence Score Here] Provide your response in text.				
Template		###Persona### ###Task### ###N-shot Examples### ###Input### ###Output###				

Table 8: Prompt components used for evaluation. We fix the content for N-shot examples, Input, Output, and Prompt Template components while systematically evaluating different variations in Persona and Task Instructions. We select the prompt component for the corresponding dataset.

Prompt	Persona	Task	N-shot Examples	Input	Output	Prompt Template
Baseline	X	X	Fixed (zeroshot)	Fixed	Fixed	Fixed
Baseline + Persona	O	X	Fixed (zeroshot)	Fixed	Fixed	Fixed
Baseline + Task Instruction	X	O (3 variations)	Fixed (zeroshot)	Fixed	Fixed	Fixed
Baseline + Persona + Task Instruction	O	O (3 variations)	Fixed (zeroshot)	Fixed	Fixed	Fixed

Table 9: Incremental prompt variations for systematic evaluation, starting from the baseline prompt and progressively incorporating persona and task instructions.

```
###Persona###
You are an expert system specializing in mental health analysis, designed to evaluate text with a highly sensitive and empathetic approach. Your expertise lies in identifying signs of mental health concerns, including anxiety, depression, and stress, by carefully analyzing nuanced language and subtle emotional cues.

###Task###
Carefully examine the post, tuning into the writer's emotional depth, mental struggles, and the underlying despair in their thoughts. Use this emotional insight to determine whether the writer is "suicidal" or suffering from "depression."

###N-shot Examples###

###Input###
Scenario: Momo loved dancing and used it as an escape from her stressful life. Today, her client refused her work multiple times which resulted in her being reprimanded by her boss. Arriving at the dance studio, she put her bag down and screamed before starting to dance. After she finished dancing, Momo closed her eyes and nodded to herself in the mirror.
Question: What emotion would "Momo" ultimately feel in this situation?
Choices: "['Relief', 'Sadness', 'Anger', 'Hopeless', 'Amusement', 'Guilt']"

###Output###
[Requirements]
Provide your response in text. Only select the label from "['Jealousy', 'Embarrassment', 'Nervousness', 'Delight', 'Disapproval', 'Guilt']". Do not generate labels that are not in the list. Your response must include 'Label: ' followed by the selected label and 'Confidence Score: ' followed by a score from 0 to 1 indicating your confidence. Only generate Label and Confidence Score. Respond according to the [Format].

[Format]
<Label>: [Your Selected Label Here]
<Confidence Score>: [Your Confidence Score Here]
```

Emobench → Baseline + Persona + Task Instruction(Clear & Direct)

```
###Persona###
You are an expert system specializing in mental health analysis, designed to evaluate text with a highly sensitive and empathetic approach. Your expertise lies in identifying signs of mental health concerns, including anxiety, depression, and stress, by carefully analyzing nuanced language and subtle emotional cues.

###Task###
Immerse yourself in the post, deeply sensing the writer's emotional state, their mental turmoil, and the underlying thoughts that guide their feelings. Let this emotional insight inform your response, answering the question with a definitive "yes" or "no".

###N-shot Examples###

###Input###
Post: Post: If you knew Garret you would never have seen this coming. The family is exhausted at this point. I just want to do anything I can to help them and this is the platform I have. Any little thing helps. I can't think of a more caring family that is totally deserving of any help they can get.
nQuestion: "Does the poster suffers from stress?"
Choices : "['yes', 'no']"

###Output###
[Requirements]
Provide your response in text. Only select the Label from "['yes', 'no']". Do not generate Label that are not in the list. Your response must include 'Label: ' followed by the selected label and 'Confidence Score: ' followed by a score from 0 to 1 indicating your confidence. Only generate Label and Confidence Score. Respond according to the [Format].

[Format]
<Label>: [Your Selected Label Here]
<Confidence Score>: [Your Confidence Score Here]
```

Dreaddit → Baseline + Persona + Task Instruction(Emotionally Descriptive)

```
###Persona###
You are an expert system specializing in mental health analysis, designed to evaluate text with a highly sensitive and empathetic approach. Your expertise lies in identifying signs of mental health concerns, including anxiety, depression, and stress, by carefully analyzing nuanced language and subtle emotional cues.

###Task###
Conduct a comprehensive assessment of the post, focusing on the writer's mental state, affective expressions, and cognitive processes. Utilize established psychological frameworks to determine the most appropriate suicide severity scale, ensuring it accurately reflects the writer's current mental condition and risk level.

###N-shot Examples###

###Input###
Post: Please, dont do it. Im sure that there is someone out there who can help you. You may not believe it, but Im sure many more people then you think will be upset if you were to commit suicide. Please dont.
Choices : "['supportive', 'indicator', 'ideation', 'behavior', 'attempt']"

###Output###
[Requirements]
Provide your response in text. Only select the Label from "['supportive', 'indicator', 'ideation', 'behavior', 'attempt']". Do not generate Label that are not in the list. Your response must include 'Label: ' followed by the selected label and 'Confidence Score: ' followed by a score from 0 to 1 indicating your confidence. Only generate Label and Confidence Score. Respond according to the [Format].

[Format]
<Label>: [Your Selected Label Here]
<Confidence Score>: [Your Confidence Score Here]
```

CSSRS-Suicide → Baseline + Persona + Task Instruction(Technical & Analytical)

Figure 6: Examples of constructed prompts

Emotion Datasets
• **EmoBench** [53], based on emotional intelligence theories [18], consists of 400 multi-label scenarios, split into 200 scenarios for each category of emotion understanding and emotion application. Emotion understanding tests a model's ability to recognize and reason about emotion, while application evaluates how well it navigates emotionally complex situations. For our research, we focused on the emotion understanding category to evaluate models for complex emotion classification. • **GoEmotions** [13] is a large-scale Reddit-based dataset with 27 fine-grained emotion categories as well as the neutral category, allowing for more detailed classification compared to traditional datasets.
Mental Health Datasets
• **Dreaddit** [64] (stress) contains Reddit posts from five domains (abuse, social, anxiety, PTSD, and financial), labeled as stressful or not, and is considered relatively easy due to its binary classification. • **SDCNL** dataset covers suicidal ideation [21]. It is a Reddit-based dataset in which posts were labeled as either 'suicidal' or 'depression.' The task is moderately difficult due to the nuanced differences between suicide and depression. • **CSSRS-Suicide** dataset [16] includes posts from 15 mental health-related subreddits, annotated by psychiatrists based on five suicide risk indicators from the Columbia-Suicide Severity Rating Scale [16]. This dataset evaluates a model's ability to classify varying suicide risk levels.

Table 10: Summary of emotion and mental health datasets

Model	Release Date	Parameters Size	Open-Source	License
gpt-4o-2024-05-1	May-2024	-	X	Proprietary
Gemini-1.5-Pro-001	Feb-2024	-	X	Proprietary
Qwen2-7B-Instruct	Jun-2024	7B	O	Apache-2.0
Mistral-7B-Instruct-v0.3	May-2024	7B	O	Apache-2.0

Table 11: Model specifications including parameters, release dates, open-source availability, and license.

Task Instruction	Model	Emobench		Goemotion		Dreaddit		SDCNL		CSSRS	
		TPR	TNR	TPR	TNR	TPR	TNR	TPR	TNR	TPR	TNR
Persona-None											
None	**Mistral**	0.3426	0.8664	0.3133	0.9747	0.6520	0.6520	0.7047	0.7047	0.2700	0.8175
	Qwen2	0.3849	0.8762	0.3192	0.9751	0.6800	0.6800	0.7434	0.7434	0.2400	0.8100
	Gemini	0.5360	0.9084	0.3759	0.9768	0.5808	0.5808	0.7050	0.7050	0.3101	0.8277
	GPT4o	0.5797	0.9157	0.4015	0.9780	0.6650	0.6650	0.7072	0.7072	0.2950	0.8237
Clear & Direct	**Mistral**	0.3665	0.8733	0.3457	0.9759	0.6050	0.6050	0.7232	0.7232	0.2700	0.8175
	Qwen2	0.4161	0.8811	0.3296	0.9754	0.6290	0.6290	0.6921	0.6921	0.3026	0.8256
	Gemini	0.5618	0.9138	0.4065	**0.9786**	0.5556	0.5556	0.7378	0.7378	0.3600	0.8400
	GPT4o	**0.6184**	**0.9226**	**0.4154**	0.9785	**0.6650**	**0.6650**	0.7100	0.7100	0.4000	0.8500
Emotionally Descriptive	**Mistral**	0.3260	0.8660	0.3372	0.9755	0.5200	0.5200	0.7050	0.7050	0.2400	0.8100
	Qwen2	0.4290	0.8827	0.3535	0.9765	0.5670	0.5670	0.7285	0.7285	0.2513	0.8128
	Gemini	0.5530	0.9120	0.3561	0.9766	0.5250	0.5250	**0.7443**	**0.7443**	0.4200	0.8550
	GPT4o	0.5586	0.9104	0.4108	0.9784	0.5850	0.5850	0.7200	0.7200	**0.4262**	**0.8567**
Technical & Analytical	**Mistral**	0.3336	0.8705	0.3506	0.9762	0.5750	0.5750	0.7078	0.7078	0.2550	0.8137
	Qwen2	0.4068	0.8780	0.3826	0.9776	0.6188	0.6188	0.7130	0.7130	0.2308	0.8077
	Gemini	0.5670	0.9156	0.3884	0.9779	0.5600	0.5600	0.6127	0.6127	0.3646	0.8410
	GPT4o	0.5952	0.9183	0.4080	0.9784	0.6650	0.6650	0.7038	0.7038	0.3971	0.8489
Persona-Expert											
None	**Mistral**	0.3421	0.8716	0.3590	0.9764	0.5700	0.5700	0.7184	0.7184	0.2800	0.8200
	Qwen2	0.4023	0.8766	0.3002	0.9743	0.6150	0.6150	**0.7626**	**0.7626**	0.3050	0.8262
	Gemini	0.5413	0.9099	0.4123	0.9782	0.5850	0.5850	0.6750	0.6750	0.3000	0.8250
	GPT4o	0.5600	0.9114	0.4084	0.9784	**0.6750**	**0.6750**	0.7188	0.7188	0.3367	0.8345
Clear & Direct	**Mistral**	0.3331	0.8702	0.3433	0.9758	0.5700	0.5700	0.7250	0.7250	0.2550	0.8138
	Qwen2	0.4201	0.8823	0.2831	0.9736	0.6237	0.6237	0.7079	0.7079	0.2615	0.8154
	Gemini	0.5517	0.9122	0.3760	0.9770	0.5850	0.5850	0.6469	0.6469	**0.4150**	**0.8538**
	GPT4o	**0.6277**	**0.9221**	0.4069	0.9783	0.6700	0.6700	0.7250	0.7250	0.3650	0.8412
Emotionally Descriptive	**Mistral**	0.3346	0.8702	0.3365	0.9755	0.5400	0.5400	0.7150	0.7150	0.2550	0.8137
	Qwen2	0.4114	0.8787	0.3359	0.9755	0.5876	0.5876	0.6979	0.6979	0.2205	0.8051
	Gemini	0.5322	0.9069	0.3829	0.9773	0.5550	0.5550	0.6492	0.6492	0.3600	0.8400
	GPT4o	0.5668	0.9111	**0.4196**	**0.9788**	0.6600	0.6600	0.7234	0.7234	0.4050	0.8512
Technical & Analytical	**Mistral**	0.3179	0.8680	0.3158	0.9748	0.5500	0.5500	0.7129	0.7129	0.2600	0.8150
	Qwen2	0.4149	0.8799	0.3633	0.9768	0.6186	0.6186	0.7027	0.7027	0.2103	0.8026
	Gemini	0.5309	0.9088	0.4031	0.9780	0.5800	0.5800	0.6492	0.6492	0.3900	0.8475
	GPT4o	0.5855	0.9144	0.4165	0.9786	0.6550	0.6550	0.7200	0.7200	0.3718	0.8431

Table 12: Evaluation results of each model using two Persona and three Task Instruction combinations, including cases where Task Instruction is set to None. The bold texts indicate the highest performance in terms of True Positive Rate (TPR) and True Negative Rate (TNR) for each dataset.