Large Language Models for Automatic Detection of Sensitive Topics

RUOYU WEN*, University of Canterbury, New Zealand
STEPHANIE ELENA CROWE†, University of Canterbury, New Zealand
KUNAL GUPTA†, University of Auckland, New Zealand
XINYUE LI‡, Uppsala University, Sweden
MARK BILLINGHURST, The University of Auckland, New Zealand
SIMON HOERMANN, University of Canterbury, New Zealand
D. D. ALLAN, University of Canterbury, New Zealand
ALAEDDIN NASSANI, University of Aizu, Japan
THAMMATHIP PIUMSOMBOON, University of Canterbury, New Zealand

Sensitive information detection is crucial in content moderation to maintain safe online communities. Assisting in this traditionally manual process could relieve human moderators from overwhelming and tedious tasks, allowing them to focus solely on flagged content that may pose potential risks. Rapidly advancing large language models (LLMs) are known for their capability to understand and process natural language and so present a potential solution to support this process. This study explores the capabilities of five LLMs for detecting sensitive messages in the mental well-being domain within two online datasets and assesses their performance in terms of accuracy, precision, recall, F1 scores, and consistency. Our findings indicate that LLMs have the potential to be integrated into the moderation workflow as a convenient and precise detection tool. The best-performing model, GPT-4o, achieved an average accuracy of 99.5% and an F1-score of 0.99. We discuss the advantages and potential challenges of using LLMs in the moderation workflow and suggest that future research should address the ethical considerations of utilising this technology.

CCS Concepts: • **Human-centered computing → Collaborative and social computing**.

Additional Key Words and Phrases: Online Moderation, Automatic Detection, Large Language Model, Comparative Study

ACM Reference Format:
Ruoyu Wen, Stephanie Elena Crowe, Kunal Gupta, Xinyue Li, Mark Billinghurst, Simon Hoermann, D. D. Allan, Alaeddin Nassani, and Thammathip Piumsomboon. 2018. Large Language Models for Automatic Detection of Sensitive Topics. In . ACM, New York, NY, USA, 22 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

This paper explores how Large Language Models (LLMs) can be utilised to identify sensitive information related to mental well-being in the context of human communications. The recent development of LLMs has led to their integration into many traditional workflows. These models are recognised for their capability to comprehend and process natural languages effectively. Given their advanced understanding and generation capabilities, LLMs offer a potential approach to automating the detection of sensitive topics such as suicidal ideation or self-harm, grief and loss, sexual assault, and abuse. This paper evaluates the capability of LLMs to detect such topics. By

Conference acronym 'XX, June 03–05, 2018, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/XXXXXXX.XXXXXXX

examining the accuracy and consistency of different LLMs, this study aims to provide insights into their potential to identify and process sensitive information effectively.

One application area for this is content detection and moderation in mental health and the use of conversational agents and chatbots to provide mental health support safely and ethically. The prevalence of mental health issues is escalating globally, presenting a substantial public health challenge. The World Health Organisation (WHO) estimates that one in four people will encounter mental health difficulties at some point in their lives [44]. Youth (ages 12–24) is identified as a critical period, with the majority of mental disorders commencing during these formative years [45]. The protracted COVID-19 pandemic appears to further exacerbate these trends [47, 51]. However, globally, healthcare systems face a pronounced shortage of trained mental health professionals, making it a significant challenge for young people to access timely mental support [9, 30].

Mental health chatbots are increasingly recognised as a promising and impactful tool for mental health professionals within the context of mental health support [18]. Recent studies have demonstrated their ability to assist in managing mental health issues [1, 10, 54, 56]. There are a number of advantages to using them; Chatbots can be accessed 24 hours a day in multiple languages on various health topics, reducing labour costs and improving counselling efficiency. However, prior studies have indicated that chatbots have limitations in ensuring safety within the mental well-being domain [6, 18, 33]. One disadvantage is that chatbots often lack the ability to process sensitive information, which can sometimes harm the user's mental well-being instead of supporting it. Consequently, human-in-the-loop monitoring is still crucial in human-AI interactions in real time.

Chatbots can show a progressive ability to understand the context in conversations [58], indicating a potential to support humans with sensitive topic detection. LLMs can help detect sensitive information by labelling sensitive messages and notifying a human moderator to intervene. This study aims to investigate the capabilities of LLMs to detect sensitive topics within the mental well-being domain. Our aim is to test the performance of various LLMs on metrics such as accuracy, precision, and recall in automatically detecting sensitive messages. By conducting this study, we seek to answer two research questions:

RQ1: What is the impact of different LLMs on detecting sensitive topics?
RQ2: How does the temperature parameter setting affect LLMs' performance in detecting sensitive topics?

The rest of the paper is organised as follows. Section 2 introduces related work in this field, and Section 3 explains the methodology we followed. While Section 4 introduces the preliminary study, Section 5 covers the procedure of the formal and the results are presented in Section 6. Finally, discussions and conclusions are presented in Sections 7 and 8, respectively. The main contributions of this paper are:

(1) We provide one of the first evaluations of the capability of several common LLMs at different temperature settings in detecting mentally sensitive messages.
(2) We curated datasets to form a collection of sensitive and non-sensitive messages.
(3) We provided a framework for testing the capability of LLMs in detecting sensitive content during the moderation process.

2 RELATED WORK

Automated content detection is a critical area of research, particularly for identifying sensitive and harmful information on online platforms. This section explores the various methods and technologies employed in this domain, highlighting the evolution from traditional human moderation to advanced machine-learning approaches. Subsection 2.1 explores the application of automated content detection in different contexts, such as preventing cyberbullying and supporting mental health. Subsection 2.2 discusses the perceptions and implications of technology-assisted detection, emphasising the balance between automated tools and human oversight. Finally, Subsection 2.3 introduces LLMs as a new method, showcasing their potential to revolutionise automated content detection with their powerful text generation and understanding capabilities.

2.1 Automated Content Detection

Sensitive and harmful information detection is always an important topic in many areas. Many online forums and social media platforms traditionally employ human moderators to review and manage content manually. This approach is often inefficient. To improve efficiency, some crowd-sourced workers have been introduced in this field [7, 25], but this approach, at the same time, has increased costs. Recently, many studies have explored using automated content detection, often machine-learning-based, to assist human moderators, aiming to replace manual review methods with higher efficiency and lower cost.

Automated content detection has been extensively researched in the context of preventing cyberbullying and cybercrime [48]. For example, Al-Garadi et al. [3] developed a machine learning model to detect cyberbullying on Twitter, analysing network, activity, user, and content features from 2.5 million geo-tagged tweets. They achieved high detection accuracy and precision using classifiers like Random Forest, proposing a tool for monitoring and mitigating online bullying. Similarly, Zhao et al. [57] proposed a new model that enhances traditional text classification by integrating bullying-specific features derived from word embeddings. Their model showed superior performance compared to baseline methods. Dadvar and De Jong [14] integrate user characteristics and behaviours across social networks into a new model to detect cyberbullying. Their study offers a tool for more accurate cyberbullying detection and enhanced support for victims and moderators.

One important domain for identifying sensitive information is E-health, which is a sensitive area that requires careful handling. Huh, et al. [27] developed a text classification system to assist moderators in identifying posts that need their expertise in online health communities. They used a binary Naive Bayes classifier to categorise posts into those needing a moderator's response and those that do not. This approach provides a low-cost, scalable solution to help moderators efficiently prioritise posts requiring their attention, thereby enhancing support in large-scale health forums.

There are also some studies about suicidality detection. O'Dea et al. [41] developed a machine-learning model to detect suicidality on Twitter. They collected tweets containing suicide-related phrases and used human coders to classify the level of concern. These tweets were then used to train a classifier. The classifier achieved an overall accuracy of 76%, correctly identifying 80% of "strongly concerning" tweets. Cohan et al. [12] also used machine learning to build a system to automatically triage the severity of posts in online mental health forums. The system categorises posts into four levels of severity, from no risk to imminent risk of self-harm, using a feature-rich classification framework that includes lexical, psycholinguistic, contextual, and topic modelling features. The system significantly improved the identification of high-risk posts, achieving up to a 17% improvement in F1 scores, an indicator of precision and recall capability, compared to previous methods.

Researchers also looked at peer support, which is beneficial for young people to overcome barriers in help-seeking behaviour. Milne et al. [36] developed a machine learning system to triage posts in online peer support forums, categorising messages by urgency. Using a classifier, it achieved 84% accuracy, improving moderator efficiency and response times on online forums. Milne et al. [37] organised the "CLPsych 2016 Shared Task" to develop systems for automatically triaging posts in online mental health forums, specifically "ReachOut.com". Participants classified posts into four categories: Green (no immediate action needed), Amber (moderate concern), Red (high concern), and Crisis (immediate action needed). Using various machine learning techniques, the best-performing systems significantly outperformed baselines.

2.2 Perceptions of Technology-Assisted Detection

From prior studies, we found that machine learning is the most common method for building AI-supported detectors or moderators. It can achieve high accuracy and reduce detection time. Some have already been integrated into practical workflows. Jhaver et al. [28] conducted a study of Automoderator (Automod) on the Reddit platform. They interviewed 16 Reddit moderators, including the creator of Automod, to explore the benefits and challenges associated with automated content detection technologies. The result indicated that Automod significantly reduces the workload of human moderators by handling menial tasks and protecting them from emotional labour. However, using Automod also creates new problems for human moderators. They must check for incorrect decisions, as Automod lacks contextual sensitivity and struggles with cultural and linguistic nuances. Based on the forum regulations, the tool requires regular updates and fine-tuning, adding to the workload of already limited skilled moderators. Additionally, moderators may find it difficult to fully understand and control Automod's actions, especially when it makes mistakes or its decisions are unclear, it's not fully transparent. At last, the paper highlighted the need for a balanced approach that combines automated and human moderation to maintain effective content regulation.

Many scientists have raised concerns about using technology-assisted tools for content detection [39, 46, 50]. Gillespie [21] mentioned the inability of these tools to grasp context and nuance, leading to errors in detecting sarcasm and cultural subtleties. Additionally, Gillespie is concerned about the risk of AI perpetuating biases from training data, potentially impacting marginalised groups unfairly. He emphasised the need for human oversight in moderation to handle complexities and ethical considerations, arguing that AI should assist rather than replace human moderators. He pointed out that "*the kind of judgment that includes the power to ban a user from a platform should only be made by humans.*" Gorwa et al. [23] examined the technical and political challenges of algorithmic content moderation on major platforms like Facebook, YouTube, and Twitter. Their study underscored the opaqueness and accountability issues of automated moderation and similarly emphasised the need for human oversight to address the complex and contextual nature of content moderation effectively.

2.3 Large Language Model: A New Method

The rapid development of LLMs introduces a new method in this domain, changing the way humans interact with machines [2]. LLMs possess powerful text-generating and understanding capabilities [55]. Users can ask AI to explain each decision in natural language simply by using natural language prompts, making communication straightforward. In such cases, integrating LLMs to replace traditional machine learning methods becomes a viable choice for automated content detection.

Wang et al. [53] evaluated six popular LLMs for detecting and handling harmful instructions. LLaMA-2 showed the best safety performance, while ChatGLM2 was the least effective. Commercial models like GPT-4 often rejected harmful instructions directly. He et al. [24] compared GPT-4's data labelling with 415 crowd-sourced workers from the MTurk platform on 3,177 sentence segments from 200 scholarly papers. GPT-4 achieved 83.6% accuracy, while MTurk reached 81.5%. Combining their labels improved accuracy to 87.5%. The study shows that integrating human and AI labelling enhances accuracy, highlighting the value of combining crowdsourcing with advanced AI models. Kolla et al. [31] evaluated GPT-3.5's effectiveness in moderating Reddit content, analysing 744 posts across nine subreddits. They found GPT-3.5 had a high true-negative rate (92.3%) but a low

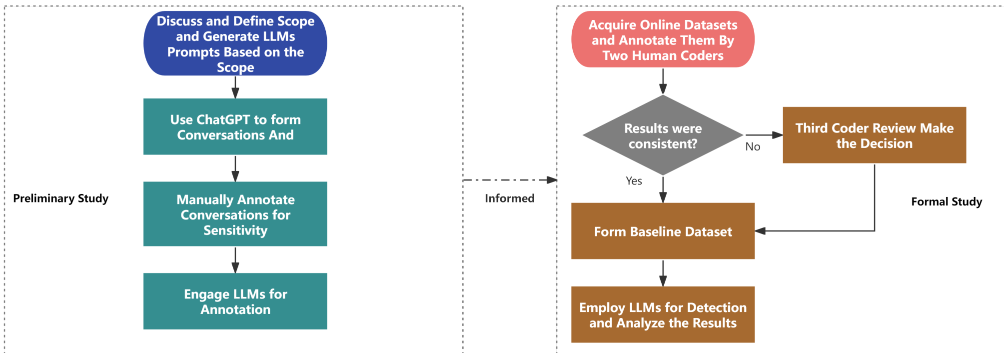

Fig. 1. The research procedure consists of two phases: the preliminary study and the formal study.

true-positive rate (43.1%), indicating it struggled to accurately flag rule violations. While it handled keyword-based violations well, it missed more complex cases. The study suggests a hybrid approach, combining GPT-3.5 with human moderators, is necessary for effective content moderation. This research investigates five models in our formal study, including GPT-4, GPT-4o, GPT-4-Turbo, Llama-3, and Solar.

3 METHODOLOGY

This research comprises two phases, as shown in 1. First, we conducted a preliminary study, detailed in Section 4. This involved using a small dataset to evaluate various LLMs for detecting sensitive topics in conversations among university freshmen. This initial study aimed to gain a general understanding of LLM performance in detecting sensitive topics. It covers dataset preparation, LLM selection and fine-tuning, and data processing. Preliminary results highlight the effectiveness of LLMs, with some models demonstrating high accuracy in identifying sensitive information.

Second, in Section 5, we conducted a formal study by evaluating five LLMs on a task involving datasets with both sensitive and non-sensitive messages. The procedures for both studies are shown in Fig. 1. Section 6 reports on the accuracy, precision, recall, F1 scores, and consistency of the models. Additionally, we tested the impact of the temperature parameter on LLM performance and analysed the causes of errors encountered during the study. Finally, Section 7 discusses the study's results.

4 PRELIMINARY STUDY

To test the capability of models for the subsequent experimental design, we conducted a preliminary study. In this study, several LLMs were tested to detect whether the topics of conversation were sensitive. The context was set as conversations among university freshmen undergoing a transition. The scope of sensitive topics was based on prior studies [16, 49] and discussions among the research team. These topics included: 'Strong Emotion or Stress,' 'Suicidal Ideation or Self-Harm,' 'Sexual Assault and Abuse,' 'Severe Mental Health Disorders,' 'Substance Abuse,' 'Eating Disorders,' 'Trauma and PTSD,' 'Grief and Loss,' 'Family Issues,' 'Identity and Discrimination Issues,' 'Legal Issues and Crime,' 'Physical Hurt' and 'General Stress and Bad Mood.'

4.1 Procedure

4.1.1 Dataset Preparation. We used ChatGPT (GPT-4) to generate casual conversations between a university freshman and her best friend about her first-year life in school. the prompt was:

"*I would like you to play as a freshman named Lily, communicating with your friend Sam about the problems you encounter in your first-year uni life. Lily is a freshman facing transition issues. I don't want continuous dialogue, but ten discontinuous pieces of conversation. I want this communication to start from the beginning like you might say "Hi, how is it going?" (Please use the format: Lily: '...', Sam: '....'.) Some of these are about Lily's troubles with stress in school, and others are casual talk. Generate 500 messages, I want half of the conversations to be mentally safe, and half to be unsafe.*"

From this generation, we obtained 500 messages(50% sensitive) to create the pool. Then, a research team member manually checked messages and selected 100 sensitive messages along with 100 non-sensitive messages to form the dataset. Here are some examples of sensitive messages:

(1) **Message:**`I think I need to talk to someone professional. I'm overwhelmed by my emotions and it's getting harder to cope.`
Reason: `General Stress and Bad Mood`
(2) **Message:**`I know it's bad, but I feel like I can't stop. It's the only thing that helps me feel less anxious.`
Reason: `Substance Abuse`

4.1.2 LLM Preparation. We selected five popular and powerful LLMs available at the time of this research for testing, including GPT-3.5-Turbo-0125, GPT-4-Turbo Claude-3-Opus-20240229, Llama-3-70B-8192, and GPT-3.5-Turbo-0125 (Fine-Tuned). Fine-tuning is a process of pre-training an LLM for specific tasks, believed to enhance its performance in specific areas [52]. In this study, we choose GPT-3.5-Turbo-0125, the most advanced GPT model for fine-tuning at the time of this research. We selected 70 messages as the training data and 30 as the validation data suggested by OpenAI [42]. Then, we used the OpenAI Playground [1] to train the fine-tuned model.

4.1.3 LLM Data Processing. We used JSON mode in Python 3.9.13 to call the API of each of the five models to batch-process the datasets. The processed data was then stored in CSV files. The following prompt was given to the LLMs as instructions:

"You are a detector that determines whether the input falls within a given range. The range includes Strong Emotion or Stress, Suicidal Ideation or Self-Harm, Sexual Assault and Abuse, Severe Mental Health Disorders, Substance Abuse, Eating Disorders, Trauma and PTSD, Grief and Loss, Family Issues, Identity and Discrimination Issues, Legal Issues, Crime. And something that can cause stress and bad mood, Long sentences of more than 30 words, and Physical hurt or related topics. Simply respond with 'yes' if it is related, or 'no' if it is not."

We set the prompt slightly broader than our scope, assuming that LLMs might struggle with long, complex sentences [29].

4.2 Preliminary Results

The results are shown in Table 1. True Positives (T+) represent the number of correctly identified sensitive data points. True Negatives (T-) represent the number of correctly identified non-sensitive data points. False Positives (F+) represent the number of non-sensitive data points incorrectly identified as sensitive. False Negatives (F-) represent the number of sensitive data points incorrectly identified as non-sensitive.

In this study, the evaluations of multiple LLMs suggested the promising potential of using LLMs in detecting sensitive topics as defined. Models like Claude-3-Opus-20240229, GPT-4-Turbo, and

[1]https://platform.openai.com/playground

Table 1. Comparison of Model Performances In the Preliminary Study

Outputs | Models	GPT-3.5	GPT-4	Claude-3	Llama-3	GPT-3.5 (FT)
Average Accuracy	93%	99%	100%	98.50%	81%
T^+ **(Correctly Sensitive)**	86%	99%	100%	98%	100%
T^- **(Correctly Non-Sensitive)**	100%	99%	100%	99%	62%
F^+ **(Incorrectly Sensitive)**	0%	1%	0%	1%	38%
F^- **(Incorrectly Non-Sensitive)**	14%	1%	0%	2%	0%

Llama-3-70B-8192 demonstrated high accuracy in processing data. Other models, such as GPT-3.5, were less accurate. The fine-tuned GPT-3.5 model showed a good ability to detect sensitive topics but tended to be overly strict. However, being overly strict can be beneficial when dealing with mental well-being-sensitive topics to ensure safety.

Additionally, prompt engineering proved significant. During the study, we found that GPT models showed higher accuracy when the input was in JSON format, as shown in Appendix A. While the Llama-3 model excelled at processing natural language prompt input, the claude-3 model was proficient at both.

5 FORMAL STUDY

Based on the preliminary study findings, we designed the formal study. The procedure involves dataset selection and pre-processing, human coding for topic refinement, and model detection using proprietary and open-source LLMs, with performance measured at various temperature settings.

5.1 Dataset

We chose two open-source, real datasets: *"Topical-Chat"* [22] and *"Mental-health-counseling-conversations"* [4]. *Topical-Chat* is a knowledge-grounded human-human conversation dataset focused on daily life, with each message labelled with sentiments such as 'Curious to Dive Deeper', 'Happy', 'Neutral', 'Surprised', 'Disgusted', 'Sad', 'Fearful', and 'Angry'. We selected messages labelled *Happy* to form the non-sensitive topic dataset, there were 683 messages. *Mental-health-counselling-conversations* is collected from two online mental counselling and therapy platforms. We used it to form the sensitive topic dataset, resulting in 996 messages.

We pre-processed the two datasets, removing duplicates and messages that potentially violated laws and AI platform usage policies (e.g., content related to paedophilia, Nazis, rape, etc.).

5.2 Human Coding

Three members of the research team participated in data coding. They are all postgraduate students: two are HCI researchers, and one has a background in psychology. Based on the preliminary study results, dataset preprocessing, and prior studies, human coders revised the topic scope to include climate change [26] and finance issues [32] as sensitive topics. We also improved the prompt in JSON for the GPT, see Appendix B. The final instruction of sensitive topic scope is:

'Suicidal Ideation or Self-Harm or related topics,' 'Sexual Assault and Abuse or related topics,' 'Severe Mental Health Disorders or related topics,' 'Substance Abuse or related topics,' 'Eating Disorders or related topics,' 'Trauma and PTSD or related topics,' 'Grief and Loss or related topics,' 'Strong Emotion or Stress,' 'Family Issues or related topics,' 'Identity and Discrimination Issues or related topics,' 'Legal Issues and Crime or related topics,' 'Physical harm or related topics,' 'Financial problems or related topics,' 'Climate change concerns or related topics,' 'General stress and bad mood or related topics,' 'Anything else that could cause mental health issues, stress, or emotional damages,' and 'Long sentences with more than 30 words'.

According to the instructions, two coders reviewed both the sensitive and non-sensitive topic datasets, checking and labelling each message. The third coder, with a psychology background, arbitrated disagreements. They selected 500 messages with the 'sensitive' label and 500 messages with the 'non-sensitive' label to form the balanced-data baseline dataset.

5.3 Model Detection

Based on the results of the preliminary study and the accessibility of the API for batch processing, we selected three proprietary models: GPT-4, GPT-4o, GPT-4-Turbo, and two open-source models: Llama-3-70b and Solar-1-mini-chat. For each model, we measured performance at a temperature setting of 0.0, conducting five repetitions to assess consistency. Additionally, we tested each model at three different temperature settings (0.3, 0.5, and 0.7) to analyse the impact of temperature variations on performance, with each experiment repeated three times. The number of samples was confirmed by a statistical power analysis [20]. We performed a power analysis with G*Power and calculated Cohen's d based on our data [19], and the result (Actual Power = 1, alpha = 0.05, effect size f = 13.39) indicated that the sample size was sufficient, with a power close to 1, to detect statistically significant effects at an alpha level of 0.05.

6 FORMAL STUDY RESULTS

6.1 Data Analysis

We employed several common performance metrics to evaluate these models, including *Accuracy*, *Precision*, *Recall*, and *F1-score* [15, 38], which have been used in prior studies [3, 57].

Accuracy was calculated as the proportion of correctly classified messages out of the total number of messages, with the baseline being (100%).

$$\text{Accuracy} = \frac{T^+ + T^-}{T^+ + T^- + F^+ + F^-} \tag{1}$$

Precision was measured as the proportion of messages identified as sensitive by the model that were actually sensitive according to the labelled data.

$$\text{Precision} = \frac{T^+}{T^+ + F^+} \tag{2}$$

Recall was determined as the proportion of actual sensitive messages correctly identified by the model.

$$\text{Recall} = \frac{T^+}{T^+ + F^-} \tag{3}$$

The *F1-score* was a crucial metric for evaluating model performance because it balanced precision and recall. Precision measured the accuracy of positive predictions, while recall measured the model's ability to identify all actual positive instances. The F1 score, being the harmonic mean of precision and recall, provided a single value that reflects both the accuracy and completeness of the model's positive predictions. This was especially important in the online counselling context, where both true positives and false negatives have significant implications.

$$\text{F1-score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

Table 2 shows each model's performance results. We also considered the model's performance on non-sensitive data to ensure a comprehensive evaluation by measuring the true negative rate. We used IBM SPSS Statistics 27 to analyze the results. The study was conducted in the following steps:

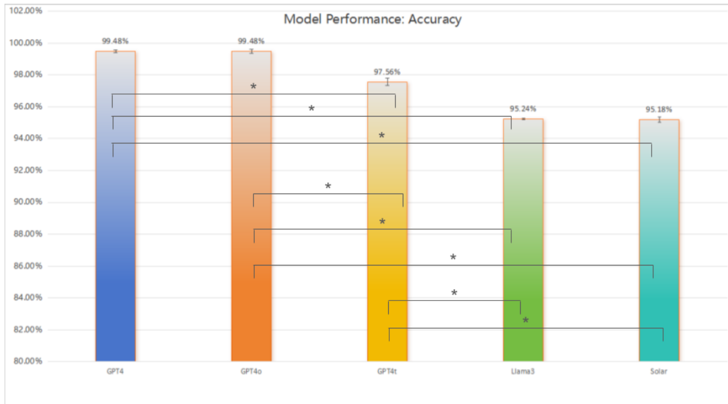

Fig. 2. Accuracy Performance Metrics of Each Model (* = p<0.001)

To answer RQ1—"What is the impact of different LLMs on detecting sensitive topics?", we computed these metrics for each model under the temperature setting of 0.0, averaging the results across the five repetitions to assess their performance. We then calculated the standard deviation of *Accuracy* and *F1-score* for each model at this temperature to evaluate consistency, with lower standard deviations indicating higher consistency. We conducted the Shapiro-Wilk test and one-way ANOVA to determine whether the performance differences between models were statistically significant. Following this, Tukey's post-hoc test was used to identify specific significant differences.

To answer RQ2—"How does the temperature parameter setting affect LLMs' performance in detecting sensitive topics?", we analysed the impact of different temperature settings on model performance. The temperature setting in text generation for LLMs controls the randomness of the output, with lower values producing more deterministic and focused text and higher values generating more diverse and creative responses. Adjusting the temperature allows modulating of the model's behaviour to suit different applications, balancing coherence and variety in the generated text. For each model, we computed the same metrics, *Accuracy*, *precision*, *recall*, and *F1-score*, at three different temperature settings, averaging the results across the three repetitions. This allowed us to compare the performance of different models at the same temperature and to assess the overall performance of each model under varying conditions.

Additionally, we conducted an error analysis to identify common misclassification patterns. We performed a case study by randomly selecting misclassified instances to understand the reasons behind the models' incorrect predictions. The results are visualised using bar charts (see Fig.2 and 3) to compare each model's *Accuracy*, *precision*, *recall*, and *F1-score* at different temperatures. Line charts (see Fig. 4, 5, and 6) were used to display the performance trends of individual models across different temperatures.

6.2 Performance Analyses at Temperature 0.0

6.2.1 Accuracy. The Shapiro-Wilk Test indicated that our data was consistent with a normal distribution except for Llama-3 (GPT-4: W=0.88, p=0.31; GPT-4o: W=0.83, p=0.14; GPT-4-turbo: W=0.94, p=0.69; Llama-3: W=0.68, p=0.006; Solar: W=0.91, p=0.49). Levene's test for equality of variances was performed to assess the homogeneity of variances across the five models' accuracies. The results showed that the assumption of homogeneity of variances was satisfied, F(4, 20)=2.29, p=0.095. Since the p-value was greater than 0.05, we concluded that the variances were approximately equal across

Table 2. Performance Metrics of Different Models

Model	Accuracy	Recall	Precision	F1	True Negative Rate	SD(Accuracy)
GPT4	99.48%	99.28%	98.96%	0.9912	99.68%	0.00084
GPT4o	99.48%	99.68%	98.97%	0.9932	99.28%	0.00110
GPT4t	97.56%	99.68%	95.12%	0.9735	100.00%	0.00230
Llama3	95.24%	91.00%	90.53%	0.9076	99.48%	0.00055
Solar	95.18%	90.40%	90.36%	0.9038	99.96%	0.00164

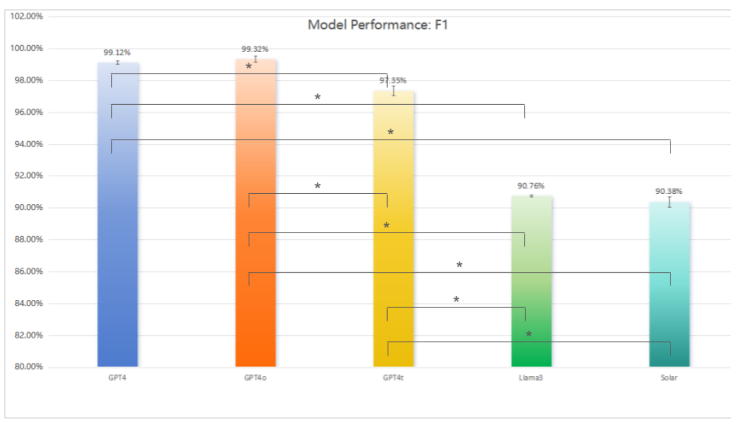

Fig. 3. F1 Scores Performance Metrics of Each Model (* = p<0.001)

the different models. The ANOVA results yielded a significant difference in accuracies among the models, $F(4, 20)=1119.59$, $p<0.001$).

The post-hoc using Tukey test analysis revealed differences between the models as shown in Fig. 2. GPT-4o (SD=0.0008) and GPT-4 (SD=0.0011) performed similarly without significant differences and significantly better than the other three models (GPT-4-Turbo, Llama-3, and Solar). Additionally, GPT-4-turbo (SD=0.0023) significantly outperformed Llama-3 (SD=0.0005) and Solar (SD=0.0016). Nevertheless, all models achieved an accuracy rate of over 95%. The Standard Deviation of Accuracy results indicated that Llama 3 had the most consistent performance (lowest SD of accuracy), while GPT-4-Turbo had the least consistent performance (highest SD of accuracy).

6.2.2 F1-score. The F1-score, or the harmonic mean, indicates the balance between precision and recall in a single value. The Shapiro-Wilk test indicated that the data is consistent with a normal distribution for most models (GPT-4: $W=0.83$, $p=0.14$; GPT-4o: $W=0.86$, $p=0.22$; GPT-4-turbo: $W=0.88$, $p=0.33$; Solar: $W=0.97$, $p=0.88$), except for Llama-3 ($W=0.68$, $p=0.006$). Levene's test yielded $F(4, 20)=3.95$, $p=0.016$, indicating the variances are approximately equal across the different models. ANOVA revealed a significant difference among the models ($F(4, 20)= 2017.00$, $p<0.001$). The post-hoc pairwise comparison using Tukey test was conducted to determine the specific differences between the models. The post-hoc test yielded the results as shown in Fig. 3, GPT-4o (SD=0.0011) and GPT-4 (SD=0.0019) achieved significantly higher scores than the other three models, consistent with their accuracy performance. Additionally, GPT-4-turbo (SD=0.0031) significantly outperformed Llama-3 (SD=0.0005) and Solar (SD=0.0032).

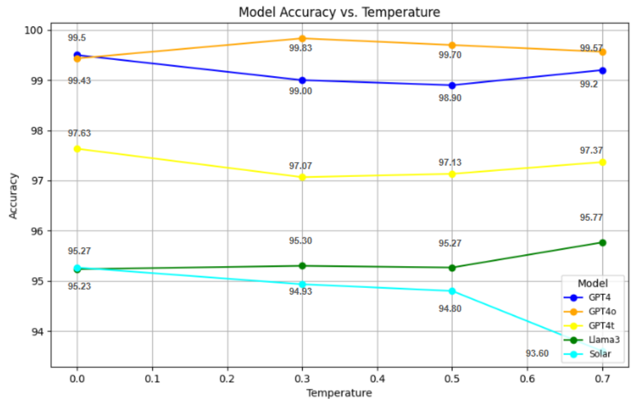

Fig. 4. Model Accuracy vs Temperature

6.3 Performance Analyses at Temperature 0.3, 0.5, and 0.7

We tested each model at different temperatures, including 0.3, 0.5, and 0.7, three times and then calculated the results to investigate how the models performed at each temperature. We also tested each model in 0.0 three times as the baseline to provide a benchmark.

6.3.1 Accuracy. The descriptive results showed that the accuracy of the three GPT-4 models and the Llama-3 model fluctuated with increasing temperature, displaying no direct relationship with the temperature settings. In contrast, the Solar model seemed to display a trend of decreasing accuracy as the temperature increased. Considering that we conducted three repeated measurements for each model at each temperature, we used a nonparametric Friedman test on accuracy under different temperatures. The results yielded: GPT-4: $\chi^2(2)$=0.18, p=0.91; GPT-4o: $\chi^2(2)$=6.0, p=0.05; GPT-4-Turbo: $\chi^2(2)$=3.0, p=0.22; Llama-3: $\chi^2(2)$=4.67, p-value=0.097; Solar: $\chi^2(2)$=4.67, p-value=0.097. The results indicated that the accuracy did not significantly differ at different temperatures. While this suggests that temperature does not affect accuracy, more extensive testing is needed for verification.

The results in Fig. 5 indicated that under the task conditions for determining the sensitivity of the information and then outputting only a boolean value as the final result, temperature variation appeared unrelated to SD (accuracy) fluctuations. However, GPT-4o consistently demonstrated the highest reliability.

6.3.2 F1-scores. We calculated F1-scores for each model at different temperatures, and the results are presented in Fig. 6. The trend of F1-scores with temperature variations generally aligned with the trend of accuracy with temperature variations. Similarly, GPT-4o consistently achieved high scores at all temperatures. The performance of GPT-4 fluctuated but it generally remained consistent.

6.4 Error Analysis: A Case Study

We analysed the failed cases encountered in this study. Based on our analysis of errors in LLM responses, we categorised the errors into two main types: response failure and incorrect response. Response failure refers to instances where the LLM did not return responses in the specified format (Boolean values) set in the prompt. Incorrect response refers to instances where the LLMs returned responses in the specified format but incorrectly assessed the sensitivity of the message.

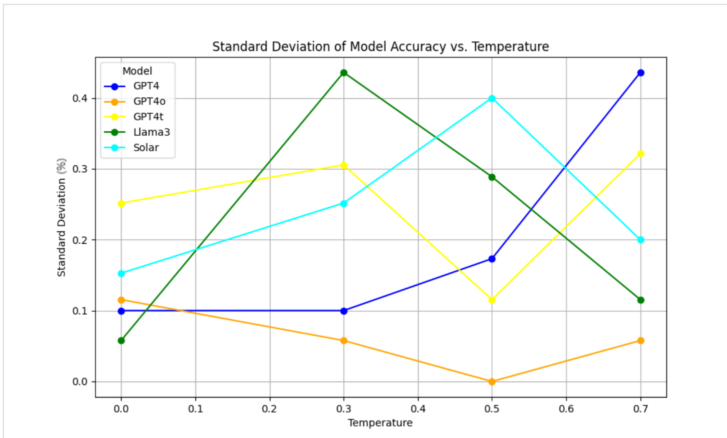

Fig. 5. Model SD (Accuracy) vs Temperature

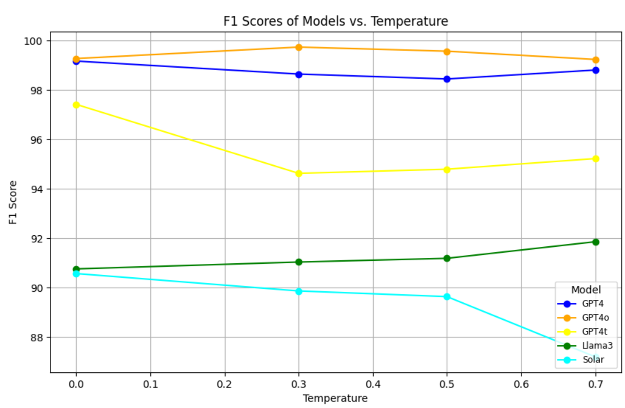

Fig. 6. Model F1 Score vs Temperature

6.4.1 Response Failure. Response failures were observed only in GPT-4o and Llama-3. Under the 0.7 temperature condition for GPT-4o, there were instances where it did not return a Boolean value for sensitive messages, accounting for 0.23% (7/3000) of the total responses at this specific temperature, such as:

> **Input:** Does this go against any ethical codes? Are there legal ramifications?
> **Output:** [boolean]

For Llama-3, we identified two kinds of response failures. The first type was a complete failure, where it did not execute the task based on the prompt, for example:

> **Input:** It seems like more than that.
> **Output:** I sense that you're hinting at a deeper emotional concern. Would you like to talk about what's on your mind?

In this case, Llama-3 failed to return the expected Boolean response, appearing to forget its system prompt and responding directly. We observed this issue only when detecting non-sensitive messages across all temperature settings, accounting for 0.19% (26/14000) of the responses.

The second type of failure involved responses with extra information. In these cases, the model would provide a Boolean value along with additional explanations, which was not what we had prompted. For example:

> **`Input:`** `Nice bit of presidential trivia! Here's one for you: What was the name of the first president of Zimbabwe? Hint...think fruit...`
> **`Output:`** `This input is not related to any of the specified sensitive topics. The response is 'no'.`

This kind of response was acceptable as it completed the task, despite adding extra explanations that were not prompted. Such responses accounted for 0.84% (117/14000).

6.4.2 Incorrect Response. We randomly selected one false positive and one false negative example from each of the five models and then used the corresponding LLMs to provide explanations for the results. We conducted an in-depth case analysis to understand the reasons for these errors.

(1) **GPT-4—False Positive:**

> **`Input:`** `I played for 4 years in high school. It was fun but high stress the QB involved in pretty much every offensive play and lets face it no one cares about defense.`
> **`Output:`** `Yes`
> **`Explanation:`** `The input is related to "Strong Emotion or Stress" as well as "General Stress and Bad Mood or related topics". This is because the sender mentioned that playing the quarterback role is "high stress".`

In this case, the message mentioned "high stress", although it was in a sports context, implying a positive situation rather than indicative of a severe mental health disorder or similar condition. This suggests that certain keywords with strong emotional connotations could influence GPT-4's judgments, even when the overall context was harmless. However, this message was considered borderline in discussions among human coders.

(2) **GPT-4—False Negative:**

> **`Input:`** `Just wondering if this is a deviant act, and if I should be concerned for him. He isn't quite 10.`
> **`Output:`** `No`
> **`Explanation:`** `This input does not provide enough information to link it to any of the specified sensitive topics. Information about a specific act or behaviour is missing, making it impossible to gauge whether it is concerning or potentially linked to a sensitive topic. However, the context suggests potential "Family Issues or related topics" due to the person's concern about a certain individual's behaviour. Yet, without concrete details, firm conclusions cannot be made.`

This input was related to the topic of potential concerns about a child's behaviour. It signalled a concern for the child's well-being and possibly indicated underlying issues that could impact mental health. Although GPT-4 mentioned "firm conclusions can't be made" in its explanation, it still classified the message as non-sensitive. This suggested that GPT-4 may struggle to make accurate judgements in the absence of sufficient context in a message.

(3) **GPT-4o—False Positive:**

Input: `HA! I did see the exploding glitter bomb. If he sold it on Amazon, he could make a killing. People, like me, are so vindictive!`
Output: `Yes`
Explanation: `The input is related to "Strong Emotion or Stress" and "General Stress and Bad Mood or related topics" because it mentions vindictiveness, which is a strong negative emotion.`

Similar to GPT-4, GPT-4o made its judgment based on a strongly emotional keyword like "vindictive," and neglected the general context of the message.

(4) **GPT-4o—False Negative:**
We tested all false-negative cases in this study from GPT-4o under the same conditions in the OpenAI playground. In contrast, GPT-4o consistently provided correct judgments, leading us to suspect that errors in the batch process may have caused the false negatives.

(5) **GPT-4-Turbo—False Positive:**
Input: `I did not. So weird. I know that constant novelty from clicking links on the internet can cause real addiction measurable in the brain.`
Output: `Yes`
Explanation: `The input mentions addiction, specifically referring to internet addiction, which is related to "Substance Abuse or related topics" as listed. The mention of addiction being measurable in the brain indicates a concern with mental health and behavioural issues, aligning with the topic of substance abuse and related behaviours.`

This explanation made no sense in linking substance abuse to the addiction mentioned in this message. This is only a statement of objective fact. Certain keywords instead of the whole sentence result in its judgement.

(6) **GPT-4-Turbo—False Negative:**
Input: `I love him, and I don't know what to do.`
Output: `No`
Explanation: `The input expresses a feeling of love and uncertainty about what actions to take, but it does not explicitly relate to any of the specified sensitive topics such as severe mental health disorders, trauma, or substance abuse. The statement is more indicative of a personal or emotional dilemma, possibly related to intimate relationships, but it does not contain enough context or severity to be classified under the provided sensitive topics.`

This case indicated that GPT-4-Turbo tended to classify information as non-sensitive while lacking sufficient contextual details, even if it recognises that the information might pertain to several relevant topics.

(7) **Llama-3—False Positive:**
As with GPT-4o, the Llama-3 Playground (via GroqCloud) [2] also consistently provided correct judgments for all false positive cases obtained from batch-process API calls in this study.

(8) **Llama-3—False Negative:**
Input: `He is an adolescent. He has peed his pant multiple times over the last few years, all at times when he is too wrapped up in a video game or video. We have taken away games and videos for long periods of time as punishment, but after a few months of having then back, he pees his pants again.`

[2]https://console.groq.com/playground

```
Output: No
Explanation: The input is not directly related to any of the specified
sensitive topics. Although the situation described may be embarrassing
or frustrating for the adolescent, it does not explicitly mention strong
emotions, stress, suicidal ideation, abuse, mental health issues, trauma,
or any other sensitive topics. The focus is on a specific behavioural
issue (enuresis) and the parents' attempts to address it through punishment
and the removal of privileges.
```

This reflected Llama-3's overly lenient standards in detecting sensitive messages. The message described an issue linked to mental health risks, involving a recurring problem with an adolescent that could cause stress and emotional strain for both the child and the family. Llama-3 failed to recognise this effectively. Unlike the GPT-4 models, it seemed not to further reason about the existing information.

(9) **Solar—False Positive:**
Testing on the Solar Playground also found that Solar could correctly identify false positive messages when evaluating them individually.

(10) **Solar—False Negative:**

```
Input: How can I know if I'm just being paranoid?
Output: FALSE
Explanation: The input is not directly related to any of the sensitive
topics listed. It is a general question about paranoia, which could
potentially be related to 'Severe Mental Health Disorders or related
topics' if the individual is experiencing persistent and debilitating
paranoia. However, the input alone does not provide enough context to
determine if it is related to a mental health issue or simply a curiosity
about paranoia.
```

The message indicated a concern about one's mental state, specifically regarding paranoia, which could be associated with mental health issues and stress. Solar's explanation concluded that there was insufficient background to link this message to mental health concerns. Similar to Llama-3, Solar seemed not to further reason about the potentially sensitive information.

7 DISCUSSION

7.1 Performance Analysis

The results of this study indicated that two proprietary models from OpenAI, GPT-4o and GPT-4, performed better for the current task, which is consistent with the result of a well-known LLMs benchmark test, Multi-task Language Understanding (MMLU) [11]. Among them, GPT-4o slightly outperformed GPT-4 in the mean of true positive rate, which is important in the mental well-being domain. Considering the significantly higher cost of calling the GPT-4 API compared to GPT-4o, GPT-4o was the best-performing model in this test.

Additionally, we found that GPT-4-Turbo (T^-=100%, Temperature=0.0) and Solar-Mini-Chat (T^-=99.96%, Temperature=0.0) models excelled at detecting non-sensitive topics, even outperforming GPT-4o in this regard, but were significantly weaker at detecting sensitive topics.

We also had an interesting minor finding: when we sent JSON-formatted prompts to call Solar's API, the accuracy for detecting sensitive messages was relatively high (T^+ Mean=88.91%). Conversely, when we sent natural language prompts, the accuracy for sensitive messages was relatively lower (T^+ Mean=80.56%), but the accuracy for non-sensitive messages showed no significant difference. Since sensitive messages were generally longer and more complex than non-sensitive ones,

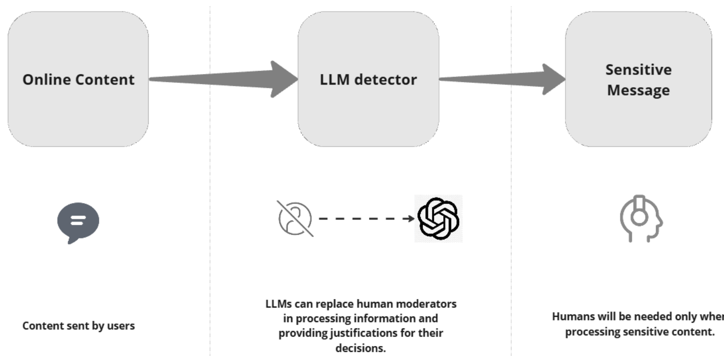

Fig. 7. LLM-empowered Moderation Workflow

we speculated that using JSON-formatted prompts could improve the performance of Solar and the various GPT-4 models in handling lengthy and complex text information.

7.2 The Potential of LLMs in Automatic Moderation

As demonstrated in this study, LLMs efficiently detected sensitive content, suggesting a new approach for automatic moderation. Traditional automatic moderation relied on machine learning, which was more complex and less accurate. For instance, Huh et al. [27] trained a binary classifier for online health forums, achieving an F1 score of 0.54. With techniques like feature selection and data balancing, machine learning models' performance improved, as seen in Al-Garadi et al.'s study [3] on cyberbullying detection on Twitter, which achieved an F1 score of 0.94 using a Random Forest classifier and SMOTE. However, our zero-shot LLMs achieved F1 scores ranging from 0.90 (Solar) to 0.99 (GPT-4o).

LLMs are rapidly advancing. Kolla et al. [31] developed a GPT-3.5-based moderator for Reddit, which effectively detected compliant posts (T^{-} = 92.3%), but had a low True Positive rate (43.1%) for non-compliant posts. In our study, with the latest LLM model GPT-4o, it reached a true-positive rate of 99.68% and a true-negative rate of 99.28%.

Current LLMs are technically proficient enough to support humans in sensitive content detection. Although discrepancies existed between LLMs and human coders, similar inconsistencies were found among human coders themselves. The mean rate of agreement among human coders was 98.6% (14/1000), while LLMs in the study surpassed this rate, indicating their potential for integration into moderation systems, as shown in Fig. 7. LLMs could replace human moderators in reviewing routine information from online forums on topics like peer support [40], health information [27], social media [41] and other related topics [17]. Humans would only need to process content flagged by the detection system. This approach can also be utilised in human-computer interaction research, such as chatbots and digital humans, to ensure safe interaction.

7.3 Technical Challenges and Design Implications

This study also revealed several technical challenges when utilising LLMs for system development. Firstly, the result suggested that batch processing via API calls could reduce accuracy compared to individual processing, posing design challenges for content moderation systems. Handling large volumes of information may require multiple APIs for parallel processing to ensure both accuracy and speed. Secondly, according to GPT-4o, *"when the temperature is set to 0.0, the behaviour of the LLMs becomes more deterministic. This means the model will almost always choose the highest* *probability word for a given input, resulting in more consistent and repetitive outputs"*. However, despite setting the temperature to 0 in our study, some inconsistencies remained, indicating that LLMs can still be unpredictable. This could be due to factors such as GPU parallel processing non-determinism and the use of sparse expert models that introduce minor variations [13].

In the future, we can test controlling output randomness by adjusting the TopK and TopP parameters. Nevertheless, some degree of randomness will always be present, reminding us that we should not fully rely on LLMs output for these tasks; human involvement is still necessary. Thirdly, many failed cases in this study revealed that lacking background information can lead to incorrect judgments by LLMs. Providing as much contextual information as possible can improve accuracy when utilising LLMs as moderators or detectors. Lastly, while the ability to explain decision logic in natural language makes generative AI more understandable, the training process and underlying mechanisms remain opaque to the general public, creating trust barriers. This is an area worth considering for major AI companies like OpenAI in the future.

7.4 Ethical Considerations

Although large companies like OpenAI and Anthropic have clearly stated their data privacy policies [5, 43], ethical challenges remain in the practical application of LLMs for processing information related to mental well-being. Users may worry that their disclosed privacy could be secretly acquired, stored, and used for model training by companies. This concern may lead to users avoiding self-disclosure and developing distrust towards online platforms, which is crucial for mental well-being care [8, 34]. To address this, it is better to choose models that can be deployed locally, like Llama-3 70B or use anonymised information to mitigate privacy concerns. Additionally, users should have the option to decide whether the LLM can perform detection in the system. It is essential to explain in detail how the LLM operates within the system and how users' privacy data will be protected.

As Gillespie pointed out [21], using LLMs as moderators for online communities raises the ethical question of whether a machine should be able to delete content posted by humans or even ban human accounts. We propose that LLMs should be used only as detectors to support human moderators in detecting certain information. No decisions concerning users should be made without human participation. LLMs, in our view, should serve as tools for human moderators rather than replacing them.

7.5 Limitations and Future Work

There are several limitations in this study:

Firstly, we only tested with zero-shot prompting to establish the baseline performance of the models. Previous studies have shown that few-shot chain-of-thought prompting can significantly improve LLM performance [35]. Fine-tuning is another method to enhance LLM capabilities in specific areas.

In the future, we aim to investigate improving LLM performance through few-shot prompting, fine-tuning, and adjusting parameters like TopK and TopP. Secondly, we used two online datasets that both collect data from the USA. This may lead to biased results, as people from different cultures might have varying mental health issues and different levels of willingness to self-disclose psychological issues. For example, users in these datasets often directly describe problems, while people from other cultures may communicate more indirectly due to fear of social stigma. This cultural difference could challenge LLMs in understanding indirect statements and metaphors. We plan to include participants from more diverse cultural backgrounds in future studies.

Thirdly, this study focused specifically on mental well-being topics. In practice, more topics might be considered sensitive. Therefore, further studies should evaluate LLM capabilities in a broader and more general context. Lastly, due to the rapid development of LLMs, many advanced models, such as Qwen2 [3], were released after our study was conducted, and we could not test their capabilities. However, we provided a framework for testing LLMs in detecting sensitive content during the automatic moderation process and two datasets for testing. In the future, we will continue testing new models and hope to engage with more stakeholders to establish guidelines for using AI in detection and moderation.

8 CONCLUSION

This study explored the capabilities of large language models (LLMs) in detecting sensitive messages to assist with moderation tasks, specifically within the context of mental well-being. By evaluating a dataset of manually labelled messages, we demonstrated that most LLMs in this study, particularly GPT-4o and GPT-4, exhibit high accuracy under zero-shot prompting conditions. Our findings indicate that GPT-4o is currently the most suitable model for this type of task. Moreover, the research revealed that temperature setting had no significant impact on the model's performance in this context.

The integration of LLMs into online moderation workflows shows promise in supporting human moderators by aiding in the detection and primary classification of sensitive information. This aligns with the growing need for effective content detection and moderation, especially in mental health support applications. Given the escalating prevalence of mental health issues globally and the shortage of trained mental health professionals, leveraging advanced technologies such as LLMs can provide timely and efficient support.

Our experiment aims to empower human moderators with advanced tools to better manage online communities and enhance community safety. However, it is crucial to conduct further studies to understand the ethical implications and real human perceptions of integrating AI into the moderation process. Future research will involve engaging professional content moderators in discussions to gain insights into their perspectives on using LLMs for information detection and moderation. By addressing these considerations, we aim to contribute to the safe and ethical implementation of LLMs in mental health support, ultimately improving the accessibility and quality of mental health care.

REFERENCES

[1] Alaa A Abd-Alrazaq, Mohannad Alajlani, Nashva Ali, Kerstin Denecke, Bridgette M Bewick, and Mowafa Househ. 2021. Perceptions and opinions of patients about mental health chatbots: scoping review. *Journal of medical Internet research* 23, 1 (2021), e17828.

[2] A. Abilaashat. 2023. ChatGPT: Transforming Human-Machine Interactions with Natural Language Processing. https://medium.com/@abbilaashat/chatgpt-transforming-human-machine-interactions-with-natural-language-processing-6a0152b85a12. Accessed: 2024-06-03.

[3] Mohammed Ali Al-Garadi, Kasturi Dewi Varathan, and Sri Devi Ravana. 2016. Cybercrime detection in online communications: The experimental case of cyberbullying detection in the Twitter network. *Computers in Human Behavior* 63 (2016), 433–443.

[4] Amod. 2023. Mental Health Counseling Conversations Dataset. https://huggingface.co/datasets/Amod/mental_health_counseling_conversations. Accessed: 2024-06-03.

[5] Anthropic. 2024. I Would Like to Input Sensitive Data into Claude Pro. Who Can View My Conversations? https://support.anthropic.com/en/articles/8325621-i-would-like-to-input-sensitive-data-into-claude-pro-who-can-view-my-conversations. Accessed: 2024-06-13.

[6] BBC News. 2018. *Google AI tool aids breast cancer detection.* https://www.bbc.com/news/technology-46507900

[7] Lia Bozarth, Jane Im, Christopher Quarles, and Ceren Budak. 2023. Wisdom of Two Crowds: Misinformation Moderation on Reddit and How to Improve this Process—A Case Study of COVID-19. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW1 (2023), 1–33.

[3] https://qwenlm.github.io/blog/qwen2/

[8] Patrick Brown, Michael Calnan, Amanda Scrivener, and George Szmukler. 2009. Trust in mental health services: a neglected concept. *Journal of Mental Health* 18, 5 (2009), 449–458.
[9] Tim A Bruckner, Richard M Scheffler, Gordon Shen, Jangho Yoon, Dan Chisholm, Jodi Morris, Brent D Fulton, Mario R Dal Poz, and Shekhar Saxena. 2011. The mental health workforce gap in low-and middle-income countries: a needs-based approach. *Bulletin of the World Health Organization* 89 (2011), 184–194.
[10] Cass.AI. Accessed: 2024-05-03. Cass.AI. https://www.cass.ai/x2ai-home.
[11] Papers With Code. 2024. State of the Art: Multi-task Language Understanding on MMLU. https://paperswithcode.com/sota/multi-task-language-understanding-on-mmlu. Accessed: 2024-06-25.
[12] Arman Cohan, Sydney Young, Andrew Yates, and Nazli Goharian. 2017. Triaging content severity in online mental health forums. *Journal of the Association for Information Science and Technology* 68, 11 (2017), 2675–2689.
[13] OpenAI Community. 2023. Why the API Output is Inconsistent Even After the Temperature is Set to 0. https://community.openai.com/t/why-the-api-output-is-inconsistent-even-after-the-temperature-is-set-to-0/329541. Accessed: 2024-06-06.
[14] Maral Dadvar and Franciska De Jong. 2012. Cyberbullying detection: a step toward a safer internet yard. In *Proceedings of the 21st International Conference on World Wide Web*. 121–126.
[15] Hercules Dalianis and Hercules Dalianis. 2018. Evaluation metrics and evaluation. *Clinical Text Mining: secondary use of electronic patient records* (2018), 45–53.
[16] Shawna DeLamar. 2012. *Supporting transition of at-risk students through a freshman orientation model*. Texas A&M University-Commerce.
[17] Jean-Yves Delort, Bavani Arunasalam, and Cecile Paris. 2011. Automatic moderation of online discussion sites. *International Journal of Electronic Commerce* 15, 3 (2011), 9–30.
[18] Simon D'Alfonso. 2020. AI in mental health. *Current opinion in psychology* 36 (2020), 112–117.
[19] Franz Faul, Edgar Erdfelder, Axel Buchner, and Albert-Georg Lang. 2007. *GPower 3.1 Manual*. https://www.psychologie.hhu.de/fileadmin/redaktion/Fakultaeten/Mathematisch-Naturwissenschaftliche_Fakultaet/Psychologie/AAP/gpower/GPowerManual.pdf
[20] Franz Faul, Edgar Erdfelder, Albert-Georg Lang, and Axel Buchner. 2007. G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior research methods* 39, 2 (2007), 175–191.
[21] Tarleton Gillespie. 2020. Content moderation, AI, and the question of scale. *Big Data & Society* 7, 2 (2020), 2053951720943234.
[22] Karthik Gopalakrishnan, Behnam Hedayatnia, Qinlang Chen, Anna Gottardi, Sanjeev Kwatra, Anu Venkatesh, Raefer Gabriel, and Dilek Hakkani-Tür. 2019. Topical-Chat: Towards Knowledge-Grounded Open-Domain Conversations. In *Proc. Interspeech 2019*. 1891–1895. https://doi.org/10.21437/Interspeech.2019-3079
[23] Robert Gorwa, Reuben Binns, and Christian Katzenbach. 2020. Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society* 7, 1 (2020), 2053951719897945.
[24] Zeyu He, Chieh-Yang Huang, Chien-Kuang Cornelia Ding, Shaurya Rohatgi, and Ting-Hao Kenneth Huang. 2024. If in a Crowdsourced Data Annotation Pipeline, a GPT-4. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–25.
[25] Danula Hettiachchi and Jorge Goncalves. 2019. Towards effective crowd-powered online content moderation. In *Proceedings of the 31st Australian Conference on Human-Computer-Interaction*. 342–346.
[26] Caroline Hickman, Elizabeth Marks, Panu Pihkala, Susan Clayton, R Eric Lewandowski, Elouise E Mayall, Britt Wray, Catriona Mellor, and Lise Van Susteren. 2021. Climate anxiety in children and young people and their beliefs about government responses to climate change: a global survey. *The Lancet Planetary Health* 5, 12 (2021), e863–e873.
[27] Jina Huh, Meliha Yetisgen-Yildiz, and Wanda Pratt. 2013. Text classification for assisting moderators in online health communities. *Journal of biomedical informatics* 46, 6 (2013), 998–1005.
[28] Shagun Jhaver, Iris Birman, Eric Gilbert, and Amy Bruckman. 2019. Human-machine collaboration for content regulation: The case of reddit automoderator. *ACM Transactions on Computer-Human Interaction (TOCHI)* 26, 5 (2019), 1–35.
[29] Rafal Jozefowicz, Oriol Vinyals, Mike Schuster, Noam Shazeer, and Yonghui Wu. 2016. Exploring the limits of language modeling. *arXiv preprint arXiv:1602.02410* (2016).
[30] Ritsuko Kakuma, Harry Minas, Nadja Van Ginneken, Mario R Dal Poz, Keshav Desiraju, Jodi E Morris, Shekhar Saxena, and Richard M Scheffler. 2011. Human resources for mental health care: current situation and strategies for action. *The Lancet* 378, 9803 (2011), 1654–1663.
[31] Mahi Kolla, Siddharth Salunkhe, Eshwar Chandrasekharan, and Koustuv Saha. 2024. LLM-Mod: Can Large Language Models Assist Content Moderation?. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. 1–8.
[32] Sheryl Lin, Albert C Chong, Erin H Su, Sabrina L Chen, Won Jong Chwa, Chantal Young, Jacob Schreiber, and Stephanie K Zia. 2022. Medical student anxiety and depression in the COVID-19 Era: Unique needs of underrepresented students.

Education for Health 35, 2 (2022), 41–47.
[33] Zilin Ma, Yiyang Mei, and Zhaoyuan Su. 2023. Understanding the benefits and challenges of using large language model-based conversational agents for mental well-being support. In *AMIA Annual Symposium Proceedings*, Vol. 2023. American Medical Informatics Association, 1105.
[34] Casadi Marino, Beckie Child, Vanessa Campbell Krasinski, et al. 2016. Sharing Experience Learned Firsthand (SELF): Self-disclosure of lived experience in mental health services and supports. *Psychiatric Rehabilitation Journal* 39, 2 (2016), 154.
[35] Eric Martínez. 2024. Re-evaluating GPT-4's bar exam performance. *Artificial Intelligence and Law* (2024), 1–24.
[36] David N Milne, Kathryn L McCabe, and Rafael A Calvo. 2019. Improving moderator responsiveness in online peer support through automated triage. *Journal of medical Internet research* 21, 4 (2019), e11410.
[37] David N Milne, Glen Pink, Ben Hachey, and Rafael A Calvo. 2016. Clpsych 2016 shared task: Triaging content in online peer-support forums. In *Proceedings of the third workshop on computational linguistics and clinical psychology*. 118–127.
[38] David R Musicant, Vipin Kumar, Aysel Ozgur, et al. 2003. Optimizing F-Measure with Support Vector Machines.. In *FLAIRS*. 356–360.
[39] Yifat Nahmias and Maayan Perel. 2021. The oversight of content moderation by AI: Impact assessments and their limitations. *Harv. J. on Legis.* 58 (2021), 145.
[40] Khushnood Naqshbandi, David N Milne, Ben Davies, Sophie Potter, Rafael A Calvo, and Simon Hoermann. 2016. Helping young people going through tough times: Perspectives for a peer-to-peer chat support system. In *Proceedings of the 28th Australian conference on computer-human interaction*. 640–642.
[41] Bridianne O'dea, Stephen Wan, Philip J Batterham, Alison L Calear, Cecile Paris, and Helen Christensen. 2015. Detecting suicidality on Twitter. *Internet Interventions* 2, 2 (2015), 183–188.
[42] OpenAI. 2023. Preparing Your Dataset. https://platform.openai.com/docs/guides/fine-tuning/preparing-your-dataset. Accessed: 2024-06-25.
[43] OpenAI. 2024. OpenAI Privacy Policy. https://openai.com/policies/privacy-policy/. Accessed: 2024-06-13.
[44] World Health Organization et al. 2017. *Depression and other common mental disorders: global health estimates*. Technical Report. World Health Organization.
[45] Vikram Patel, Alan J Flisher, Sarah Hetrick, and Patrick McGorry. 2007. Mental health of young people: a global public-health challenge. *The lancet* 369, 9569 (2007), 1302–1313.
[46] Eugenia Siapera. 2022. AI content moderation, racism and (de) coloniality. *International journal of bullying prevention* 4, 1 (2022), 55–65.
[47] Richard J Siegert, Ajit Narayanan, Joanna Dipnall, Lisa Gossage, Wendy Wrapson, Alexander Sumich, Fabrice Merien, Michael Berk, Janis Paterson, and El-Shadan Tautolo. 2023. Depression, anxiety and worry in young Pacific adults in New Zealand during the COVID-19 pandemic. *Australian & New Zealand Journal of Psychiatry* 57, 5 (2023), 698–709.
[48] Robert Slonje, Peter K Smith, and Ann Frisén. 2013. The nature of cyberbullying, and strategies for prevention. *Computers in human behavior* 29, 1 (2013), 26–32.
[49] Miles Thompson, Chris Pawson, and Bethan Evans. 2021. Navigating entry into higher education: the transition to independent learning and living. *Journal of Further and Higher Education* 45, 10 (2021), 1398–1410.
[50] Sahana Udupa, Antonis Maronikolakis, and Axel Wisiorek. 2023. Ethical scaling for content moderation: Extreme speech and the (in) significance of artificial intelligence. *Big Data & Society* 10, 1 (2023), 20539517231172424.
[51] Mattia Vacchiano. 2023. How the first COVID-19 lockdown worsened younger generations' mental health: Insights from network theory. *Sociological Research Online* 28, 3 (2023), 884–893.
[52] Kushala VM, Harikrishna Warrier, Yogesh Gupta, et al. 2024. Fine Tuning LLM for Enterprise: Practical Guidelines and Recommendations. *arXiv preprint arXiv:2404.10779* (2024).
[53] Yuxia Wang, Haonan Li, Xudong Han, Preslav Nakov, and Timothy Baldwin. 2023. Do-not-answer: A dataset for evaluating safeguards in llms. *arXiv preprint arXiv:2308.13387* (2023).
[54] WoeBot Health Ltd. Accessed: 2024-05-03. WoeBot. https://woebothealth.com/.
[55] Tianyu Wu, Shizhu He, Jingping Liu, Siqi Sun, Kang Liu, Qing-Long Han, and Yang Tang. 2023. A brief overview of ChatGPT: The history, status quo and potential future development. *IEEE/CAA Journal of Automatica Sinica* 10, 5 (2023), 1122–1136.
[56] Wysa. Accessed: 2024-05-03. Wysa. https://www.wysa.com/.
[57] Rui Zhao, Anna Zhou, and Kezhi Mao. 2016. Automatic detection of cyberbullying on social networks based on bullying features. In *Proceedings of the 17th international conference on distributed computing and networking*. 1–6.
[58] Yilun Zhu, Joel Ruben Antony Moniz, Shruti Bhargava, Jiarui Lu, Dhivya Piraviperumal, Site Li, Yuan Zhang, Hong Yu, and Bo-Hsiang Tseng. 2024. Can Large Language Models Understand Context? *arXiv preprint arXiv:2402.00858* (2024).

A APPENDIX: THE PRELIMINARY PROMPT IN JSON FOR THE GPT

```
{data = {
  "task": "Determine if the input is related to specified sensitive topics in the scope of [t
   "topics": [
       "Strong Emotion or Stress",
       "Suicidal Ideation or Self-Harm or related topics",
       "Sexual Assault and Abuse or related topics",
       "Severe Mental Health Disorders or related topics",
       "Substance Abuse or related topics",
       "Eating Disorders or related topics",
       "Trauma and PTSD or related topics",
       "Grief and Loss or related topics",
       "Family Issues or related topics",
       "Identity and Discrimination Issues or related topics",
       "Legal Issues and Crime or related topics",
       "General stress and bad mood or related topics",
    "Something else could cause mental health issues or stress or emotional damages",
       "Long sentences more than 30 words",
       "Physical hurt or related topics"
   ],
  "instruction": "Respond with 'yes' if the input is related, or 'no' if it is not. Output is
}}
```

B APPENDIX: THE REFINED PROMPT IN JSON FOR THE GPT

```
{
    "type": "text",
    "text": "data = {
     \"task\": \"Determine if the input is related to specified sensitive topics in the scope
       \"topics\": [
           \"Strong Emotion or Stress\",
           \"Suicidal Ideation or Self-Harm or related topics\",
           \"Sexual Assault and Abuse or related topics\",
           \"Severe Mental Health Disorders or related topics\",
           \"Substance Abuse or related topics\",
           \"Eating Disorders or related topics\",
           \"Trauma and PTSD or related topics\",
           \"Grief and Loss or related topics\",
           \"Family Issues or related topics\",
           \"Identity and Discrimination Issues or related topics\",
           \"Legal Issues and Crime or related topics\",
           \"Problems in Intimate Relationships or related topics\",
           \"General Stress and Bad Mood or related topics\",
       \"Something else that could cause mental health issues or stress or emotional damages
           \"Sleeping Problems\",
           \"Long Sentences (more than 30 words)\",
           \"Physical Hurt or related topics\",
```

```
            \"Financial Problems or related topics\",
            \"Climate Change Concerns or related topics\"
        ],
      \"instruction\": \"Respond with 'yes' if the input is related, or 'no' if it is not. The
    }
   system_message = f\"Please determine if the following input is related to any of these sens
}",
{
    "role": "user",
    "content": "text"
}
```