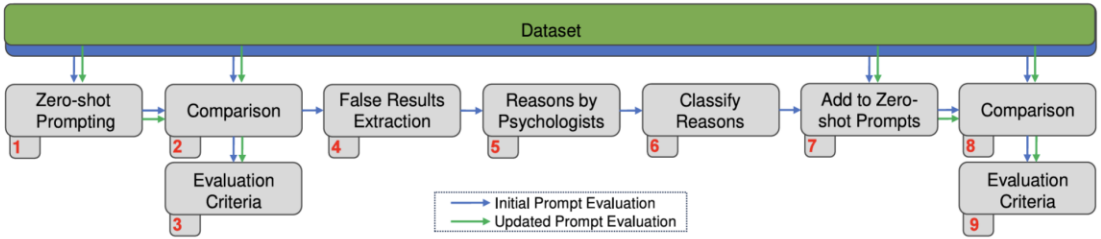
Stress Detection Through Prompt Engineering with a General-Purpose LLM

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

Presented by: Abdullah Mamun
Date: 15 July 2026

Email: a.mamun@asu.edu



Paper

abdullah-mamun.com

𝕏: @AB9Mamun

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.

Summary: A New Frontier in Mental Health AI

This study presents a novel framework using psychologist-informed prompt engineering to significantly improve the ability of the general-purpose AI model, GPT-4, to detect stress in social media posts.

17% Accuracy Boost

The method increased GPT-4's accuracy from 72% to an impressive 89%, surpassing specialized models.

80% Fewer False Positives

It dramatically reduced the model's tendency to incorrectly label non-stressful posts as stressful.

Human-Readable Rationales

A key benefit is the model's ability to generate clear explanations for its decisions, crucial for validation by mental health professionals.

The Challenge: Adapting General AI for Specialized Medical Tasks

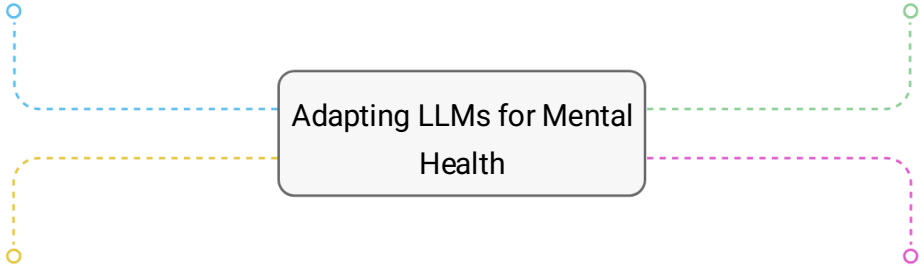
Using powerful, general-purpose Large Language Models (LLMs) for sensitive and nuanced domains like mental health presents several key challenges that hinder their direct application.

Cost of Specialization

Fine-tuning models like GPT-4 is computationally expensive and resource-intensive.

Closed-Source Models

Proprietary models like GPT-4 cannot be directly modified, limiting customization.



The Need for Explainability

Psychologists and patients need to trust model outputs, requiring transparent reasoning, not just an answer.

Lower Initial Accuracy

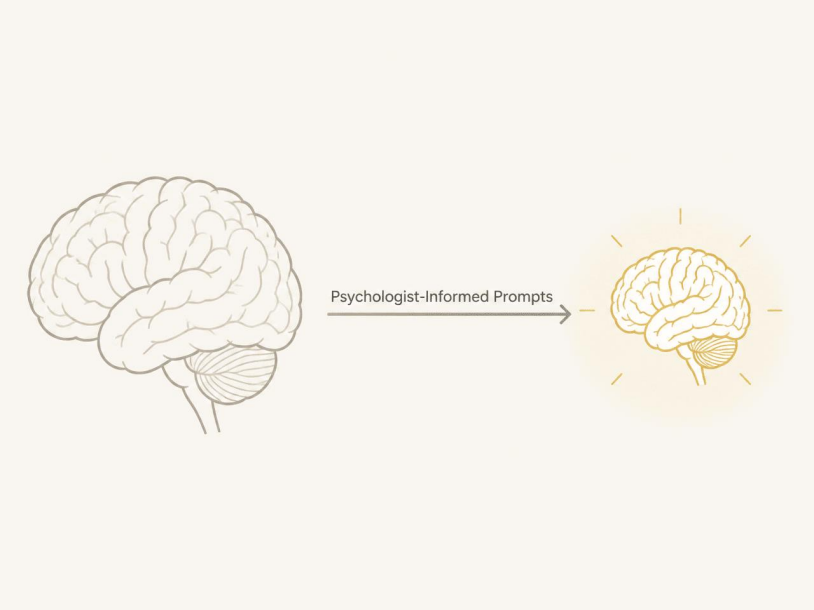
General-purpose models often overlook subtle cues that domain-specific models are trained to catch.

A Resource-Efficient Solution: Prompt Engineering

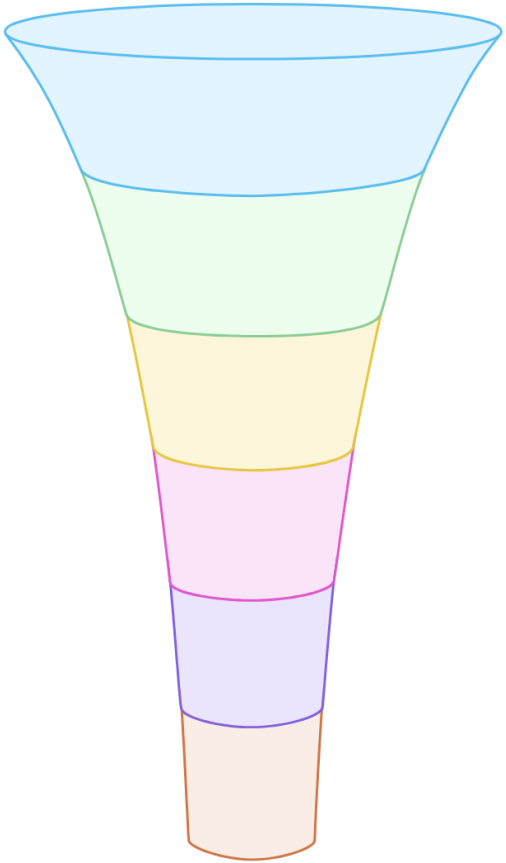
Prompt engineering offers a clever and efficient way to tailor a general-purpose LLM for a specific task without altering the model itself. By carefully crafting the instructions (the 'prompt'), we can guide the model's focus and reasoning.

The Study's Innovation

This research introduces an iterative refinement process where prompts are enhanced with 'hints' derived from psychologist analysis of the model's errors. This 'expert-in-the-loop' approach bridges the gap between general AI and specialized knowledge.



How The Study Was Conducted



Start with the Dreaddit dataset, a collection of Reddit posts labeled for stress.

Apply an initial 'zero-shot' prompt to GPT-4 to establish a baseline performance.

Analyze the model's errors with the help of psychologists to understand why they occurred.

Develop a set of 'hints' from the psychologists' insights to guide the model.

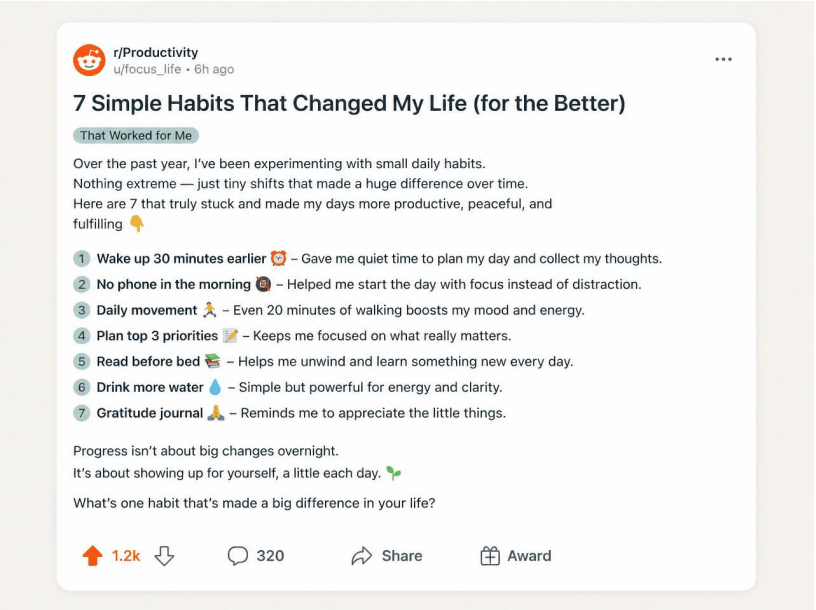
Incorporate these hints into a new, refined prompt.

Evaluate the refined prompt on a separate test set and compare results against the baseline and other specialized models.

The Foundation: The Dreaddit Dataset

The study utilized the Dreaddit dataset, a specialized resource created for stress analysis in social media. Its unique characteristics make it an ideal testbed for this research.

- Source: Reddit communities.
- Content: 3,553 annotated text segments.
- Rich & Nuanced: Posts are long-form (avg. 420 tokens), capturing complex emotional narratives and context.
- Robust Annotation: Each post was labeled by at least five annotators to ensure reliability.
- Authenticity: The data was minimally preprocessed to retain natural language, including emojis, slang, and misspellings.



The Core Method: Iterative Prompt Refinement

The methodology follows a systematic, nine-stage loop to iteratively improve the model's performance by learning from its mistakes with expert guidance.

1

Step 1: Zero-Shot Prompting

An initial, generic prompt is applied to the dataset to get a baseline performance reading.

2

Steps 2 & 3: Error Extraction & Analysis

All incorrectly classified posts (false positives and false negatives) are collected. Psychologists then analyze these errors to find their root causes.

3

Steps 4, 5, 6: Hint Development

The psychologists' insights are distilled into concise, actionable 'hints'. These hints are then ranked and formatted to be added to the prompt.

4

Steps 7, 8, 9: Refined Evaluation

The prompt, now enhanced with the expert hints, is used to re-evaluate the model on an independent test set, and the improved performance is measured.

Example of Hint

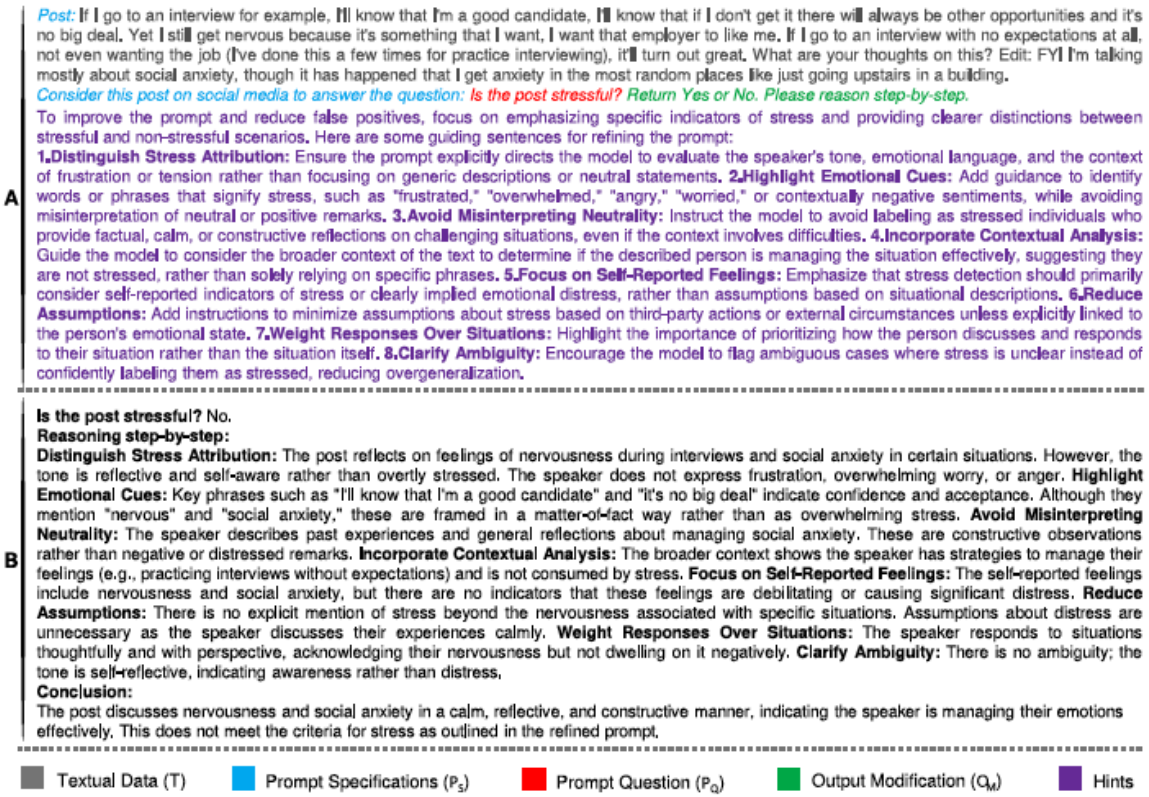


Fig. 3. Example of prompt engineering for a non-stressful text: GPT-4, with hints, correctly classified the text. Part A shows the prompt with psychologist-informed hints, and Part B presents GPT-4's response and reasoning.

From Errors to Insights: Developing the Hints

The initial error analysis revealed that GPT-4 had a strong bias towards caution, resulting in a high rate of false positives—it was seeing stress where there was none. 90% of the initial errors were of this type.

The 8 Key Hints for GPT-4

Psychologists identified eight common misinterpretation patterns and created hints to correct them. These hints instructed the model on how to better distinguish true stress from related but distinct emotional states.

Distinguish Stress Attribution

Focus on the speaker's tone and emotional language, not just neutral descriptions of difficult situations.

Highlight Emotional Cues

Identify specific words like 'frustrated' or 'overwhelmed' while avoiding misinterpretation of neutral remarks.

Incorporate Contextual Analysis

Consider if the person is managing the situation effectively, which suggests they are not stressed.

Focus on Self-Reported Feelings

Prioritize direct indicators of stress from the user, rather than making assumptions based on the situation.

Results: The Dramatic Impact of Hints

Incorporating the psychologist-informed hints led to a significant improvement across all key performance metrics, transforming GPT-4 into a highly accurate and reliable tool for stress detection.

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**

Accuracy jumped by 17%, and Precision saw a massive 21.6% increase, indicating the model became much better at correctly identifying only the truly stressful posts.

Benchmarking Against the Specialists

The enhanced GPT-4 was not only better than its baseline self; it also outperformed several domain-specific models that were explicitly fine-tuned on mental health datasets.

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

Discussion: Strengths and Limitations

Key Strengths

- Cost-Effective: A highly efficient alternative to expensive fine-tuning.
- Expert-Driven: The 'expert-in-the-loop' process is crucial for capturing nuance in sensitive domains.
- Transparent Rationales: Generates human-readable explanations, fostering trust and clinical utility.
- Model-Agnostic: The methodology can be adapted to other LLMs, including open-weight models.

Identified Limitations

- Post-Hoc Explanations: Relies on the model explaining itself, which is not true algorithmic transparency.
- Dataset Specificity: The results are based on the Reddit platform; generalizability to Twitter or Instagram is unknown.
- Potential Label Noise: The dataset relies on human annotators, which can introduce subjectivity.
- Proprietary Model Risk: Using a closed-source model like GPT-4 poses challenges for long-term reproducibility.

Conclusion & Future Directions

Conclusion

Prompt engineering, guided by domain experts, is a powerful, scalable, and transparent strategy for adapting general-purpose LLMs to specialized, sensitive tasks like mental health monitoring. It offers a practical path to leveraging advanced AI while maintaining human oversight.

Future Work

- Validate the approach on open-weight models like Llama to improve accessibility and transparency.
- Test the methodology on datasets from diverse social media platforms.
- Incorporate more advanced prompting techniques (e.g., Chain-of-Thought) to further improve reasoning.
- Develop hybrid methods to enhance true explainability beyond post-hoc rationales.

Thank You!

Email: a.mamun@asu.edu

Paper

abdullah-mamun.com

𝕏: @AB9Mamun