Evaluating Large Language Models for Analyzing Suicidal Social Media Health Narratives

Sumnan Azade & Salim Sazzed*

Georgia Southern University, Statesboro, GA, USA
*ssazzed@georgiasouthern.edu

Abstract. The proliferation of social media health platforms has led to the widespread availability of data encompassing mental health discussions and personal narratives, highlighting the need for scalable and cost-effective solutions for analyzing them. Recent advancements in large language models (LLMs) have shown promise in analyzing complex textual data at scale. This study investigates the effectiveness of advanced LLMs in extracting key information and insights from emotionally charged, real-world mental health text, with a particular focus on posts related to suicidal ideation. We assess the performance of six leading LLMs—Gemini, GPT, Claude, Llama, DeepSeek, and Mistral AI—across two distinct set of tasks: (1) keyword extraction tasks to identify mental and physical health conditions, and (2) binary classification tasks to detect references to pivotal life events, such as family abuse, substance use, and prior suicide attempts. Our evaluation highlights the moderate to strong performance of LLMs across both task sets, with Gemini demonstrating the most consistent and effective performance among the LLMs. Out of the five tasks, it outperforms the other LLMs in three, although the differences are not always statistically significant when compared to the second-best LLMs, as indicated by the McNemar and Wilcoxon Signed-Rank tests. These findings underscore the potential of LLMs in the large-scale analysis of mental health data, as well as the variability in their performance, which could guide the development of effective mental health tools and interventions.

1 Introduction

The rapid expansion of health-focused discussion forums on social media has created extensive digital archives of user-generated content, offering firsthand accounts of individuals' experiences with mental health challenges [1,2]. Platforms hosting these discussions provide real-time access to narratives surrounding issues such as suicidal ideation, depression, and self-harm, presenting valuable opportunities to understand the lived experiences of those affected [3]. However, the volume, variability, and emotional complexity of these posts pose significant challenges for manual analysis, which is often labor-intensive, inconsistent, and difficult to scale [4]. Training machine learning models from scratch demands

extensive manual annotations and high computational power to achieve satisfactory accuracy for specific tasks, making the process highly resource-intensive. These challenges underscore the need for scalable, automated solutions capable of analyzing sensitive health-related content with precision and nuance.

In recent years, large language models (LLMs) have begun offering effective solutions to address various challenges in health research [5,6,7]. In particular, LLMs can play a pivotal role in addressing the growing prevalence of social media health data, enabling large-scale analysis and generating meaningful insights. Recent studies have explored the potential of LLMs for analyzing and generating insights from mental health data, in both clinical and non-clinical settings [8,9].

In this study, we assess the capabilities of large language models (LLMs) in analyzing and generating insights from social media suicidal ideation posts. Suicide is a significant public health concern, with over 720,000 people dying by suicide each year [1]. The role of social media platforms in suicide research has garnered increasing attention, as these platforms offer a unique opportunity to detect early signs of distress, informing interventions and strategies that could ultimately save lives. To this end, we evaluate the performance of six state-of-the-art LLMs in extracting meaningful information and deriving insights from these posts. We have developed a targeted set of questions designed to test the LLMs' ability to interpret narratives containing complex, personal, and sensitive content and extract key information. These questions address two distinct types of analytical tasks:

- **Taskset 1:** Keyword extraction tasks, such as identifying mental (T1) and physical health (T2) conditions in the posts, aimed at evaluating the ability of LLMs to extract crucial information.
- **Taskset 2:** Binary classification tasks focused on identifying significant events, such as family abuse (T3), substance use (T4), and prior suicide attempts (T5), within posts. These tasks assess the semantic understanding and inference capabilities of the LLMs.

Our findings show that, across multiple tasks, Gemini delivers the most effective and consistent performance in identifying and inferring relevant information from suicidal narratives, ranking as the top classifier in 3 out of 5 tasks. Other LLMs, such as GPT and DeepSeek, follow closely in terms of consistency and effectiveness, often emerging as the second-best or top performers. However, no single model consistently outperforms the others, and differences between the top two LLMs are often statistically insignificant. Overall, for each task, at least one LLM demonstrates effective performance, underscoring the potential of LLMs to extract valuable insights from emotionally charged mental health narratives. This highlights the promise of LLMs in supporting the development of digital mental health technologies, especially considering the vast volume of health-related data available on social media.

[1] `https://www.who.int/news-room/fact-sheets/detail/suicide`

2 Corpus Creation

We leverage suicidal posts written in the Reddit discussion forum `r/SuicideWatch` [2], where individuals openly share their experiences with severe mental health challenges, including suicidal ideation and attempts. These posts offer first-person narratives that capture complex emotional and psychological experiences, providing a rich and unique data source for evaluating large language models (LLMs) in sensitive mental health contexts. The dataset is publicly available on Kaggle.

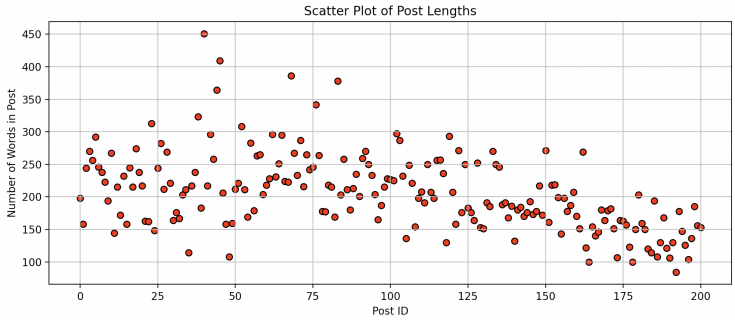

Fig. 1. Word length distribution of the 201 samples, ranging from 80 to 450 words, with a mean of 206.35, standard deviation of 59.32, and median of 203.00.

To evaluate the capabilities of LLMs, we created a corpus of 201 samples, each ranging from 100 to 450 words (see Fig 1), which were manually annotated for five tasks. These samples contain key information for tasks such as keyword extraction and classification, while maintaining a feasible word range for manual annotations as ground truth. This curated corpus enables a nuanced evaluation of LLMs' ability to process emotionally charged discourse.

3 Methodology

To assess the performance of various LLM models, a set of tasks is defined. Ground truth is established through manual annotation, and LLM outputs are generated using carefully crafted prompts.

3.1 Tasks Designed for Assessing LLM Capabilities

We designed the following two sets of tasks to assess the ability of LLMs to comprehend and extract relevant information from social media health text.

Keyword extraction tasks: These tasks involve the extraction of lexical items, such as mentions of specific mental or physical health conditions, from the given text. For example, extracting terms like "depression," "bipolar disorder" or "diabetes" mentioned in the samples.

- **Task 1:** What mental health conditions or illnesses are explicitly mentioned in the text?
- **Task 2:** What physical health conditions or illnesses are explicitly mentioned in the text?

[2] `https://www.reddit.com/r/SuicideWatch/`

Classification tasks: These tasks require the LLMs to perform binary classification, detecting the presence or absence of key issues or experiences, such as family abuse or references to past suicide attempts.

- **Task 3:** Does the text reference any family abuse or related experiences?
- **Task 4:** Does the text mention substance use (e.g., alcohol, drugs)?
- **Task 5:** Does the text reference past suicide attempts?

Each category presents distinct cognitive demands, with varying levels of comprehension and reasoning. Keyword extraction tasks focus on lexical identification, while classification tasks require higher-order decision-making by synthesizing contextual information to identify the presence or absence of complex mental health issues.

3.2 Manual Annotation Guidelines

Ground truth for each task is established through rigorous manual annotation by an annotator, followed by expert validation. The annotator is instructed to adhere to the following detailed guidelines:

General instructions for all tasks:

- **Thorough comprehension:** Read each text sample completely before annotating.
- **Precision:** Annotations must be as specific and accurate as possible, in strict adherence to the provided guidelines.

Task 1. Guidelines for mental health conditions

- Identify and list recognized mental health conditions.
- Focus on diagnostic terms (e.g., "bipolar disorder," "major depressive disorder").
- Exclude vague terms like "sadness" or "stress"
- **Format:** Comma-separated list. If none, state "None."

Task 2. Guidelines for physical health conditions

- Identify and list physical health conditions.
- Include physical health conditions (e.g., "hepatitis", "migraine ").
- Exclude primarily mental health conditions, even if physical symptoms are present.
- **Format:** Comma-separated list. If none, state "None."

Task 3. Guidelines for family abuse

- Determine if the text mentions family abuse.
- Identify implicit and explicit references to physical, emotional, sexual abuse, or abusive family dynamics.
- **Format:** "Yes" or "No".

Task 4. Guidelines for labeling substance use

- Identify references to substance use (e.g., alcohol, recreational drugs, prescription drug misuse).
- Look for explicit or implicit references to any form of substance use.
- **Format:** "Yes" or "No."

Task 5. Guidelines for past suicidal attempts

- Determine if the text includes references to past suicidal attempts.
- Annotate implicit or explicit mentions of past attempts (self-injurious acts with intent to die).
- **Format:** "Yes" or "No."

Expert Validation: The annotated samples were reviewed and validated by an expert with over 7 years of experience in Natural Language Processing (NLP) and text annotation. The inter-annotator agreement for the five tasks ranged from 0.85 to 0.95, demonstrating a high level of consistency and reliability across the tasks.

3.3 Prompts Used for LLMs Across Different Tasks

Keyword extraction task:

- **Task 1:** Mental health conditions or illnesses are explicitly mentioned in the text.
  Prompt for LLM: "Please identify and list only recognized mental health conditions or disorders explicitly mentioned in the following text. Exclude any emotional states or subjective terms. Provide the response as a comma-separated list of terms referring to mental health conditions or disorders."
- **Task 2:** Physical health conditions or illnesses are explicitly mentioned in the text.
  Prompt for LLM: "Please identify and list only the physical health conditions or illnesses explicitly mentioned in the following text. Provide the response as a comma-separated list of relevant terms referring to physical conditions or illnesses."

Classification tasks:

- **Task 3:** Mentions of any family abuse or related experiences.
  Prompt for LLM: "Does the individual describe any experiences related to family abuse or mention abuse within a family context in the given text? Respond with 'Yes' or 'No'."
- **Task 4:** Mentions of substance use (e.g., alcohol, drugs).
  Prompt for LLM: "Does the given text contain any reference to substance use (such as alcohol, drugs, prescription drugs)? Respond with 'Yes' or 'No'."
- **Task 5:** Mentions of past suicidal attempts.
  Prompt for LLM: "Are there references to previous suicide attempts in the given text? Respond with 'Yes' or 'No'."

3.4 Selection of Large Language Models (LLMs)

We employed six state-of-the-art large language models (LLMs)—specific versions listed below—accessed via public APIs with default settings, to extract and infer key insights from social media discussions on mental health. We selected these models for their demonstrated effectiveness in various real-world NLP applications [10], with our choice of specific versions further guided by their free public availability, ease of use, and documented performance.

OpenAI's *GPT-4.1-mini*, launched on April 14, 2025, is a compact yet powerful model within the GPT-4.1 series, offering a strong balance of intelligence,

speed, and cost-efficiency [11]. Google *Gemini 2.0 Flash-001*, released on February, 2025, is a fast and efficient model in the Gemini 2.0 family, optimized for speed and cost-effectiveness [12]. Anthropic's *Claude 3 Haiku*, released on March 4, 2024, is the fastest model in the Claude 3 series, excelling in real-time tasks like customer support and content moderation [13].Mistral AI's *Mistral-Small-3.2-24B-Instruct*, updated on June 20, 2025, is a versatile model designed for practical, real-world tasks [3]. Meta's *llama-3.3-70b-instruct* is an instruction-tuned model enhancing performance across a wide range of NLP tasks [4]. Deepseek's *deepseek-chat-v3-0324* is a specialized conversational model optimized for contextually complex dialogues [5].

4 Results and Discussion

4.1 Evaluation

We evaluated the performance of large language models (LLMs) using established quantitative metrics to rigorously assess their effectiveness in keyword extraction and classification tasks.

Keyword Extraction Tasks: Keywords determined through manual annotations for each sample serve as the ground truth. The LLM outputs are standardized to ensure they follow the same format as the manual annotations. In this process, phrases such as "Borderline Personality Disorder" and "BPD" are treated as equivalent, as they refer to the same condition. Additionally, all keywords are lowercased and whitespace removed for comparison.

For the keyword extraction task, true positives (TP), false negatives (FN), and false positives (FP) are computed and subsequently used to calculate precision (P), recall (R), and F1 score.

Precision (Pre.): Proportion of correctly identified keywords among all keywords predicted (TP / (TP + FP)).

Recall (Rec.): Proportion of relevant keywords detected out of all ground truth terms (TP / (TP + FN)).

F1 Score: Harmonic mean of precision and recall.

Precision (Pre.), Recall (Rec.), F1-score, and Accuracy (Acc.) are calculated for each sample, and the average values (Avg.) across the entire corpus are subsequently reported.

Classification Tasks: For binary classification tasks, such as identifying the presence of family abuse or substance use, model performance is evaluated using Precision (Pre), Recall (Rec), and F1-score (macro, MR; weighted, WG) for both classes (Class 1 and Class 2). Manual annotation reveals that the class distributions across the three tasks are significantly skewed toward the *No* class.

1. Family abuse: 13 *Yes* and 61 *No*.
2. Substance abuse: 13 *Yes* and 61 *No*.
3. Previous suicide attempts: 14 *Yes* and 60 *No*.

[3] `https://docs.mistral.ai`
[4] `https://github.com/meta-llama`
[5] `https://arxiv.org/abs/2401.02954`

4.2 Predictions by LLMs on a Sample

The following sample illustrates significant mental health challenges experienced by an individual. Table 1 presents the outputs from six LLMs across five tasks for this sample.

> *I've been suicidal since the age of 8 im now 19 and nothing has ever helped you could give me the most money in the world and id still want to kill myself, I dont want to be here anymore im done with my life i have no purpose to be here i hate my life and everyone around me im miserable. Ive never had a easy upbringing growing up with a bipolar mother being bullied, struggled with bad mental health, being threatened to put into care if i dont stop being suicidal. i've attempted suicide 3 times all those times i got caught. I have BDD, ADHD, ASBERGERS and depression i can't live like this anymore, i was promised help by my doctor but been waiting over a year and still no letter about me getting help i feel hopeless. I dont see any purpose being here the only reason im alive is im scared it will fail. Id try go missing and never get found but i dont want for people to see what i looked like. Im sick of life :(I dont understand why its so bad to kill urself? people will get over it nothing would change the world would still be the same. My mom always says shes sick of me and my depressive thoughts so why not?*

Table 1. Outputs from LLMs for the above Sample

Task	**GPT**	**Gemini**	**Claude**	**Mistralal**	**Llama**	**DeepSeek**
Mental Health Keyword	Bipolar disorder, BDD, ADHD, ASBERGERS ,depression	Bipolar disorder, BDD, ADHD, ASBERGERS ,depression	bipolar disorder, BDD, ADHD, ASBERGERS ,depression	Bipolar disorder, BDD, ADHD, ASBERGERS ,depression	BDD, ADHD, ASBERGERS ,depression	BDD, ADHD, ASBERGERS ,depression
Physical Health Keyword	None	None	Bipolar disorder, depression	None	depression	depression
Family Abuse	Yes	Yes	Yes	No	Yes	Yes
Substance Use	No	No	No	No	No	No
Past Suicide	Yes	Yes	Yes	Yes	Yes	Yes

4.3 Comparative Analysis and Statistical Tests

Table 2 compares the performance of six large language models (LLMs) in extracting mental and physical health-related keywords, using precision, recall, F1-score, and accuracy as evaluation metrics.

Gemini performs best across both taks. In mental health, it achieves highest F1-score (0.81) and accuracy (0.80), indicating strong and reliable keyword extraction. For physical health, its performance is a bit lower, but it still remains the best among all LLMs. GPT, DeepSeek, and Mistral exhibit similar performance, particularly in the keyword extraction task related to mental health. Llama demonstrates the lowest level of effectiveness across both tasks.

Wilcoxon Signed-Rank Test was performed to assess whether Gemini demonstrates a statistically significant improvement over the second-best performing LLM across both tasks. For mental health keyword extraction, the difference with the second-best performing DeepSeek yields a p-value of 0.0094, indicating a statistically significant difference. In contrast, for physical health, no significant difference (p-value: 0.2597) is observed between Gemini and Claude.

Table 2. Performance Comparison of LLMs in Extracting Mental and Physical Health Keywords

Model	Mental Health				Physical Health			
	Avg. Pre.	**Avg. Rec.**	**Avg. F1**	**Avg. Acc.**	**Avg. Pre.**	**Avg.Rec.**	**Avg. F1**	**Avg. Acc.**
GPT	0.71	0.77	0.73	0.7	0.57	0.61	0.58	0.56
Gemini	0.81	0.82	**0.81**	**0.8**	0.65	0.68	**0.66**	**0.64**
Claude	0.6	0.66	0.62	0.59	0.6	0.63	0.61	0.59
Mistral	0.71	0.76	0.73	0.71	0.59	0.62	0.6	0.58
Llama	0.64	0.72	0.67	0.64	0.43	0.49	0.45	0.43
DeepSeek	0.72	0.77	0.74	0.72	0.55	0.59	0.56	0.54

Table 3. Performance Comparison of Large Language Models in three classification tasks

Model	Substance Use			Past Suicidal Attempt			Family Abuse		
	Pre. Cls:1,2	**Rec. Cls:1,2**	**F1 MR/WG**	**Pre. Cls:1,2**	**Recall Cls:1,2**	**F1 MR/WG**	**Pre. Cls:1,2**	**Rec. Cls:1,2**	**F1 MR/WG**
GPT	[0.96,0.38]	[0.79,0.82]	0.69/0.82	[0.97,0.54]	[0.86,0.85]	**0.78/0.87**	[0.97,0.69]	[0.91,0.87]	**0.85/0.9**
Gemini	[0.96,0.36]	[0.76,0.82]	0.68/0.8	[0.96,0.42]	[0.77,0.85]	0.71/0.81	[0.97,0.68]	[0.9,0.89]	**0.85/0.91**
Claude	[0.93,0.2]	[0.5,0.79]	0.49/0.6	[0.95,0.33]	[0.68,0.82]	0.63/0.74	[0.95,0.46]	[0.77,0.82]	0.72/0.80
Mistral	[0.94,0.36]	[0.79,0.71]	0.67/0.81	[0.96,0.44]	[0.8,0.82]	0.72/0.82	[0.95,0.61]	[0.88,0.79]	0.80/0.87
Llama	[0.99,0.41]	[0.77,0.96]	**0.72/0.83**	[1.0, 0.28]	[0.5, 1.0]	0.55/0.63	[0.99,0.53]	[0.8,0.97]	0.78/0.85
DeepSeek	[0.99,0.41]	[0.79,0.93]	**0.72/0.83**	[1.0,0.37]	[0.66,1.0]	0.67/0.75	[0.98,0.56]	[0.83,0.92]	0.8/0.86

Table 3 compares the performance of six large language models (LLMs) in classifying Substance Use, Past Suicidal Attempts, and Family Abuse. Gemini and GPT achieve the highest macro F1 score for Family Abuse (0.85). GPT performs best in classifying past suicidal attempts (F1 = 0.78), while Llama and DeepSeek outperform others in the substance use classification task. Mistral performs consistently well across all three tasks, though it does not achieve the highest score in any of them. Claude is the lowest-performing LLM across all three tasks.

McNemar's test was employed to compare the performance of the top-performing LLM against the second-best (or, in the case of a tie, the top LLM against the third-best) to evaluate whether significant differences existed. For the substance use task, the p-value between Llama and GPT was 0.823, for the past suicidal attempt task, it was 0.072 between GPT and Mistral, and for the family abuse task, it was 0.210 between Gemini and Mistral—all non-significant. These results indicate no statistically significant differences between the top LLMs, suggesting that none significantly outperform their closest competitors.

4.4 Findings and Implications

Gemini is found to be the most balanced and effective model for both mental and physical health keyword extraction tasks. However, all the LLMs, including Gemini, perform below average for the physical health keyword extraction task. Our investigation suggests that this is often due to a failure to distinguish between mental and physical health, with models mistakenly categorizing mental health as physical health. Additionally, they sometimes infer health conditions from the text, even though the prompt explicitly instructs them to check for explicit mentions.

In classification tasks, the performance of the LLMs is less consistent across tasks. For example, although Llama shows the best performance along with

DeepSeek for the substance use task, it performs poorly for the past suicide task and moderately for the family abuse task. The same applies to DeepSeek: it performs best in the substance use task but underperforms for the past suicide attempt task. Similar to keyword extraction tasks, Claude is the lowest-performing LLM across all classification tasks. Among all the LLMs, Gemini shows consistent performance, providing the highest results for two tasks and ranking among the top three in the substance use task.

The results provide a default performance of LLMs on emotionally charged health-related text. However, as observed, there is no clear winner that consistently outperforms the others. Overall, Gemini emerges as the most consistent and reliable model for health text analysis under the default settings. Other LLMs, such as Llama, show strong performance on specific tasks but do not consistently lead across all tasks. Statistical significance tests, including the Wilcoxon Signed-Rank Test for keyword extraction tasks and the McNemar test for classification tasks, reveal that, in most cases, the difference between the top-performing LLM and the second-best is not statistically significant. This further suggests that no single LLM is the definitive winner across all tasks; rather, performance varies depending on the specific task. For each task, at least one LLM shows good or acceptable performance, highlighting the potential of LLMs. Note that, for this study, our goal was to evaluate LLM performance under default settings without any fine-tuning. Fine-tuning with labeled data is likely to improve performance, and we plan to explore this further in future work.

Observations on LLM Behavior: We observed a few issues while generating outputs from the LLMs:

- **Prompt adherence:** The models often fail to fully adhere to the prompt, particularly in keyword extraction tasks (e.g., identifying specific health conditions). When asked to identify explicitly mentioned conditions, the models occasionally infer conditions or provide symptoms instead of directly extracting the specified health conditions.
- **Model version variability:** The results may vary depending on the model version used. Additionally, web-based versions can yield slightly different outcomes due to factors such as caching or the history of prior interactions.

5 Summary and Conclusion

This study evaluates the capabilities of six advanced large language models (LLMs) in extracting key insights from social media narratives related to suicidal ideation, which could facilitate the development of effective mental health tools and support the growing field of digital mental health research. Using a task-oriented framework, we assessed each model's ability to extract relevant keywords and classify sensitive phenomena, such as family abuse, substance use, and prior suicide attempts.

Our findings indicate that LLMs, particularly Gemini, perform high level of consistency and effectiveness in both keyword extraction tasks (identifying key mental and physical health conditions) and classification tasks, closely followed by GPT and DeepSeek. The results also suggest that LLM performance often varies depending on the tasks considered, and that, in many cases, no significant

difference is observed among the top-performing LLMs. These results highlight the significant potential of LLMs for scalable and timely analysis of large-scale social media health data, providing a feasible alternative to manual methods and underscoring the promise of LLMs as essential tools for developing intelligent, automated systems for mental health monitoring and intervention. This study focuses on a limited set of tasks using a small dataset, constrained by the 10-page restriction. Future research will build on these findings by developing more complex tasks and utilizing a larger dataset, along with a broader array of large language models (LLMs), to conduct more comprehensive evaluations. Furthermore, we will fine-tune LLMs with annotated data to evaluate the extent to which this enhances performance.

Ethical considerations: This study ensures the exclusion of all identifiable information, in accordance with ethical research standards.

References

1. John A Naslund, Ameya Bondre, John Torous, and Kelly A Aschbrenner. Social media and mental health: benefits, risks, and opportunities for research and practice. *Journal of technology in behavioral science*, 5:245–257, 2020.
2. Tatsawan Timakum, Qing Xie, and Soobin Lee. Identifying mental health discussion topic in social media community: subreddit of bipolar disorder analysis. *Frontiers in Research Metrics and Analytics*, 8:1243407, 2023.
3. Iram Fatima, Hamid Mukhtar, Hafiz Farooq Ahmad, and Kashif Rajpoot. Analysis of user-generated content from online social communities to characterise and predict depression degree. *Journal of Information Science*, 44(5):683–695, 2018.
4. Muskan Garg. Mental health analysis in social media posts: a survey. *Archives of Computational Methods in Engineering*, 30(3):1819–1842, 2023.
5. Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R Pfohl, Heather Cole-Lewis, et al. Toward expert-level medical question answering with large language models. *Nature Medicine*, pages 1–8, 2025.
6. Cheng Peng, Xi Yang, Aokun Chen, Kaleb E Smith, Nima PourNejatian, Anthony B Costa, Cheryl Martin, Mona G Flores, Ying Zhang, Tanja Magoc, et al. A study of generative large language model for medical research and healthcare. *NPJ digital medicine*, 6(1):210, 2023.
7. Anmol Arora and Ananya Arora. The promise of large language models in health care. *The Lancet*, 401(10377):641, 2023.
8. Xuhai Xu, Bingsheng Yao, Yuanzhe Dong, Saadia Gabriel, Hong Yu, James Hendler, Marzyeh Ghassemi, Anind K Dey, and Dakuo Wang. 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, 2024.
9. Kailai Yang, Tianlin Zhang, Ziyan Kuang, Qianqian Xie, Jimin Huang, and Sophia Ananiadou. Mentallama: interpretable mental health analysis on social media with large language models. In *Proceedings of the ACM Web Conference 2024*, pages 4489–4500, 2024.
10. Changkun Mao, Yuanyuan Bao, Yuanyuan Yang, and Yongsheng Cao. Application of chatgpt in pediatric surgery: opportunities and challenges. *International Journal of Surgery*, 110(5):2513–2514, 2024.
11. OpenAI. GPT-4 Technical Report. 2023.
12. Google DeepMind. Gemini 2.0 Flash-001 Technical Report. 2025.
13. Anthropic. Claude 3 Model Family: Opus, Sonnet, Haiku. 2024.