Delft University of Technology

Stress detection through prompt engineering with a general-purpose LLM

Esmi, Nima; Shahbahrami, Asadollah; Nabati, Yasaman; Rezaei, Bita; Gaydadjiev, Georgi; de Jonge, Peter

DOI
10.1016/j.actpsy.2025.105462

Publication date
2025

Document Version
Final published version

Published in
Acta Psychologica

Citation (APA)
Esmi, N., Shahbahrami, A., Nabati, Y., Rezaei, B., Gaydadjiev, G., & de Jonge, P. (2025). Stress detection through prompt engineering with a general-purpose LLM. *Acta Psychologica*, *260*, Article 105462. https://doi.org/10.1016/j.actpsy.2025.105462

Acta Psychologica 260 (2025) 105462

Contents lists available at ScienceDirect

Acta Psychologica

journal homepage: www.elsevier.com/locate/actpsy

Stress detection through prompt engineering with a general-purpose LLM

Nima Esmi [a,b,*], Asadollah Shahbahrami [c,b], Yasaman Nabati [c], Bita Rezaei [c], Georgi Gaydadjiev [d], Peter de Jonge [e]

[a] Bernoulli Institute, University of Groningen, Groningen, The Netherlands
[b] ISRC, Khazar University, Baku, Azerbaijan
[c] University of Guilan, Guilan, Iran
[d] Delft University of Technology, Delft, The Netherlands
[e] University of Groningen, Groningen, The Netherlands

ARTICLE INFO

Keywords:
Large language model
Prompt engineering
Stress recognition
Social media
GPT-4

ABSTRACT

Advancements in large language models (LLMs) have opened new avenues for mental health monitoring through social media analysis. In this study, we present an iterative prompt engineering framework that significantly enhances the performance of the general-purpose LLM, GPT-4, for stress detection in social media posts, leveraging psychologist-informed hints. This approach achieved a substantial 17% accuracy improvement from 72% to 89% for the January 2025 version of GPT-4, alongside an 80% reduction in false positives compared to baseline zero-shot prompting. Our method not only surpassed domain-specific models like Mental-RoBERTa by 5% but also uniquely generates human-readable rationales. These rationales are crucial for mental health professionals, assisting them in understanding and validating the model's outputs—a key benefit for sensitive mental health applications. These results highlight prompt engineering as a resource-efficient, transparent strategy to adapt general-purpose LLMs for specialized tasks, offering a scalable solution for mental health monitoring without the need for costly fine-tuning.

1. Introduction

Large Language Models (LLMs) have revolutionized applications in healthcare, education, and industry by processing vast, diverse datasets with billions to trillions of parameters (Beghetto, Ross, Karwowski, & Glăveanu, 2025; Caruccio et al., 2024; Zahid et al., 2024; Zhang et al., 2025). In mental health, LLMs enable early detection of stress or depression through analysis of social media posts, offering scalable insights into emotional states (Chang, Shi et al., 2024). General-purpose LLMs, for instance GPT-4 with its trillion-parameter architecture, provide strong capabilities in generalization and explainability compared to domain-specific models. These include, specifically, Mental-RoBERTa, which are fine-tuned on targeted datasets derived from Reddit's mental health communities (Chang, Wang et al., 2024; Chebbi, Kniesel, Abdennadher, & Dimarzo, 2024; Devlin, Chang, Lee, & Toutanova, 2019; Ji et al., 2022; Raffel et al., 2020; Yang et al., 2023). However, domain-specific models often outperform general-purpose ones in specialized tasks due to their tailored training, yet they require resource-intensive, expert-labeled datasets (Gandy, Ivanitskaya, Bacon, & Bizri-Baryak, 2025). Fine-tuning large models like GPT-4 is equally challenging, constrained by computational costs and closed-source architectures, limiting customization for sensitive domains (Yang, Tao et al., 2024). Explainability is crucial in mental health, where stakeholders, including psychologists and patients, demand transparent reasoning to trust model outputs (Tufano, Dabić, Mastropaolo, Ciniselli, & Bavota, 2024). Explainability, which involves generating human-readable justifications that psychologists can validate, is distinct from interpretability, which requires insight into internal mechanisms such as feature weighting—often infeasible with proprietary models such as GPT-4. By utilizing GPT-4's explainability, this study ensures outputs align with psychological expertise, fostering trust and supporting responsible deployment in stress detection. These challenges highlight the need for innovative approaches to adapt general-purpose LLMs without extensive retraining (Wang et al., 2024).

Prompt engineering offers a resource-efficient solution to tailor general-purpose LLMs for domain-specific tasks through carefully designed prompts, bypassing the need for fine-tuning (Priyadarshana, Senanayake, Liang, & Piumarta, 2024). Techniques, specifically zero-shot and few-shot learning, enable LLMs to perform tasks with minimal

* Corresponding author at: Bernoulli Institute, University of Groningen, Groningen, The Netherlands.
E-mail addresses: n.esmi.rudbardeh@rug.nl (N. Esmi), shahbahrami@guilan.ac.ir (A. Shahbahrami), yas.nabauti@gmail.com (Y. Nabati), bita.rezaeidogaheh@gmail.com (B. Rezaei), g.n.gaydadjiev@tudelft.nl (G. Gaydadjiev), peter.de.jonge@rug.nl (P. de Jonge).

https://doi.org/10.1016/j.actpsy.2025.105462
Received 2 April 2025; Received in revised form 16 August 2025; Accepted 20 August 2025
Available online 28 August 2025

or no task-specific training by embedding task descriptions or examples in prompts (Brown et al., 2020; Kojima, Gu, Reid, Matsuo, & Iwasawa, 2022). These methods have enhanced LLM performance in diverse natural language processing (NLP) tasks, such as sentiment analysis and question answering, by eliciting robust reasoning capabilities (Pal, Bhattacharya, Lee, & Chakraborty, 2024; Yang, Zhang et al., 2024). In mental health, where data heterogeneity and ambiguity (e.g., nuanced expressions or emojis) pose challenges, prompt engineering can guide models to focus on domain-relevant cues, reducing computational barriers while maintaining scalability (Chung et al., 2024; Taori et al., 2023). Despite its promise, the application of prompt engineering to stress detection remains underexplored, particularly in leveraging expert feedback to refine model outputs (Priyadarshana et al., 2024).

This study employs prompt engineering to enhance GPT-4's performance in detecting stress in social media posts from the Dreaddit dataset, labeled by psychologists as "stressful" or "non-stressful" (Ji et al., 2022). The dataset's complexity, with ambiguous posts and diverse linguistic patterns, tests GPT-4's ability to distinguish subtle stress indicators. We iteratively refined prompts using psychologist-informed hints, incorporating feedback from misclassifications to improve reasoning and accuracy. This approach achieved a 17% accuracy improvement, surpassing Mental-RoBERTa by 5%, without accessing GPT-4's internal parameters (Brown et al., 2020; Kojima et al., 2022). By leveraging GPT-4's explainability, we aligned outputs with mental health expertise, demonstrating a scalable, transparent method for adapting general-purpose LLMs to sensitive tasks. Our findings highlight prompt engineering's potential to bridge the gap between generalization and specialization, offering a practical solution for mental health applications (Chung et al., 2024; Taori et al., 2023).

The remainder of this paper is organized as follows: Section 2 reviews related work, focusing on prompt engineering for general-purpose and domain-specific LLMs for mental health. Section 3 details our methodology, including the dataset, preprocessing, prompt design, experimental setup, and evaluation steps. Section 4 presents the results, analyzing error patterns, the impact of psychologist-informed hints, and comparisons with domain-specific models and zero-shot and few-shot baselines. Section 5 discusses key findings, including performance differences across GPT-4 versions, limitations in handling ambiguous cases, and implications for explainability. Finally, Section 6 concludes with key insights and directions for future research.

2. Related work

This section reviews the literature relevant to our study, focusing on advancements in prompt engineering and the application of LLMs in mental health. It explores prompt engineering techniques for general-purpose LLMs, such as zero-shot and few-shot learning, and their role in enhancing performance for tasks like stress detection. Additionally, domain-specific LLMs tailored for mental health are discussed, highlighting their strengths, limitations, and resource requirements.

2.1. Prompt engineering for general-purpose LLMs

General-purpose LLMs, such as PaLM, GLaM, Llama, Mistral, Grok, and GPT-4, demonstrate remarkable versatility across tasks like translation, summarization, question answering, and sentiment analysis, processing human-like text with trillions of parameters (Chuang, Tang, Jiang, & Hu, 2024; Lee, Bahukhandi, Liu, & Ma, 2025; Shao, Yu, Wang, & Yu, 2023; Singhal et al., 2025; Viggiato & Bezemer, 2024; Zhu, Pan, & Xiong, 2024). In mental health, these models analyze social media posts to detect stress or provide conversational support, capturing complex language patterns and contextual nuances (Kortemeyer, 2024; Tufano et al., 2024; Zhang, Deng, Liu, Pan, & Bing, 2024). Their explainability, generating transparent reasoning, is critical for stakeholder trust in sensitive domains like psychology (Chang, Shi et al., 2024). However, their general-purpose design often results in lower accuracy for specialized tasks compared to fine-tuned models, as they may overlook subtle mental health cues, such as implicit stress indicators (Bauer et al., 2024; Yang, Tao et al., 2024). Additionally, closed-source architectures and high fine-tuning costs limit customization, particularly for researchers with constrained resources (Gandy et al., 2025), while data noise, such as ambiguous or mislabeled posts, exacerbates challenges in mental health applications (Pal et al., 2024).

Prompt engineering has transformed the adaptability of general-purpose LLMs for domain-specific tasks, offering a cost-effective alternative to fine-tuning (Priyadarshana et al., 2024; Wang et al., 2024). Recent advancements highlight the efficacy of zero-shot and few-shot learning paradigms. For example, zero-shot learning, introduced by Brown et al. (2020), enables LLMs to perform tasks without training by using descriptive prompts, while Kojima et al. (2022) showed that few-shot learning, which incorporates labeled examples, yields robust reasoning comparable to that of fine-tuned models. These approaches have enhanced LLM performance in tasks like sentiment analysis and question answering (Pal et al., 2024; Yang, Zhang et al., 2024). In mental health, where social media posts exhibit ambiguity (e.g., nuanced emotions, emojis, or slang), prompt engineering guides models to prioritize relevant cues, such as emotional intensity or contextual triggers (Chung et al., 2024; Taori et al., 2023). Techniques like chain-of-thought prompting further improve reasoning by encouraging step-by-step analysis, critical for complex tasks like stress detection (Priyadarshana et al., 2024). Despite these advances, applying prompt engineering to mental health remains underexplored, particularly in integrating expert feedback to refine prompts iteratively. Our study addresses this gap by designing prompts with psychologist-informed hints, optimizing GPT-4's accuracy for stress detection in Dreaddit posts, aligning with broader prompt engineering literature (Yang, Zhang et al., 2024). This approach demonstrates that prompt engineering can effectively tailor general-purpose LLMs for sensitive tasks, offering a scalable, resource-efficient solution without altering model architecture.

2.2. Domain-specific LLMs for mental health

Domain-specific LLMs, fine-tuned on mental health datasets, are customized to detect emotional states, stress, or disorders, ensuring ethical, privacy-conscious, and contextually relevant outputs (Hu et al., 2024). Mental-RoBERTa, built on RoBERTa and fine-tuned with Reddit posts from communities like r/depression and r/Anxiety, excels in classifying stress, depression, and suicidal ideation by leveraging domain-specific linguistic patterns (Ji et al., 2022). MentalQLM, a lightweight model with 0.5 billion parameters, employs instruction tuning and dual Low-rank Adaptation (LoRA) for efficient binary and multi-class classification, supporting real-time mental health applications (ShiJiayu et al., 2024). Mental-Flan-T5 utilizes chain-of-thought reasoning and instruction tuning to analyze complex texts, adapting to mental health tasks with robust few-shot performance (Chung et al., 2024; Xu et al., 2024). Mental-Alpaca, optimized for user-friendly interactions, enhances contextual reasoning for diverse mental health scenarios (Taori et al., 2023; Xu et al., 2024). However, these models require fine-tuning, demanding large, expert-labeled datasets and significant computational resources, which are often inaccessible in low-resource settings (Gandy et al., 2025). Data heterogeneity, such as varying expressions of stress across platforms, further complicates training (Yang, Tao et al., 2024). Our approach mitigates these challenges by using prompt engineering to achieve comparable performance with GPT-4, bypassing the need for resource-intensive fine-tuning.

3. Methodology and experimental design

In this section, the dataset, specifically designed for stress detection in social media posts, is first explored. Subsequently, the minimal preprocessing approach adopted to preserve the data's authenticity and nuanced emotional and contextual cues is described. This strategy ensures the model processes posts as they appear in real-world social media environments, thereby enhancing the practical applicability of our stress detection methodology.

3.1. Dataset and preprocessing

To develop and evaluate our prompt engineering approach for stress detection, we utilized the Dreaddit dataset, a Reddit-based resource specifically designed for stress analysis in social media posts (Xu et al., 2024). The dataset comprises 3553 annotated segments derived from 2929 posts, with an average length of 420 tokens per post, distinguishing it from shorter-form platforms like Twitter. These segments, sourced from Reddit communities, capture a diverse range of linguistic patterns, emotional narratives, and situational triggers, making Dreaddit an ideal testbed for evaluating LLMs like GPT-4 in detecting nuanced stress indicators. The dataset is split into an initial prompt evaluation set (2838 segments, 80%) and an independent test set (715 segments, 20%), maintaining near-balance with 51.6% and 52.4% stressful labels, respectively. Each segment was annotated by at least five annotators, with labels determined through majority voting to ensure robustness despite subjective interpretations of stress. This annotation process mitigates challenges such as linguistic diversity (e.g., slang, idioms), cultural nuances, and potential label noise, which are common in user-generated content (Raffel et al., 2020). Prompt hints are explicit, actionable insights derived from error analysis, serving as strategic suggestions to guide the LLM's reasoning. To prevent data leakage, these hints were derived exclusively from the evaluation set during error analysis, with updated prompts evaluated on the independent test set. Dreaddit's rich, context-heavy content provides a robust foundation for validating our prompt engineering methodology, offering deeper insights into stress detection compared to traditional mental health datasets.

Given the dataset's complexity, we adopted a minimal preprocessing approach to preserve the authenticity of social media posts and capture nuanced emotional and contextual cues critical for stress detection (Esmi, Shahbahrami, Gaydadjiev, & de Jonge, 2025). All textual elements, including emojis, hashtags, misspellings, and punctuation, were retained to reflect the natural tone and affective significance of the posts. Emojis, such as sadness (😔) or anxiety (😥), are particularly important for signaling stress, especially among younger users, and were preserved to leverage their emotional weight. No tokenization, normalization, or spelling corrections were applied, as such interventions could alter the posts' emotional intent or contextual meaning. Data noise, such as ambiguous phrases or inconsistent formatting, was addressed by relying on GPT-4's robust contextual understanding, effectively guided by carefully designed prompts. This strategy ensures that the model processes posts as they appear in real-world social media environments, enhancing the practical applicability of our stress detection methodology. By preserving the dataset's inherent complexity, our approach aligns with Dreaddit's design to reflect authentic user-generated content, enabling robust analysis of diverse, nuanced expressions of stress.

3.2. Prompt engineering framework

To adapt GPT-4 for stress detection on the dataset, we developed a structured prompt engineering framework, grounded in established principles (Liu & Chilton, 2022). Fig. 1 illustrates a zero-shot prompt example, comprising, part A, a social media post with an associated question, and part B, GPT-4's response to that question. The zero-shot prompt (P_{ZS}), designed to obtain binary classifications ("Yes" for stressful, "No" for non-stressful), comprises four components: a task description (T), the social media post (P_S), a classification query (P_Q), and an output modifier (O_M). The task description (T) instructs GPT-4 to classify posts based on linguistic and emotional cues, such as tone, sentiment, or trigger words relevant to mental health (e.g., expressions of anxiety or distress). The social media post is presented verbatim to preserve its raw content, ensuring that contextual nuances, including emojis and slang, are retained. The classification query requests a binary "Yes" or "No" response to minimize ambiguity, while the output modifier enforces standardized outputs for consistency and reproducibility, as shown in Eq. (1). This structure, minimizes response variability and ensures that GPT-4 focuses on stress detection across diverse posts.

$$P_{ZS} = T + P_S + P_Q + O_M \tag{1}$$

While the initial prompt provided a clear framework, it lacked domain-specific guidance, leading to errors in capturing subtle stress indicators, such as implicit anxiety or situational stressors. To address this, we implemented an iterative refinement process informed by psychologist expertise, following a structured error analysis as depicted in Fig. 2. After zero-shot prompting (Step 1), we extracted posts with false positive (FP) and false negative (FN) outcomes from the initial prompt evaluation set (Steps 2 and 3). These misclassified posts were reviewed by psychologists, who identified root causes of errors, such as misinterpretation of neutral linguistic cues or failure to detect contextual triggers (e.g., financial or social stressors) (Esmi et al., 2025). The psychologists' analyses were grouped into categories, including linguistic ambiguity, contextual oversight, and emoji misinterpretation, and ranked by frequency of occurrence. Each category was distilled into concise, hints to guide prompt updates. For example, one hint instructed GPT-4 to prioritize situational context (e.g., references to job loss) over isolated emotional keywords, reducing false positives in ambiguous posts (Steps 4, 5, and 6) in Fig. 2.

These psychologist-informed hints (H) were incorporated into the prompt, enhancing its ability to capture stress-related patterns, as shown in Eq. (2) (Step-7). The updated prompt structure explicitly guided GPT-4 to focus on domain-relevant cues, such as emotional intensity or situational triggers. For instance, a hint might direct the model to consider phrases like "overwhelmed with work" as indicators of stress, even in the absence of explicit emotional terms. This iterative refinement process was also evaluated to measure the impact of psychologists' hints on GPT-4's responses (Steps-8, 9).

$$P_{ZS} = T + P_S + P_Q + O_M + H \tag{2}$$

This approach highlights the importance of iterative, domain-specific prompt refinement in addressing the challenges of stress detection. By leveraging psychologist-informed hints, our approach bridges the gap between general-purpose LLMs and specialized tasks, offering a scalable, resource-efficient alternative to fine-tuning (Priyadarshana et al., 2024).

3.3. Experimental design

To evaluate the efficacy of the prompt engineering framework for stress detection, experiments utilized GPT-4, specifically the January 2025 version of the OpenAI Application Programming Interface (API) (Kojima et al., 2022). The model was configured with specific parameters to balance response creativity and coherence, thus ensuring reliable binary classifications. We set the `temperature` to 0.7 based on preliminary empirical evaluations. This value allowed moderate response variability, which was beneficial for eliciting nuanced indicators of stress in diverse social media contexts. At the same time, reproducibility was ensured by constraining the task to binary outputs ("Yes" or "No") with a standardized prompt design. In this way, the

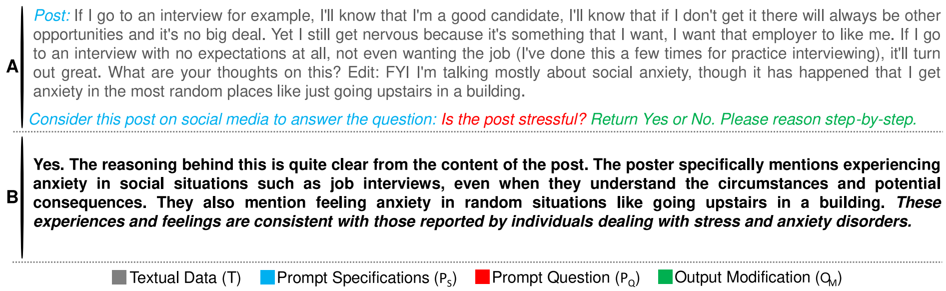

Fig. 1. Structure of the initial prompt for stress detection, showing textual components: task description (T), social media post (P_S), classification query (P_Q), and output modifier (O_M), as per Eq. (1).

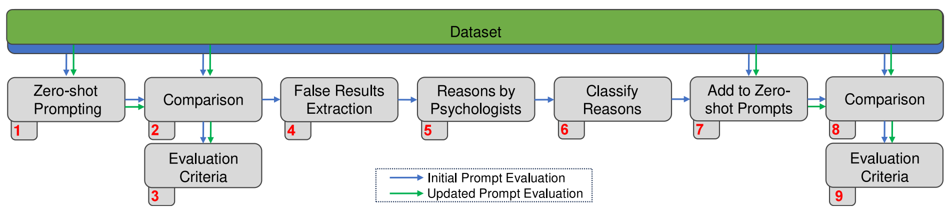

Fig. 2. Proposed approach with nine stages for enhancing GPT-4's stress detection via prompt engineering. Blue lines indicate steps on the initial prompt evaluation set before hints, and green lines represent steps on the test set with updated prompts. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

slightly higher temperature encouraged sensitivity to subtle linguistic cues while avoiding inconsistent classifications, thus maintaining both robustness and task focus across runs. The `top-p` parameter was set to 0.9 to support diverse yet relevant token sampling. A `frequency_penalty` of 0.0 was applied, as the binary output format does not require the suppression of repetition and thus avoids penalizing repeated tokens. These settings were chosen to optimize GPT-4's performance for concise, consistent outputs, while accommodating the linguistic diversity of the dataset (Xu et al., 2024).

For the Llama model, the psychologist-informed prompt engineering protocol was not fully applied due to practical limitations, with its inclusion primarily serving as an open-weight baseline. The model was executed with the following parameters: `ture` 0.7, `top-p` 0.8, and `frequency-penalty` 0.0. Furthermore, the model's `context window` was limited to 2048 tokens. Model inference was conducted via the Together AI API for the Llama 3.1 405B model (Arora, Sayeed, Licorish, Wang, & Treude, 2024; Together, 2025).

For all domain-specific baseline models (M-Alpaca, M-Flan-T5, M-QLM, and M-RoBERTa), a consistent fine-tuning protocol was applied to ensure reproducibility and optimize performance on the dataset. All models were fine-tuned using Parameter-Efficient Fine-Tuning (PEFT) methods, primarily LoRA, with a learning rate of 5e−5. Training was conducted for 3 epochs, utilizing an AdamW optimizer with a weight decay of 0.01 and a warmup ratio of 0.05. A per-device batch size of 4 was used, with gradients accumulated over 4 steps to achieve an effective batch size of 16. Mixed precision training (FP16/BF16) was employed for computational efficiency. For consistent results across all runs, a seed of 42 was set, and the best performing checkpoint for each model was selected based on its F1-score on the validation set.

Experiments were executed on a cloud-based platform equipped with sufficient computing resources to handle API requests efficiently, minimizing latency and ensuring scalability, which was critical for concurrently processing the 3553 segments of the dataset. This setup allowed for efficient management of numerous API calls without significant queuing or delays. We monitored API response times and error rates to maintain stability, addressing potential challenges in real-time deployment scenarios. Each social media post was processed in a single API call, with no additional preprocessing or post-processing beyond the output modifier (O_M) defined in the prompt structure. This approach ensured consistency across evaluations and supported reproducibility despite GPT-4's proprietary nature (Brown et al., 2020; Together, 2025). The experimental protocol involved applying the initial and refined prompts to the dataset, split into an evaluation set (80%) and an independent test set (20%). The initial zero-shot prompt, formatted as Eq. (1), was first applied to the evaluation set to establish a baseline performance (Step 1) in Fig. 2. Following error analysis and prompt refinement with psychologist-informed hints (Steps 4–6), the updated prompt, based on Eq. (2) was applied to the test set to assess performance improvements (Steps 8–9). This two-phase approach ensured that prompt refinements were derived solely from the evaluation set, preventing data leakage and enhancing generalizability. The binary responses generated by GPT-4 were compared against Dreaddit's ground-truth labels to evaluate classification performance, with results analyzed using confusion matrices and performance metrics. This experimental design facilitated a robust evaluation of our prompt engineering approach, enabling systematic comparisons between zero-shot and hint-enhanced prompting scenarios. By leveraging a stable API configuration and scalable computing resources, we ensured efficient processing of the dataset while maintaining methodological rigor (Priyadarshana et al., 2024).

3.4. Evaluation and performance metrics

To evaluate the effectiveness of our prompt engineering framework for stress detection using GPT-4 on the dataset, we implemented a systematic evaluation process, as outlined in Steps 2, 3, 8, and 9 of our methodology. The evaluation involved comparing GPT-4's binary outputs against the ground-truth labels provided by the dataset (Xu et al., 2024). The process was conducted in two phases: first, applying the initial zero-shot prompt (Eq. (1)) to the evaluation set to establish a baseline, and second, applying the refined prompt with psychologist-informed hints (Eq. (2)) to the independent test set. This two-phase approach ensured that prompt refinements were derived solely from the evaluation set, preventing data leakage and supporting generalizability (Kojima et al., 2022).

A

Post: If I go to an interview for example, I'll know that I'm a good candidate, I'll know that if I don't get it there will always be other opportunities and it's no big deal. Yet I still get nervous because it's something that I want, I want that employer to like me. If I go to an interview with no expectations at all, not even wanting the job (I've done this a few times for practice interviewing), it'll turn out great. What are your thoughts on this? Edit: FYI I'm talking mostly about social anxiety, though it has happened that I get anxiety in the most random places like just going upstairs in a building.
Consider this post on social media to answer the question: Is the post stressful? Return Yes or No. Please reason step-by-step.
To improve the prompt and reduce false positives, focus on emphasizing specific indicators of stress and providing clearer distinctions between stressful and non-stressful scenarios. Here are some guiding sentences for refining the prompt:
1.Distinguish Stress Attribution: Ensure the prompt explicitly directs the model to evaluate the speaker's tone, emotional language, and the context of frustration or tension rather than focusing on generic descriptions or neutral statements. **2.Highlight Emotional Cues:** Add guidance to identify words or phrases that signify stress, such as "frustrated," "overwhelmed," "angry," "worried," or contextually negative sentiments, while avoiding misinterpretation of neutral or positive remarks. **3.Avoid Misinterpreting Neutrality:** Instruct the model to avoid labeling as stressed individuals who provide factual, calm, or constructive reflections on challenging situations, even if the context involves difficulties. **4.Incorporate Contextual Analysis:** Guide the model to consider the broader context of the text to determine if the described person is managing the situation effectively, suggesting they are not stressed, rather than solely relying on specific phrases. **5.Focus on Self-Reported Feelings:** Emphasize that stress detection should primarily consider self-reported indicators of stress or clearly implied emotional distress, rather than assumptions based on situational descriptions. **6.Reduce Assumptions:** Add instructions to minimize assumptions about stress based on third-party actions or external circumstances unless explicitly linked to the person's emotional state. **7.Weight Responses Over Situations:** Highlight the importance of prioritizing how the person discusses and responds to their situation rather than the situation itself. **8.Clarify Ambiguity:** Encourage the model to flag ambiguous cases where stress is unclear instead of confidently labeling them as stressed, reducing overgeneralization.

B

Is the post stressful? No.
Reasoning step-by-step:
Distinguish Stress Attribution: The post reflects on feelings of nervousness during interviews and social anxiety in certain situations. However, the tone is reflective and self-aware rather than overtly stressed. The speaker does not express frustration, overwhelming worry, or anger. **Highlight Emotional Cues:** Key phrases such as "I'll know that I'm a good candidate" and "it's no big deal" indicate confidence and acceptance. Although they mention "nervous" and "social anxiety," these are framed in a matter-of-fact way rather than as overwhelming stress. **Avoid Misinterpreting Neutrality:** The speaker describes past experiences and general reflections about managing social anxiety. These are constructive observations rather than negative or distressed remarks. **Incorporate Contextual Analysis:** The broader context shows the speaker has strategies to manage their feelings (e.g., practicing interviews without expectations) and is not consumed by stress. **Focus on Self-Reported Feelings:** The self-reported feelings include nervousness and social anxiety, but there are no indicators that these feelings are debilitating or causing significant distress. **Reduce Assumptions:** There is no explicit mention of stress beyond the nervousness associated with specific situations. Assumptions about distress are unnecessary as the speaker discusses their experiences calmly. **Weight Responses Over Situations:** The speaker responds to situations thoughtfully and with perspective, acknowledging their nervousness but not dwelling on it negatively. **Clarify Ambiguity:** There is no ambiguity; the tone is self-reflective, indicating awareness rather than distress.
Conclusion:
The post discusses nervousness and social anxiety in a calm, reflective, and constructive manner, indicating the speaker is managing their emotions effectively. This does not meet the criteria for stress as outlined in the refined prompt.

Textual Data (T) Prompt Specifications (P_S) Prompt Question (P_Q) Output Modification (O_M) Hints

Fig. 3. Example of prompt engineering for a non-stressful text: GPT-4, with hints, correctly classified the text. Part A shows the prompt with psychologist-informed hints, and Part B presents GPT-4's response and reasoning.

For each phase, GPT-4's responses were categorized into four outcomes: True Positive (TP, correctly identified stressful posts), True Negative (TN, correctly identified non-stressful posts), False Positive (FP, non-stressful posts incorrectly classified as stressful), and False Negative (FN, stressful posts incorrectly classified as non-stressful). These outcomes were used to construct confusion matrices for both the zero-shot and hint-enhanced prompting scenarios, providing a structured framework to analyze classification performance (Brown et al., 2020). The confusion matrices capture the distribution of TP, TN, FP, and FN cases, enabling a detailed assessment of the model's ability to distinguish between stressful and non-stressful posts. This approach facilitated the identification of error patterns, such as false positives due to misinterpretation of neutral cues, which informed prompt refinements.

Performance was assessed using four standard metrics: Accuracy, Precision, Recall, and F1 Score, calculated from the confusion matrices (Kojima et al., 2022). These metrics computed for both prompting scenarios to quantify the impact of psychologist-informed hints on GPT-4's performance (Esmi et al., 2025).

4. Experimental results

This section examines the psychologists' reasoning behind GPT-4's errors, analyzes the impact of incorporating this reasoning into zero-shot prompting, and compares GPT-4's performance with domain-specific models tuned for mental health analysis (Xu et al., 2024).

4.1. Error analysis and hint development

On the initial prompt evaluation set, GPT-4 exhibited a 33% error rate, primarily due to false positives, which accounted for 90% of errors (Esmi et al., 2025). Psychologists analyzed these errors (Steps 4–6) in Fig. 2, identifying eight hint sentences, ranked by frequency of occurrence (44%, 26%, 10%, 6%, 4%, 4%, 4%, 2%), as shown in Fig. 3. These hints addressed common misclassification causes, such as overreliance on neutral linguistic cues or failure to detect contextual triggers (e.g., financial stressors). For example, one hint instructed GPT-4 to prioritize situational context over isolated emotional keywords, reducing FPs in ambiguous posts (Yang, Tao et al., 2024). This process highlights the critical role of expert-guided error analysis in refining prompts for mental health applications (Chung et al., 2024).

Table 1
Performance comparison between zero-shot prompting and added hints.

Method	Acc. (%)	Pre. (%)	Rec. (%)	F1. (%)
Zero-shot	72.0	66.3	**94.6**	77.9
Added hints	**89.0**	**87.9**	91.4	**89.5**

4.2. Impact of hints on performance

Fig. 4 presents the confusion matrix for the test set in two scenarios: zero-shot prompting and after incorporating hints. In the zero-shot scenario, FPs dominated errors, reflecting GPT-4's tendency to misinterpret neutral posts as stressful.

Adding hints significantly reduced FPs (from 133 to 12 cases), though it slightly increased FNs by 12 cases, indicating a trade-off in sensitivity (Kojima et al., 2022). This shift suggests that hints improved GPT-4's ability to discern nuanced stress indicators, aligning its outputs more closely with human annotations (Pal et al., 2024).

Table 1 compares performance metrics between zero-shot and added hints scenarios. The added hints scenario improved accuracy by 17% (from 72.0% to 89.0%), precision by 21.6% (from 66.3% to 87.9%), and F1 score by 11.6% (from 77.9% to 89.5%), despite a slight recall decrease (from 94.6% to 91.4%) due to increased FNs (Liu & Chilton, 2022).

To validate the generalizability of our prompt engineering approach and address concerns about overfitting to Dreaddit's patterns, we conducted a 5-fold cross-validation experiment. The dataset was divided into five equal folds, with each fold serving as the test set while the remaining four were used for deriving hints (Steps 4–6) in Fig. 2. While hints were initially derived from the training data of each fold, our analysis of these derivations revealed significant commonalities; consequently, a single, representative set of eight hints was applied consistently across all cross-validation folds. Updated prompts were evaluated on the test fold (Steps 7–9) in Fig. 2, and this process was repeated for each fold. The average accuracy across folds was 88.5% (standard deviation 1.2%), compared to 72.0% for zero-shot prompting, confirming consistent performance gains.

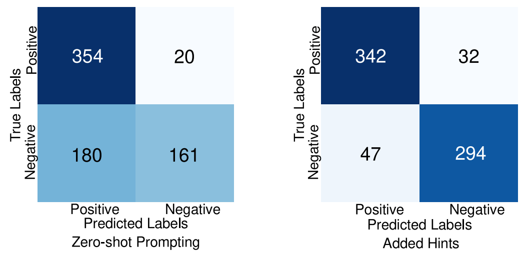

Fig. 4. Confusion matrix for test set posts: zero-shot prompting vs. using hints. Adding hints significantly reduced the false positive rate.

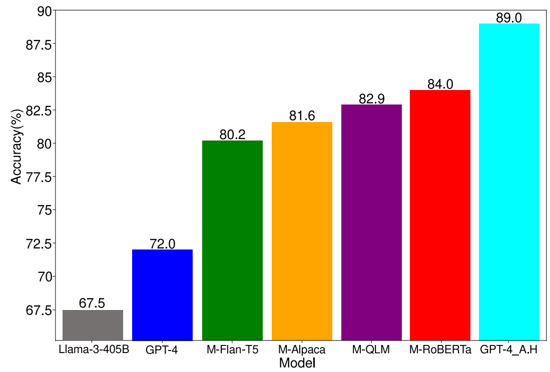

Fig. 5. Comparison of GPT-4 in zero-shot and Added Hint (A.H) modes with fine-tuned mental health models on the Dreaddit test set.

Table 2
Comparison of models for stress detection on Dreaddit, sorted by accuracy. Models are evaluated by type (General-Purpose, G-P; Domain-Specific, D-S), parameter count (Para. in billions), fine-tuning requirement (F-T), accuracy, key features, pre-training data, and pros and cons.

Model	Type	Para.	F-T	Acc.	Key features	Pre-train data	Pros & Cons
Llama-3-405B	G-P	405	No	67.5	Open source	Diverse datasets	Lower accuracy, fine-tunable
GPT-4	G-P	1700	No	72.0	General capabilities	Diverse datasets	Low initial accuracy, no fine-tuning
M-Alpaca	D-S	7	Yes	80.2	Contextual reasoning	Mental health data	Strong performance, instruction-limited
M-Flan-T5	D-S	11	Yes	81.6	Few-shot learning	Mental health data	Good accuracy, resource-heavy
M-QLM	D-S	7	Yes	82.5	LoRA adaptation	Mental health data	Lightweight, less powerful
M-RoBERTa	D-S	0.5	Yes	84.0	Transfer learning	Reddit	High accuracy, resource-intensive
GPT-4_A.H	G-P	1700	No	**89.0**	Prompt engineering	Diverse datasets	High accuracy, no fine-tuning

4.3. Comparison with domain-specific models

Fig. 5 compares GPT-4's accuracy in zero-shot and Added Hints (A.H) scenarios with four domain-specific LLMs fine-tuned for mental health: Mental-RoBERTa, Mental-Flan-T5, Mental-Alpaca, and Mental-QLM. GPT-4 (A.H) achieved the highest accuracy (89.0%), surpassing Mental-RoBERTa (84.0%) by 5% and others by larger margins (Ji et al., 2022; ShiJiayu et al., 2024). The zero-shot GPT-4 (72.0%) had the lowest accuracy, underscoring the necessity of hints for competitive performance (Taori et al., 2023).

Table 2 provides a detailed comparison of models for stress detection, including General-Purpose (G-P) and Domain-Specific (D-S) LLMs, sorted by accuracy. Llama-3-405B, a G-P model with 405 billion parameters, achieved 67.5% accuracy in zero-shot settings, limited by its lack of D-S tuning (Taori et al., 2023). GPT-4 (zero-shot) scored 72.0%, while D-S models such as Mental-RoBERTa (0.5B parameters, 84.0%) and Mental-Flan-T5 (11B parameters, 81.6%) benefited from fine-tuning on mental health data (Chung et al., 2024). However, GPT-4 (A.H) reached 89.0% without fine-tuning, highlighting prompt engineering's cost-efficiency compared to resource-intensive fine-tuning (Priyadarshana et al., 2024). Mental-QLM (7B parameters, 82.5%) and Mental-Alpaca (7B parameters, 80.2%) showed strong performance but required significant computational resources (ShiJiayu et al., 2024). These results demonstrate that prompt engineering can optimize general-purpose LLMs to outperform specialized models, especially in resource-constrained settings (Gandy et al., 2025).

4.4. Comparison with zero-shot and few-shot baselines

To contextualize our prompt engineering approach, we compared it to standard zero-shot and few-shot learning baselines using GPT-4 on the dataset test set (Brown et al., 2020). The zero-shot baseline used a simple prompt instructing GPT-4 to classify posts as stressful or non-stressful without hints. The few-shot baseline included five labeled examples (three stressful, two non-stressful) from the evaluation set to guide predictions, as shown below: **Few-shot sample:**

Classify the following post as stressful (Yes) or non-stressful (No). Examples:
1. Post: "Feeling crushed by deadlines and no one cares". Answer: Yes
2. Post: "Just got a promotion, so excited!" Answer: No
3. Post: "I'm anxious about my exams and failing". Answer: Yes
4. Post: "Had a great day at the beach!" Answer: No
5. Post: "Can't handle this stress anymore, I'm breaking". Answer: Yes
Now classify:
Post: "I'm so overwhelmed with work and can't sleep". Answer:

As depicted in 3, our method, with psychologist-informed hints, achieved 89.0% accuracy, compared to 72.0% for zero-shot and 78.5% for few-shot baselines. The few-shot approach improved over zero-shot by providing contextual examples, but it underperformed compared to our method, which leveraged expert-derived hints to address nuanced stress indicators (Kojima et al., 2022). It is worth noting that the few-shot examples, primarily structured around emotional valence (positive vs. negative), might have biased the model towards sentiment detection rather than the multifaceted construct of stress,

Table 3
Comparison of prompt engineering with zero-shot and few-shot baselines on Dreaddit test set.

Method	Accuracy (%)
Zero-shot (No Hints)	72.0
Few-shot (5 Examples)	78.5
Prompt engineering (With Hints)	89.0

potentially contributing to its relatively weaker performance. This emphasizes the superiority of iterative, domain-specific prompt refinement over generic prompting strategies, aligning with advancements in task-specific optimization (Shao et al., 2023).

5. Discussion

The following discussion evaluates the effectiveness of our prompt engineering approach for stress detection using GPT-4 on the dataset. We analyze key findings, including performance variations across GPT-4 versions, limitations in handling ambiguous cases, and the implications of relying on post-hoc explanations. These insights highlight the strengths and challenges of adapting general-purpose LLMs for mental health applications through prompt engineering.

5.1. Different GPT-4 versions comparison

Our study on stress detection using GPT-4 with the dataset revealed notable performance differences between the January 2025 and May 2025 model versions (Xu et al., 2024). In January, psychologist-informed hints improved classification accuracy by 17%, from 72% to 89%, highlighting GPT-4's reliance on explicit guidance to detect stress-related cues like emotional intensity or situational triggers (Esmi et al., 2025). Conversely, the May version achieved a baseline accuracy of 87% without hints, with hints yielding a modest 3% improvement to 90%. This suggests significant advancements in the May model, likely due to enhanced training on diverse, social media-like texts or architectural refinements improving generalization (Brown et al., 2020). The diminished impact of hints indicates that the model has internalized many stress indicators previously provided externally, reducing their necessity. The model's high initial accuracy may be limited by unclear or noisy data in Dreaddit's user-generated posts (Pal et al., 2024). To improve further, future prompts should focus on complex, unclear posts where the model struggles. This shows the need to update prompt engineering strategies to work with more advanced models, ensuring prompts tackle specific weaknesses in mental health applications (Liu & Chilton, 2022).

5.2. GPT-4 incorrect classification analysis with Grok

Despite psychologist-informed hints, GPT-4 occasionally misclassified non-stressful Dreaddit posts as stressful, revealing limitations in our prompt engineering approach (Kojima et al., 2022). We analyzed 20 misclassified posts where updated prompts altered GPT-4's output but still failed, using Grok due to resource constraints preventing comparisons with fine-tuned models like Mental-RoBERTa (Ji et al., 2022). Grok agreed with GPT-4 in most cases but identified deficiencies in five, such as overreliance on neutral linguistic cues or misinterpretation of emojis, aligning with psychologists' critiques of prompt specificity (Esmi et al., 2025). Fig. 6 illustrates a case where GPT-4 misinterpreted neutral expressions (Part A: prompt with hints; Part B: GPT-4's response; Part C: Grok's analysis highlighting keyword overemphasis). This suggests that prompts require further refinement to handle ambiguous or contextually nuanced posts robustly.

5.3. Post-hoc explanations

Our prompt engineering approach significantly improved GPT-4's stress detection performance, but its reliance on post-hoc explanations for explainable outputs poses challenges (Chang, Shi et al., 2024). These explanations, while valuable for validating outputs and building stakeholder trust in mental health applications, lack true algorithmic transparency, as GPT-4's closed-source nature obscures internal mechanisms like attention patterns or parameter weights (Gandy et al., 2025). This limitation hinders full understanding of decision-making processes, critical in sensitive domains where precise reasoning is essential (Yang, Tao et al., 2024). Additionally, as of April 2025, potential discontinuation risks for GPT-4 threaten reproducibility, a broader challenge with proprietary models (Bauer et al., 2024). To address this, our methodology is designed to be model-agnostic, adaptable to open-weight models like Llama or Mistral, leveraging general reasoning capabilities rather than model-specific features (Taori et al., 2023).

5.4. Limited dataset

The Dreaddit dataset, with 3553 annotated segments, provides a robust foundation for evaluating our prompt engineering approach, capturing nuanced stress expressions in Reddit posts (Hu et al., 2024). However, its platform-specific linguistic and contextual patterns may limit generalizability to other social media platforms like Twitter or Instagram, which differ in post length, user demographics, and expression styles (Priyadarshana et al., 2024). For instance, Twitter's concise format or Instagram's visual-heavy content may require tailored prompts to detect stress effectively. Additionally, dataset's reliance on human annotations introduces potential label noise, which may affect model performance (Pal et al., 2024). Incorporating cross-platform datasets could enhance data quality and generalizability.

5.5. Reasoning in just incorrect cases

Our study has focused on analyzing GPT-4's reasoning primarily for incorrect classifications, aiming to identify prompt improvement opportunities (Esmi et al., 2025). While for correct classifications, GPT-4 typically provided reasoning aligned with expert expectations (as verified by psychologists), we did not systematically evaluate the soundness of this reasoning or quantify the model's weighting of factors (e.g., emotional versus contextual cues). This approach, though pragmatic for prompt refinement, potentially masked logical discrepancies even in accurate predictions (Shao et al., 2023). This restricts a comprehensive understanding of GPT-4's decision-making process, which is critical for mental health applications demanding robust reasoning (Kortemeyer, 2024).

Furthermore, this limitation underscores a broader challenge in interpreting LLM outputs: discerning genuine reasoning from sophisticated pattern matching. Recent work highlights the complexities of detecting and measuring reasoning in LLMs, questioning whether observed behaviors reflect true cognitive processes or merely “the illusion of thinking” (Shojaee et al., 2025). Studies on the measurement of reasoning in LLMs emphasize the need for rigorous evaluation frameworks beyond simple task performance (Marjanović et al., 2025). Similarly, ongoing debates around methods like Chain-of-Thought prompting caution against over-interpreting step-by-step outputs as definitive proof of reasoning (Chen et al., 2025). While our prompt engineering aims to guide the model towards more structured problem-solving, our current analytical framework does not conclusively differentiate between these underlying mechanisms. Future work should incorporate more advanced methodologies, inspired by these discussions, to critically evaluate the nature of reasoning exhibited by LLMs in high-stakes domains like mental health.

Fig. 6. Prompt engineering example for a non-stressful Dreaddit post misclassified by GPT-4 despite hints. Part A: Prompt with psychologist-informed hints. Part B: GPT-4's response and reasoning. Part C: Grok's analysis highlighting deficiencies.

5.6. Explainability

In this study, explainability refers to the model's generation of human-readable, post-hoc verbal rationales for its stress detection outputs. These explanations are invaluable for mental health professionals to understand and validate model predictions. However, large language models like GPT-4 are inherently opaque due to their proprietary and black-box nature, which limits direct access to their internal mechanisms. Consequently, quantitatively measuring the fidelity (how accurately the explanation reflects model behavior) and faithfulness (how well it reflects true internal decision-making) of these rationales remains a significant challenge. This is a general and widely recognized limitation within the field of eXplainable AI (XAI) for large language models. Nevertheless, the qualitative utility of providing interpretable justifications is paramount in sensitive domains such as mental health, where trust and clinical validation are essential.

6. Conclusions

This research demonstrates the efficacy of prompt engineering in tailoring GPT-4 for stress detection in social media, achieving a 17% accuracy increase to 89% on the Dreaddit dataset, surpassing domain-specific models like Mental-RoBERTa. By integrating psychologist-informed hints, our approach significantly reduced false positives and generated human-readable rationales that prove crucial for fostering trust and aiding professionals in mental health applications. The methodology's model-agnostic design ensures adaptability to other large language models, enhancing accessibility for resource-constrained settings. Future work should validate this approach on open-weight models, such as Llama or Mistral, to enhance accessibility and transparency, ensuring long-term reproducibility for resource-constrained communities. Additionally, validating the methodology across diverse datasets from multiple platforms, like Twitter and Instagram, will confirm the robustness of psychologist-informed hints across varied linguistic styles and conventions, strengthening real-world stress detection capabilities. Incorporating advanced prompting techniques, such as chain-of-thought prompting, could further improve contextual understanding, reducing errors in complex cases and encouraging a more critical examination of whether observed behaviors reflect actual reasoning or sophisticated pattern matching. Analyzing outputs across all classifications using techniques like attention visualization or feature attribution will ensure consistency and reliability, refining prompt design to enhance GPT-4's ability to handle complex stress detection tasks and improve stakeholder trust. Moreover, hybrid methods combining prompt engineering with interpretability techniques could mitigate transparency issues while maintaining high performance, leveraging tools like Grok for external validation to guide iterative improvements in stress detection.

CRediT authorship contribution statement

Nima Esmi: Writing – original draft, Visualization, Validation, Supervision, Resources, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Asadollah Shahbahrami:** Writing – review & editing, Conceptualization. **Yasaman Nabati:** Validation, Formal analysis. **Bita Rezaei:** Validation, Formal analysis. **Georgi Gaydadjiev:** Writing – review & editing. **Peter de Jonge:** Writing – review & editing.

Ethical approval

This study utilized the publicly available Dreaddit dataset, which consists of anonymized social media posts previously annotated for stress detection. No direct interaction with human participants, collection of biological samples, or generation of new personal data occurred during this research. As the analysis relied solely on pre-existing, de-identified data, ethical approval from an Institutional Review Board (IRB) or equivalent body was not required. The research adheres to ethical guidelines by ensuring the privacy and anonymity of the original data contributors, consistent with standard practices for secondary data analysis in psychological and computational research.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

References

Arora, C., Sayeed, A. I., Licorish, S., Wang, F., & Treude, C. (2024). Optimizing large language model hyperparameters for code generation. http://dx.doi.org/10.48550/arXiv.2408.10577, arXiv.

Bauer, B., Norel, R., Leow, A., Rached, Z. A., Wen, B., & Cecchi, G. (2024). Using large language models to understand suicidality in a social media–based taxonomy of mental health disorders: Linguistic analysis of reddit posts. *JMIR Mental Health*, *11*(1), Article e57234. http://dx.doi.org/10.2196/57234.

Beghetto, R. A., Ross, W., Karwowski, M., & Glăveanu, V. P. (2025). Partnering with AI for instrument development: Possibilities and pitfalls. *New Ideas in Psychology*, *76*, Article 101121. http://dx.doi.org/10.1016/j.newideapsych.2024.101121.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., et al. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, *33*, 1877–1901.

Caruccio, L., Cirillo, S., Polese, G., Solimando, G., Sundaramurthy, S., & Tortora, G. (2024). Claude 2.0 large language model: Tackling a real-world classification problem with a new iterative prompt engineering approach. *Intelligent Systems with Applications*, *21*, Article 200336. http://dx.doi.org/10.1016/j.iswa.2024.200336.

Chang, C., Shi, W., Wang, Y., Zhang, Z., Huang, X., & Jiao, Y. (2024). The path from task-specific to general-purpose artificial intelligence for medical diagnostics: A bibliometric analysis. *Computers in Biology and Medicine*, *172*, Article 104892. http://dx.doi.org/10.1016/j.compbiomed.2024.104892.

Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., et al. (2024). A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, *15*(3), 1–45. http://dx.doi.org/10.1145/3641289.

Chebbi, A., Kniesel, G., Abdennadher, N., & Dimarzo, G. (2024). Enhancing named entity recognition for agricultural commodity monitoring with large language models. In *Proceedings of the machine learning and systems* (pp. 208–213).

Chen, Y., Benton, J., Radhakrishnan, A., Uesato, J., Denison, C., Schulman, J., et al. (2025). Reasoning models don't always say what they think. http://dx.doi.org/10.48550/arXiv.2505.05410, arXiv.

Chuang, Y.-N., Tang, R., Jiang, X., & Hu, X. (2024). Spec: A soft prompt–based calibration on performance variability of large language models in clinical notes summarization. *Journal of Biomedical Informatics*, *151*, Article 104820. http://dx.doi.org/10.1016/j.jbi.2024.104820.

Chung, H. W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., et al. (2024). Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, *25*(70), 1–53, http://jmlr.org/papers/v25/23-0870.html.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *vol. 1*, In *Proceedings of the North American chapter of the association for computational linguistics* (pp. 4171–4186).

Esmi, N., Shahbahrami, A., Gaydadjiev, G., & de Jonge, P. (2025). Suicide ideation detection based on document dimensionality expansion. *Computers in Biology and Medicine*, *192*, Article 110266. http://dx.doi.org/10.1016/j.compbiomed.2024.110266.

Gandy, L. M., Ivanitskaya, L. V., Bacon, L. L., & Bizri-Baryak, R. (2025). Public health discussions on social media: Evaluating automated sentiment analysis methods. *JMIR Formative Research*, *9*(1), Article e57395. http://dx.doi.org/10.2196/57395.

Hu, J., Dong, T., Luo, G., Ma, H., Zou, P., Sun, X., et al. (2024). PsycoLLM: Enhancing LLM for psychological understanding and evaluation. *IEEE Transactions on Computational Social Systems*, *11*(2), 1–13. http://dx.doi.org/10.1109/TCSS.2024.3497725.

Ji, S., Zhang, T., Ansari, L., Fu, J., Tiwari, P., & Cambria, E. (2022). MentalBERT: Publicly available pretrained language models for mental healthcare. In *Proceedings of the language resources and evaluation conference* (pp. 7184–7190).

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2022). Large language models are zero-shot reasoners. *vol. 35*, In *Advances in neural information processing systems* (pp. 22199–22213).

Kortemeyer, G. (2024). Performance of the pre-trained large language model GPT-4 on automated short answer grading. *Discover Artificial Intelligence*, *4*(1), 1–8. http://dx.doi.org/10.1007/s44163-024-00001-8.

Lee, Y.-T., Bahukhandi, A., Liu, D., & Ma, K.-L. (2025). Toward dataset-scale and feature-oriented evaluation of text summarization in large language model prompts. *IEEE Transactions on Visualization and Computer Graphics*, *31*(1), 481–491. http://dx.doi.org/10.1109/TVCG.2024.3456398.

Liu, V., & Chilton, L. B. (2022). Design guidelines for prompt engineering text-to-image generative models. In *Proceedings of the conference on human factors in computing systems* (pp. 1–23).

Marjanović, S. V., Patel, A., Adlakha, V., Aghajohari, M., BehnamGhader, P., Bhatia, M., et al. (2025). DeepSeek-R1 thoughtology: Let's think about LLM reasoning. http://dx.doi.org/10.48550/arXiv.2504.07128, arXiv.

Pal, S., Bhattacharya, M., Lee, S.-S., & Chakraborty, C. (2024). A domain-specific next-generation large language model (LLM) for biomedical engineering and research. *Annals of Biomedical Engineering*, *52*(3), 451–454. http://dx.doi.org/10.1007/s10439-024-00000-0.

Priyadarshana, Y. H. P., Senanayake, A., Liang, Z., & Piumarta, I. (2024). Prompt engineering for digital mental health: A short review. *Frontiers in Digital Health*, *6*, 1–7. http://dx.doi.org/10.3389/fdgth.2024.1410947.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., et al. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, *21*(140), 1–67, http://jmlr.org/papers/v21/20-074.html.

Shao, Z., Yu, Z., Wang, M., & Yu, J. (2023). Prompting large language models with answer heuristics for knowledge-based visual question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 14974–14983).

ShiJiayu, J., Wang, Z., Zhou, J., Liu, C., Sun, P. Z., Zhao, E., et al. (2024). MentalQLM: A lightweight large language model for mental healthcare based on instruction tuning and dual LoRA modules. *MedRxiv*, http://dx.doi.org/10.1101/2024.24319755.

Shojaee, P., Mirzadeh, I., Alizadeh, K., Horton, M., Bengio, S., & Farajtabar, M. (2025). The illusion of thinking: Understanding the strengths and limitations of reasoning models via the lens of problem complexity. http://dx.doi.org/10.48550/arXiv.2506.06941, arXiv.

Singhal, K., Tu, T., Gottweis, J., Sayres, R., Wulczyn, E., Amin, M., et al. (2025). Toward expert-level medical question answering with large language models. *Nature Medicine*, *31*, 1–8. http://dx.doi.org/10.1038/s41591-024-03423-7.

Taori, R., Gulrajani, I., Zhang, T., Dubois, Y., Li, X., Guestrin, C., et al. (2023). Stanford alpaca: An instruction-following llama model. http://dx.doi.org/10.48550/arXiv.2303.17580, arXiv.

Together, A. I. (2025). Llama-3.1-405b model. Retrieved from https://www.together.ai/models/llama-3-1-405b.

Tufano, R., Dabić, O., Mastropaolo, A., Ciniselli, M., & Bavota, G. (2024). Code review automation: Strengths and weaknesses of the state of the art. *IEEE Transactions on Software Engineering*, *50*(2), 338–353. http://dx.doi.org/10.1109/TSE.2023.3348172.

Viggiato, M., & Bezemer, C.-P. (2024). Leveraging the OPT large language model for sentiment analysis of game reviews. *IEEE Transactions on Games*, *16*(2), 493–496. http://dx.doi.org/10.1109/TG.2023.3313121.

Wang, L., Chen, X., Deng, X., Wen, H., You, M., Liu, W., et al. (2024). Prompt engineering in consistency and reliability with the evidence-based guideline for LLMs. *Npj Digital Medicine*, *7*, 41. http://dx.doi.org/10.1038/s41746-024-01029-4.

Xu, X., Yao, B., Dong, Y., Gabriel, S., Rainbow, G., Yu, H., et al. (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), 1–32. http://dx.doi.org/10.1145/3643540.

Yang, R., Tan, F., Lu, W., Thirunavukarasu, A. J., Ting, S. W., & Liu, N. (2023). Large language models in health care: Development, applications, and challenges of large language models in healthcare. *Health Care Science*, *2*(4), 255–263. http://dx.doi.org/10.1007/s44200-023-00045-6.

Yang, M., Tao, Y., Cai, H., & Hu, B. (2024). Behavioral information feedback with achievement gap large language models for mental disorders: Perspectives and insights. *IEEE Transactions on Computational Social Systems*, *11*(3), 3026–3044. http://dx.doi.org/10.1145/3643540.

Yang, H., Zhang, M., Tao, S., Li, S., Wang, M., Wei, D., et al. (2024). Knowledge-prompted estimator: A novel approach to explainable machine translation assessment. In *Proceedings of the conference on advanced communications technology* (pp. 305–310).

Zahid, I., Joudar, S., Albahri, A., Albahri, O., Alamoodi, A., Santamaría, J., et al. (2024). Unmasking large language models by means of OpenAI GPT-4 and Google AI: A deep instruction-based analysis. *Intelligent Systems with Applications*, *23*, Article 200431. http://dx.doi.org/10.1016/j.iswa.2024.200431.

Zhang, X., Cao, R., Wang, Y., Li, S., Xu, H., Gao, K., et al. (2025). A unified prompt-based framework for few-shot multimodal language analysis. *Intelligent Systems with Applications*, *26*, Article 200498. http://dx.doi.org/10.1016/j.iswa.2025.200498.

Zhang, W., Deng, Y., Liu, B., Pan, J., & Bing, L. (2024). Sentiment analysis in the era of large language models: A reality check. In *Findings of the association for computational linguistics* (pp. 3881–3906).

Zhu, S., Pan, L., & Xiong, D. (2024). FEDS-ICL: Enhancing translation ability and efficiency of large language models by optimizing demonstration selection. *Information Processing & Management*, *61*(5), Article 103825. http://dx.doi.org/10.1016/j.ipm.2024.103825.