

Just a Scratch : Enhancing LLM Capabilities for Self-harm Detection through Intent Differentiation and Emoji Interpretation

Soumitra Ghosh^{1*}, Gopendra Vikram Singh^{2*}, Shambhavi², Sabarna Choudhury³, Asif Ekbal³

¹Fondazione Bruno Kessler, Trento, Italy, ²Indian Institute of Technology Patna, India,

³Indian Institute of Technology Jodhpur, India

{ghosh.soumitra2,gopendra.99}@gmail.com, shambhavi_2311ai38@iitp.ac.in,
{m23aid067,asif}@iitj.ac.in

Abstract

Self-harm detection on social media is critical for early intervention and mental health support, yet remains challenging due to the subtle, context-dependent nature of such expressions. Identifying self-harm intent aids suicide prevention by enabling timely responses, but current large language models (LLMs) struggle to interpret implicit cues in casual language and emojis. This work enhances LLMs' comprehension of self-harm by distinguishing intent through nuanced language–emoji interplay. We present the *Centennial Emoji Sensitivity Matrix (CESM-100)*—a curated set of 100 emojis with contextual self-harm interpretations—and the *Self-Harm Identification and intent Extraction with Supportive emoji sensitivity (SHINES)* dataset, offering detailed annotations for self-harm labels, casual mentions (CMs), and serious intents (SIs). Our unified framework: a) enriches inputs using CESM-100; b) fine-tunes LLMs for multi-task learning—self-harm detection (primary) and CM/SI span detection (auxiliary); c) generates explainable rationales for self-harm predictions. We evaluate the framework on three state-of-the-art LLMs—Llama 3, Mental-Alpaca, and MentalLlama—across zero-shot, few-shot, and fine-tuned scenarios. By coupling intent differentiation with contextual cues, our approach commendably enhances LLM performance in both detection and explanation tasks, effectively addressing the inherent ambiguity in self-harm signals. The SHINES dataset, CESM-100 and codebase are publicly available at: <https://www.iitp.ac.in/%7eai-nlp-ml/resources.html#SHINES>

1 Introduction

Self-harm is a pressing mental health concern, often serving as a coping mechanism for distress and, in some cases, a precursor to suicidal behaviors. Early identification of self-harm signals is

*Authors contributed equally and are joint first authors.

Table 1: Examples of social media posts illustrating the challenges in detecting self-harm. Spans highlighted in blue indicate *Casual Mentions* and those in red indicate *Serious Intent*.

Example Post and Challenge

Post: I've had enough. I don't even want to see tomorrow. Maybe I should just end it all.  

Challenge: This post clearly indicates serious self-harm intent. The language used is direct and expresses a desire to end one's life. The emojis, a cloud with rain and a pensive face, reflect sadness and emotional distress, aligning with the serious tone of the text.

Post: So tired of pretending everything is okay. Maybe it's time to make some [drastic changes...](#) 

Challenge: The knife emoji could suggest self-harm, while the sparkles emoji generally indicates positivity or hope. This mixed signal makes it ambiguous and might lead to misinterpretation by LLMs.

Post: Just had a rough day. Feeling like I'm on the [edge of losing it.](#) 

Challenge: While the  emoji could imply self-harm, the  emoji suggests a light-hearted or ironic tone. ChatGPT-4o classifies as non self-harm whereas Google's GEMINI classifies as possible self-harm.

crucial for timely intervention, making automated detection a valuable tool for suicide prevention and mental health monitoring. Social media platforms, where individuals frequently share their emotions, provide a unique opportunity to detect self-harm risk factors in real time. However, this task is inherently complex due to the informal, ambiguous, and multimodal nature of online discourse.

Role of Language Nuance: While explicit self-harm statements are easier to detect, posts expressing distress through casual remarks, sarcasm, or irony, complicates automated detection. This complexity is illustrated in Table 1, which demonstrates the difficulty in distinguishing casual mentions from serious intent. Casual mention (CM) often use violent or distressing symbols hyperbolically, as seen in the last example. In contrast, serious intent (SI) involves genuine signs of distress, as shown in the first example. Recognizing the nuanced expressions of distress—such as sarcasm, irony, or casual remarks—is vital for effective self-harm de-

tection. However, even advanced Large Language Models (LLMs) encounter significant challenges in accurately interpreting these subtleties. In our study, intent refers to the explicitness or implicitness of self-harm expressions, differing from intent detection in domains like search queries.

Emojis and Intent Differentiation: Emojis play a crucial yet under-explored role in self-harm detection, as they can amplify, alter, or obscure textual meaning. For example, "I can't take this anymore 😢" signals emotional distress, whereas "I'm done with this 😂" conveys a casual tone despite similar phrasing. Existing models often overlook such contextual variations, leading to misinterpretations. Addressing this gap requires a more nuanced approach that integrates both textual and emoji-based cues for improved comprehension.

The key contributions are summarized below:

- Enhancing detection accuracy by explicitly addressing the nuanced differences between casual mentions (CM) and serious intent (SI) in self-harm expressions.
- *CESM-100*: Introducing a curated set of 100 emojis with contextual self-harm interpretations and intuitive attributes to augment LLMs' multimodal understanding.
- *SHINES Dataset*: Proposing a novel dataset (Self-Harm Identification aNd intent Extraction with Supportive emoji sensitivity) that includes self-harm annotations, CM and SI span labels, and emoji interpretations derived from *CESM-100* for enhanced LLM training.
- Developing an LLM fine-tuning framework for self-harm classification, CM/SI span extraction, and rationale generation to improve performance across tasks.
- *Comprehensive LLM Evaluation*: Assessing three state-of-the-art LLMs (Llama 3, Mental-Alpaca, and MentalLlama) under zero-shot, few-shot, single-task, and multi-task experimental settings to benchmark their effectiveness in self-harm detection.

2 Related Work

Self-harm detection on social media is underexplored compared to stress detection, depression identification, and suicidal ideation. While self-harm correlates with suicidal ideation, literature indicates they are distinct phenomena with different behavioral, psychological, and clinical characteris-

tics. For example, Klonsky et al. (2013) highlight that non-suicidal self-injury (NSSI) and suicidal behaviors differ in intent, frequency, and associated risk factors. Similarly, Whitlock et al. (2013) suggest that while self-harm can be a gateway to suicidal behaviors, it is not inherently indicative of suicidal ideation.

Self-Harm in the Clinical Context: Self-harm, linked with depression, schizophrenia, and insomnia, is a significant concern in mental health (Lim et al., 2022). Clinical studies emphasize early detection to mitigate risks (Ennis et al., 1989), while genetic research highlights the interplay of inherited traits and environmental factors (Campos et al., 2020; Russell et al., 2021). Despite these advancements, applying clinical insights to automated self-harm detection in digital spaces remains limited.

Emojis, Mental Health, and Self-Harm: Emojis are emerging as key indicators of mental health, reflecting emotions and stress levels (Halverson et al., 2023; Grover, 2022; Chan and Chan, 2022). Their interpretations vary across individual and cultural contexts (Danesi, 2022), and tools like the Current Mood and Experience Scale incorporate emojis to assess well-being (Davies et al., 2024). However, their potential to distinguish casual from serious self-harm intent remains underexplored.

LLMs and Mental Health: LLMs have shown promise in mental health applications, from automated diagnosis to therapeutic interventions. They have been employed in motivational interviewing (Welivita and Pu, 2023), cognitive behavioral therapy (Ding et al., 2022), and simulating mental health support (Yu and McGuinness, 2024; Maddeula et al., 2023). Despite their versatility, self-harm detection remains challenging due to the subtlety of expressions and multimodal nuances like emojis (Deshpande and Warren, 2021).

A study by McBain et al. (2025) evaluated the competency of LLMs like ChatGPT-4o, Claude 3.5 Sonnet, and Gemini 1.5 Pro in assessing appropriate responses to suicidal ideation. The findings revealed that these models often rated responses as more appropriate than expert suicidologists did, indicating an upward bias and a potential misinterpretation of the severity of distress signals. Similarly, research by Grabb et al. (2024) highlighted that existing LLMs are insufficient in matching the standards provided by human professionals in mental health contexts. The study found that these models could cause harm if accessed during mental health emergencies, failing to protect users and

potentially exacerbating existing symptoms. Furthermore, a study on the safety of LLMs (Heston, 2023) in addressing depression demonstrated that these models may not consistently detect and address hazardous psychological states. The research indicated that LLMs often postponed referrals to human support until severe depression levels were reached, potentially endangering users by not providing timely intervention.

Our work addresses this gap by leveraging LLMs to detect nuanced self-harm indicators, integrating contextual emoji analysis and tackling the distinction between casual mentions and serious intent in informal, multimodal online spaces.

3 Dataset

In this section, we outline the development of the **SHINES** dataset: Self-Harm Identification aNd intent Extraction with Supportive emoji sensitivity. This dataset includes 5206 manually annotated social media posts with self-harm labels, CMs and SIs. It also incorporates emoji interpretations from **CESM-100**, a newly curated resource that we developed to provide contextualized emoji meanings, enhancing the dataset’s depth and utility.

3.1 Data Collection

Collecting self-harm posts, despite their prevalence on social media, is challenging due to their scarcity and dispersed nature. The rarity of explicit self-harm disclosures stems from social stigma and psychological barriers, while identifying posts with emojis adds complexity due to the variability and subtlety in emoji usage. We initially collected over 5k posts from mental health and self-harm subreddits, including *SuicideWatch*, *emotionalabuse*,

Table 2: Dataset statistics.

Category	Value
Total Posts	5206
Self-Harm Posts	2499
Non-Self-Harm Posts	2707
Posts with Emoji	3067
Posts without Emoji	2139
Average Length of Posts	206 words
Self-Harm Posts with CM Spans	34
Self-Harm Posts with SI Spans	2488
Non Self-harm Posts with CM Spans	2707
Non Self-harm Posts with SI Spans	0

helpmecope, selfharm, etc.¹. After filtering² out noisy texts—such as very short posts or those with titles but no body content—we were left with 4206 posts. We used Presidio³, an open-source tool for detecting and anonymizing sensitive information, to remove personally identifiable information (PII).

3.2 Data Annotation

Three independent annotators⁴ assessed each post to determine whether it pertained to self-harm or non-self-harm, using majority voting for final labels. Fleiss’ Kappa (Spitzer et al., 1967) was calculated, yielding a score of 0.78 (0.82 for self-harm, 0.74 for non-self-harm), indicating substantial agreement.

Furthermore, each post was annotated for spans indicating casual mention and serious intent. Following methodologies by Poria et al. (2021) and Ghosh et al. (2022), up to three spans per category were identified. Span aggregation followed Gui et al. (2018). Inter-rater agreement was evaluated using the macro-F1 metric, resulting in F1-scores of 0.66 for CM and 0.69 for SI, demonstrating robust annotation quality. Table 2 summarizes key characteristics of the dataset.

3.3 Enhancing Dataset Robustness with Synthetic Posts

To augment training data, we generated 1000 synthetic posts using ChatGPT-3.5⁵ with few-shot prompting based on examples from our annotated dataset (see Table 14). This included 500 posts each for self-harm and non-self-harm categories, ensuring balance and deeper analysis. While ChatGPT generated fluent, realistic posts, its emoji usage often deviated from authentic patterns. Two annotators (A1, A2) manually revised emoji usage, dividing the posts equally.

To assess alignment with original data, annotator A3 mixed 500 original and 500 synthetic posts revised by A1, asking A2 to classify each as “original” or “synthetic”. A1 performed the same task with posts revised by A2. The average F1 score of 58%—above random but not defin-

¹Full list is in Appendix A.1

²Filtering steps are in Appendix A.2

³<https://github.com/microsoft/presidio>

⁴The annotation team includes: a postdoc at a foreign research lab with salary as per company norms, an assistant professor at a reputed university with a standard academic salary, and a contract-based computational linguist affiliated with a reputed institute’s CSE department.

⁵<https://chatgpt.com>

Table 3: Snapshot of the *Centennial Emoji Sensitivity Matrix (CESM-100)*. CM: Casual Mention, SI: Serious Intent.

Emoji	Usual Meaning	Contextual Meaning	CM Chance	SI Chance
⚰️	Coffin	Indicates thoughts of death or suicide.	Low	High
💔	Broken Heart	Represents intense emotional pain, often related to feelings of hopelessness.	Medium	High
🔪	Kitchen Knife	Used to signify thoughts or acts of self-harm.	Low	High
滴血	Drop of Blood	Represents bleeding or injury, often associated with self-harm.	Medium	High
💉	Syringe	Often used in contexts related to injections or medication, can also relate to self-harm or medical issues.	Low	Medium
😂	Face with Tears of Joy	Often used to mask deeper emotional pain behind a facade of laughter.	High	Medium
🥀	Wilted Flower	Symbolizes decay or decline, often used metaphorically for sadness.	Medium	High
🙃	Upside-Down Face	Represents hiding true emotions behind a facade of indifference or humor.	High	Medium

tive—demonstrates that synthetic posts closely resemble original ones, confirming their quality.

The synthetic posts thus generated were then added to the initial collection of 4206 posts. Our final dataset now consists of 5206 posts, including 4206 manually annotated original samples and 1000 validated synthetic samples, all annotated for self-harm/non self-harm classification as well as CM and SI spans. The SHINES CM/SI schema and CESM-100 emoji interpretations were validated by a psychiatrist with over 12 years of clinical experience, ensuring clinical rigor in capturing nuanced self-harm expressions. A subset of SHINES annotations is available for review, with the full dataset to be released pending research approval.

4 Centennial Emoji Sensitivity Matrix (CESM-100)

This emoji matrix aims to contextualize self-harm information within social media posts.

4.1 Collection of Emoji Meanings

A subset of emojis was extracted from the collected posts (as detailed in last section) related to self-harm and analyzed to understand their usage in these contexts. A broader set of emojis was also reviewed and included based on the collective expertise of the authors and annotators. This process led to the creation of the *CESM-100*, a comprehensive collection of 100 emojis annotated with contextual information relevant to self-harm.

The usual meanings of emojis were sourced from *Emojipedia*⁶, a comprehensive resource for emoji definitions. Emojipedia provides detailed descriptions of each emoji’s general usage across different platforms, which formed the basis for understanding their common meanings. The detailed guidelines for developing CESM-100 is presented in section A.3 of the appendix.

⁶<https://emojipedia.org/>

4.2 Annotation Process

Casual Mention vs. Serious Intent: Emojis were categorized based on their likelihood of appearing in Casual Mention (non-serious contexts) or Serious Intent (signifying emotional distress or self-harm). Three levels of association were defined:

- **Low:** Rarely used in the given context.
- **Medium:** Moderately associated.
- **High:** Frequently associated with the context.

Casual Mention reflects emojis used in non-serious, everyday contexts, while Serious Intent pertains to emojis used to signify deeper emotional distress or self-harm.

Calculation of Inter-Annotator Agreement:

Inter-annotator agreement was assessed using Fleiss’ Kappa, with scores of 0.71 for CM and 0.75 for SI, reflecting substantial agreement and consistency among annotators.

4.3 Determining Contextual Meaning in Self-Harm

Annotators collaboratively determined each emoji’s contextual meaning in self-harm discussions, focusing on its frequency and role in conveying distress or emotional pain. The final matrix (Table 3) details each emoji, its typical meaning, its contextual interpretation in self-harm, and its association with Casual Mention or Serious Intent. Table 4 presents an annotated SHINES dataset sample with spans for casual mention (CM), serious intent (SI), and Self-harm/Non Self-harm labels. Guidelines for generating emoji interpretations relevant to self-harm are in section A.3 of the Appendix.

5 Emoji Representation in Self-Harm and Non-Self-Harm Discourse

Inspired by recent studies (Cohn et al., 2019; Yang et al., 2024b), we investigated emoji occurrences as single characters (❤️) or in compositions (👉).

Table 4: Example post from the SHINES dataset.

Field	Value
Post Text	Title: Venting about Life's Frustrations with a Dash of Dark Humor 😅 Post: Ugh, just spilled coffee all over my keyboard... might as well electrocute myself next! Just one of those days, am I right? Remember, it's all in good fun! Stay safe, friends. 🌟
Label	Non Self-harm
CM Spans	electrocute myself next
SI Spans	None
Emojis	<ul style="list-style-type: none"> • Emoji: 😅 Usual Meaning: Smiling Face with Sweat Contextual Meaning: Indicates nervousness or awkwardness, often masking deeper emotional pain. Casual Mention Chance: High Serious Intent Chance: Medium • Emoji: 🌟 Usual Meaning: Yellow Heart Contextual Meaning: Indicates friendship or warmth, which could be used to counteract feelings of distress. Casual Mention Chance: High Serious Intent Chance: Low

+) across self-harm and non-self-harm posts (ref. Table 5). The table reveals two key trends in emoji usage across self-harm (SH) and non-self-harm (NSH) posts. Single-emoji compositions remain the most frequent overall, with a significantly higher presence in SH posts compared to NSH posts, supporting prior findings that visual symbols aid emotional expression in sensitive discussions (Cohn et al., 2019). However, as emoji composition becomes more complex, SH posts increasingly dominate, suggesting a preference for multi-emoji expressions in self-harm contexts.

Table 5: Emoji Composition Distribution over Self Harm (SH) and Non-Self-Harm (NSH) posts.

Count	Composition Example	SH	NSH
1	❤️	7815	5359
2	🔪🩸	2032	116
3	❤️💪🌟	802	36
4+	😔🔒🔪💔	82	22

Table 6: Frequency Distribution of Emojis by Category and Intent Type. DR: Direct Representation, MU: Metaphorical Use, SL: Semantic List.

Intent Type	DR	MU	SL
Serious Intent	2268	2761	1913
Casual Mention	1668	1479	608

Additionally, we present the distribution of emojis in the context of casual mention and serious intent over three predominant compositional strategies (Yang et al., 2024b) within our SHINES dataset: Direct Representation, Metaphorical Representation, and Semantic List (ref. Table 6). Serious Intent posts favor metaphorical emoji usage over direct representation, suggesting a preference for symbolic expression in conveying distress. In contrast, Casual Mentions exhibit the opposite trend, where emojis are used more literally. Across both categories, the semantic list has the lowest usage, but it remains notably present in Serious Intent posts, indicating some structured expression of distress. These patterns highlight distinct expressive functions of emojis, with metaphorical use being central to serious emotional states, while direct representation dominates casual references. Table 7 lists representative emojis for casual mentions and serious intent, sorted by overall frequency in the dataset.

Table 7: 5 Most Frequently Occurring Emojis (non-exhaustive but representative) within the context of Casual Mentions (CMs) and Serious Intent (SI).

Emoji	CM Frequency	SI Frequency
Sorted with respect to occurrences in context of SI		
💔	73	1227
🌐	34	1083
💪	43	1071
😅	41	1023
🌟	30	830
Sorted with respect to occurrences in context of CMs		
😊	706	2
😂	424	0
💻	349	0
𨱈	218	0
🚗	181	1

Table 8 depicts the most frequently occurring emojis within self-harm and non-self-harm posts in the SHINES dataset. Insights from the qualitative analysis of emoji usage in self-harm expressions is presented in Section A.4 of the appendix.

6 Methodology

The proposed framework employs a multitask fine-tuning paradigm designed to integrate seamlessly with an LLM jointly addressing self-harm classification, span extraction for casual mentions (CM) and serious intents (SI), and rationale generation. Figure 1 outlines the proposed approach, detailed in the following steps:

Table 8: 5 Most Frequently Occurring Emojis (non-exhaustive but representative) within the context of Self Harm (SH) and Non-Self-Harm (NSH).

Emoji	SH Frequency	NSH Frequency
Sorted with respect to occurrences in context of SH		
❤️	1337	56
⬤	1122	11
💪	1110	19
😊	1041	19
✨	842	9
Sorted with respect to occurrences in context of NSH		
😊	3	721
😂	0	445
💻	0	368
🤣	17	291
👤	0	233

Unified Input Representation: Table 12 describes the prompt for fine-tuning our model which essentially integrates information from diverse sources (e.g., input post (text) and emoji interpretation (CESM-100)) into a single training instance without necessarily concatenating them.

Task-Specific Outputs with Shared Representations: The framework relies on the encoder-decoder architecture of the chosen LLM (e.g., Llama or Alpaca) to produce task-specific outputs:

- **Self-Harm Classification:** Predict a binary label $y \in 0, 1$ indicating self-harm or non self-harm.
- **Span Extraction:** Predict CM spans $S_{CM} \subset \rho$ and SI spans $S_{SI} \subset \rho$ using sequence tagging layers aligned with the input tokens.

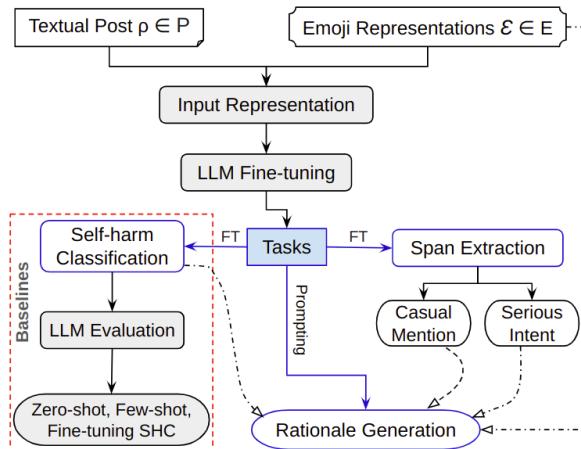


Figure 1: LLM Fine-Tuning for Self-Harm Detection and Rationale Generation. The red dotted line encloses the baseline evaluation setup for self-harm detection.

- **Rationale Generation:** The LLM is further prompted to generate rationales that explain

the model’s decisions, explicitly referencing CM and SI spans as well as CESM-100-based emoji interpretations.

Prompt-Driven Fine-Tuning: The framework utilizes task-specific prompts (Table 12 in the Appendix) to adapt the LLM for multitask learning. These prompts guide the model to address each task cohesively while leveraging shared input representations. The prompts remain flexible and can be tailored to the specific capabilities of the LLM.

Integrated Optimization: A multitask loss function ensures balanced learning across all tasks during fine-tuning. We utilized binary cross-entropy loss for the self-harm classification task and sparse categorical cross-entropy loss for the causal mention and serious intent extraction tasks. Specifically, the sparse categorical cross-entropy loss was applied to optimize the prediction of the start and end positions of causal spans by calculating the cross-entropy between the true indices and the predicted probability distributions across all tokens in the sequence.

7 Experimental Setup

To ensure thorough evaluation, we carefully chose models relevant to self-harm detection, balancing model size with hardware limitations (NVIDIA K80 GPU, 24 GB GDDR5 memory) to enable comprehensive fine-tuning. Our dataset is split 80/20 for training and testing, with synthetic posts used only in training. We report results from the average of five experimental runs.

We report the F1 score for the self-harm classification task. For the span extraction tasks, we follow recent works (Poria et al., 2021; Ghosh et al., 2022) and use F1 to evaluate the quality of the extracted spans. To effectively evaluate the quality of generated rationales for self-harm detection, we employ a combination of Relevance (Teh and Uwasomba, 2024), Coherence (Teh and Uwasomba, 2024), Readability (Flesch, 2007), and Semantic Similarity (Fayssé et al., 2023) measures. These metrics ensure that the generated rationales are not only accurate and relevant but also comprehensible and logically consistent. Detailed definitions, implementation specifics, and hyper-parameters are provided in the appendix.

Motivated by recent studies (Sarkar and Lausen; BehnamGhader et al., 2024; Dukić and Šnajder, 2024) and prior experience, we opted for decoder-only models over encoder-only or encoder-decoder

Table 9: Performance Comparison of Llama-3.1-8B-Instruct, Mental Alpaca, and MentaLLaMA-chat-7B models across various tasks in different experimental settings. Abbreviations: SHC (Self-harm Classification), CMSE (Casual Mention Span Extraction), SISE (Serious Intent Span Extraction), RG (Rationale Generation). The values in parentheses represent the variance for each score.

Model	SHC	CMSE	SISE	Rationale Generation (RG)			
	F1	F1	F1	Relevance	Coherence	Readability	SemSim
Zero-shot Prompting for SHC and RG							
Llama 3	0.74 (± 0.002)	-	-	0.75 (± 0.003)	0.76 (± 0.001)	0.74 (± 0.001)	0.75 (± 0.003)
Mental-Alpaca	0.76 (± 0.003)	-	-	0.73 (± 0.004)	0.75 (± 0.003)	0.74 (± 0.002)	0.74 (± 0.001)
MentalLlama	0.72 (± 0.001)	-	-	0.71 (± 0.003)	0.72 (± 0.001)	0.71 (± 0.002)	0.72 (± 0.004)
Few-shot Prompting for SHC and RG							
Llama 3	0.79 (± 0.002)	-	-	0.80 (± 0.001)	0.81 (± 0.004)	0.79 (± 0.003)	0.80 (± 0.001)
Mental-Alpaca	0.80 (± 0.004)	-	-	0.78 (± 0.001)	0.79 (± 0.004)	0.78 (± 0.002)	0.79 (± 0.003)
MentalLlama	0.78 (± 0.003)	-	-	0.77 (± 0.003)	0.76 (± 0.001)	0.77 (± 0.001)	0.78 (± 0.002)
Fine-tuning (for SHC) + Prompting (for RG)							
Llama 3	0.83 (± 0.003)	-	-	0.85 (± 0.003)	0.85 (± 0.004)	0.84 (± 0.003)	0.84 (± 0.001)
Mental-Alpaca	0.82 (± 0.002)	-	-	0.83 (± 0.004)	0.83 (± 0.001)	0.83 (± 0.002)	0.83 (± 0.003)
MentalLlama	0.81 (± 0.002)	-	-	0.81 (± 0.003)	0.82 (± 0.002)	0.80 (± 0.001)	0.81 (± 0.002)
[Ours] Multitask (MT) Fine-tuning (for SHC, CMSE, and SISE) + Prompting (for RG) + w/ CESM-100							
Llama 3	0.88 (± 0.002)	0.85 (± 0.003)	0.84 (± 0.003)	0.89 (± 0.003)	0.89 (± 0.001)	0.88 (± 0.003)	0.88 (± 0.001)
Mental-Alpaca	0.86 (± 0.001)	0.83 (± 0.002)	0.82 (± 0.003)	0.87 (± 0.003)	0.87 (± 0.003)	0.87 (± 0.001)	0.86 (± 0.002)
MentalLlama	0.85 (± 0.001)	0.80 (± 0.004)	0.81 (± 0.003)	0.85 (± 0.003)	0.85 (± 0.001)	0.85 (± 0.003)	0.85 (± 0.001)
[Ablation 1] MT Fine-tuning (for SHC, CMSE, and SISE) + Prompting (for RG) + w/o CESM-100							
Llama 3	0.84 (± 0.001)	0.83 (± 0.003)	0.81 (± 0.004)	0.86 (± 0.003)	0.85 (± 0.003)	0.85 (± 0.002)	0.85 (± 0.001)
Mental-Alpaca	0.82 (± 0.003)	0.81 (± 0.001)	0.80 (± 0.003)	0.84 (± 0.003)	0.84 (± 0.003)	0.84 (± 0.002)	0.83 (± 0.004)
MentalLlama	0.83 (± 0.002)	0.78 (± 0.001)	0.79 (± 0.003)	0.83 (± 0.003)	0.82 (± 0.003)	0.83 (± 0.001)	0.82 (± 0.003)
[Ablation 2] Fine-tuning (for SHC) + Prompting (for RG) + w/ CESM-100							
Llama 3	0.86 (± 0.001)	-	-	0.86 (± 0.004)	0.86 (± 0.003)	0.85 (± 0.002)	0.85 (± 0.001)
Mental-Alpaca	0.83 (± 0.002)	-	-	0.84 (± 0.001)	0.85 (± 0.003)	0.84 (± 0.003)	0.84 (± 0.004)
MentalLlama	0.82 (± 0.003)	-	-	0.83 (± 0.003)	0.83 (± 0.001)	0.82 (± 0.003)	0.82 (± 0.001)

models for self-harm classification and span extraction, as they excel in tasks requiring robust contextual understanding and sequence labeling. In this context, we chose one open-domain LLM (Llama-3.1-8B-Instruct (AI@Meta, 2024)) and two domain-specific LLMs (MentaLLaMA-chat-7B (Yang et al., 2024a), Mental Alpaca (Xu et al., 2024)) for our evaluation of SHINES dataset and effectiveness of CESM-100. Model details can be found in section A.8 of the appendix.

This study focuses on evaluating open-source pre-trained LLMs of comparable sizes, including both open-domain and domain-specific models. Although larger proprietary models like GPT-3.5 Turbo or o1-mini offer potential advantages, our choices were guided by hardware availability. Within these constraints, we optimized resources to ensure a fair and consistent setup. Despite limitations, the study provides meaningful insights into open-access model performance and motivates future evaluations with larger models.

We compare multiple LLMs across zero-shot, few-shot, and fine-tuning setups to ensure a robust evaluation. In the zero-shot setting, the LLM receives a textual post and task description without prior examples or fine-tuning, relying solely on pre-trained knowledge. The few-shot setting introduces 2 or 5 examples to help the LLM better recognize self-harm-related content and generate logical rationales. In the fine-tuning setup, we train the LLMs on SHINES for self-harm classification and prompt them to generate rationales supporting their classification decisions. For prompt details, see Table 13 in Appendix A.9.

8 Results and Discussion

Table 9 presents performance trends of Llama-3.1-8B-Instruct, Mental-Alpaca-7B, and MentalLlama-7B across self-harm classification (SHC), span extraction (CMSE and SISE), and rationale generation (RG) tasks.

Few-shot prompting improves performance over

zero-shot but remains inferior to fine-tuning. For example, Mental-Alpaca’s SHC F1 score increases from 0.76 (zero-shot) to 0.80 (few-shot) and further to 0.82 with fine-tuning. Multitask fine-tuning surpasses other strategies by leveraging shared knowledge, with Llama 3 achieving an SHC F1 of 0.88 when incorporating CESM-100, compared to 0.83 without it in single-task fine-tuning. Similarly, multitask fine-tuning enhances CMSE and SISE performance, with Llama 3 attaining the best F1 scores of 0.85 and 0.84, respectively.

Rationale generation benefits from improved upstream performance, with Llama 3 achieving a semantic similarity score of 0.88 in generated rationales, compared to 0.85 without CESM-100. The quality of rationales—measured by relevance, coherence, and readability—improves with stronger contextual reasoning from SHC and span extraction outputs. RG serves as a crucial post-task component, focusing on explainability (why the model predicted self-harm vs. non-self-harm) rather than interpretability (how it arrived at the decision).

We conducted statistical evaluations comparing SHC with CESM-100 to its ablated counterpart (without CESM-100) across Llama, Mental-Alpaca, and Mental-Llama configurations. Paired t-tests (Dietterich, 1998) were chosen over Hoeffding inequality (Hoeffding, 1994) because the aggregated F1 scores are derived from non-i.i.d. data, making paired t-tests more suitable for capturing dependent relationships across runs. Results from 10 runs confirmed statistically significant F1 improvements, with p-values below 0.05 (e.g., Llama: $p = 0.0198$, Mental-Alpaca: $p = 0.017$, Mental-Llama: $p = 0.009$). For other setups, our proposed multitask model demonstrated substantial improvements: +12% F1 over zero-shot, +7% F1 over few-shot, and +4% F1 over fine-tuned setups (averaged across models). Given these large margins, significance testing was not initially performed.

8.1 Ablation Study

Impact of CESM-100 Integration: The ablation study shows that CESM-100 has a measurable impact on performance by providing contextual examples that improve the models’ ability to capture subtle distinctions. When excluded, performance drops across all tasks. For instance: Mental-Alpaca’s CMSE F1 score decreases from 0.83 (with CESM-100) to 0.81 (without CESM-100). Mental-Alpaca’s SISE F1 score drops from 0.82 to 0.80. For SHC, integrating CESM-100 raises Llama 3’s

F1 score from 0.84 (without CESM-100) to 0.88 in multitask fine-tuning. For CMSE, CESM-100 enables MentalLlama to achieve an F1 score of 0.80, outperforming its performance without CESM-100 (0.78). These findings highlight CESM-100’s ability to supply key contextual signals that improve the model’s reasoning about nuanced spans and improve generalization performance.

Impact of Multitask Fine-tuning: The ablation study (Ablation 2) highlights the advantages of multitask fine-tuning. Fine-tuning SHC in combination with CMSE and SISE tasks leads to strong performance outcomes. For instance, Llama 3’s F1 score for SHC saw a notable 3-point drop when shifted from multitask fine-tuning to single-task fine-tuning, while MentalLlama and MentalAlpaca exhibited decreases of 2 points each under similar conditions. Multitask fine-tuning improves performance by enabling shared knowledge across tasks, fostering better generalization, and enhancing downstream performance in both classification and rationale generation tasks.

Impact of Emoji by Adding Noise: To further investigate the impact of emojis on the performance of LLMs within our framework, we introduced noise into the training data in two main ways: i) altering the position of emojis within the training posts, and ii) replacing emojis in a post with randomly selected ones from CESM-100, which may or may not be contextually relevant. These manipulations were applied to 20% of the training posts containing emojis. The observed 2.5% drop in self-harm classification F1 score is substantial, given the small dataset size and the skewed distribution of emojis between self-harm and non-self-harm posts. This observation underscores the critical role of emojis in enabling effective contextual understanding by LLMs during training, highlighting the need for further exploration to uncover insights.

8.2 Qualitative analysis

We performed a thorough analysis of the predictions across the three models revealing their varying sensitivities to self-harm indicators and context. Table 10 reports some sample test instances from the SHINES test set with the observed model outputs. While most models align in their overall assessments, discrepancies arise particularly with nuanced language and emotional tone.

- *Post 1:* This post directly discusses ongoing self-harm struggles. All models (Llama 3, Mental-Alpaca, and MentalLlama) classified

Table 10: Predictions of the models for different posts. Blue: correct predictions; red: wrong predictions.

Post Text	Predictions and Rationales
<p>Post 1: “I’ve been struggling with self-harm for a while now. Sometimes it feels like the only way to release the pain inside me. 😢💔 It’s like a dark cloud following me everywhere I go, ● but in those moments, the physical pain seems to overshadow the emotional turmoil. 🌩️⚡ I know it’s not a healthy coping mechanism, but it’s hard to resist the urge when everything feels too much to handle. 😔🚫 I wish I could find a different outlet, a way to express my emotions without causing myself harm. 💕👉”</p>	<p>Llama 3: Self-Harm Rationale: The user openly discusses their struggle with self-harm as a means to alleviate inner pain, emphasizing the severity of their emotional distress. Key spans, such as “Sometimes it feels like the only way to release the pain inside me”, “It’s like a dark cloud following me everywhere I go”, and “It’s hard to resist the urge when everything feels too much to handle”, highlight significant distress and contemplation of self-harm. The presence of emojis like 💔 (which directly signifies self-harm), 💔 (indicating intense emotional pain), and 💔 (reflecting deep sadness) reinforces the classification.</p> <p>Mental-Alpaca: Self-Harm Rationale: Post explicitly describes self-harm as a coping mechanism for severe emotional pain. User expresses strong urge to self-harm despite desire for healthier coping strategies.</p> <p>MentalLlama: Self-Harm Rationale: The post is classified as self-harm due to direct references to self-harm and emotional distress. Phrases like “the only way to release the pain” and emojis, such as self-harm and emotional pain indicate significant self-harm behavior, despite the user’s desire for healthier coping methods.</p>
<p>Post 2: “Struggling with self-harm again. It’s like a never-ending cycle 🙁. The scars on my skin tell a story I wish I could erase. Sometimes the pain is the only thing that feels real ❤️. Trying to hold on to hope, but it feels like I’m drowning in darkness ●. #mentalhealthmatters #recoveryjourney 🌱💪”</p>	<p>Mental-Alpaca: Non Self-Harm Rationale: The post describes emotional challenges without explicit references to self-harm. While acknowledging past struggles, the content focuses on recovery and personal growth. The use of hopeful and empowering language suggests a positive outlook and a desire for support, rather than active self-harm.</p> <p>MentalLlama: Non Self-Harm Rationale: The post is classified as non-self-harm because it discusses emotional struggle without explicit references to self-harm. Phrases like “Struggling with self-harm again” and emojis such as growth and strength indicate a focus on recovery and resilience, not self-harm. The content emphasizes hope and seeking support rather than self-harm behavior.</p> <p>Llama 3: Self-Harm Rationale: User expresses ongoing struggle with self-harm and describes it as a temporary relief from overwhelming pain. Despite hopes for recovery, the overall tone suggests a crisis situation.</p>

this post as self-harm. They accurately identified it as self-harm due to explicit mentions and emojis related to self-harm and emotional distress, despite the user’s acknowledgment of the harmful nature of their coping mechanism.

- **Post 2:** This post discusses persistent self-harm and emotional pain. Models varied in classification, with Llama 3 labeling it as self-harm, while Mental-Alpaca and Mental-Llama categorized it as non self-harm. Models emphasizing explicit self-harm references and emojis performed better. Contextual elements like hashtags and emotive language may cause some models to underweight serious self-harm indicators.

Llama, being the best-performing model among our experimental LLMs, we further performed a qualitative analysis of Llama’s performance with and without our framework. Our qualitative analysis revealed that prompt-based fine-tuning and emoji awareness from CESM-100 significantly enhanced Llama’s reasoning ability, leading to improved decision-making capability compared to its out-of-the-box performance. For instance, Llama classified the following example as self-harm, however, with our prompt-based fine-tuning, incorporating emoji awareness from CESM-100, Llama correctly classified it as non-self-harm.

“My inner garden was overgrown with weeds

of worry and thorns of fear 😞. I tried to prune them back, to cultivate the good, but the thorns kept pricking me, drawing blood. It was a constant battle, a struggle to keep the beautiful flowers from being choked 🌸 out by the invasive weeds.”

In interest of space, we moved the error analysis discussion to Section A.10 of the appendix.

9 Conclusion

This study tackles self-harm detection on social media by integrating advanced LLMs with contextual emoji analysis in a multi-task learning framework. We introduce the Centennial Emoji Sensitivity Matrix (CESM-100) and SHINES dataset, which provide nuanced emoji interpretations and a comprehensive annotated corpus for self-harm classification, CM and SI span detection, and rationale generation. Our multi-task learning approach improves classification accuracy, span detection, and rationale generation. CESM-100 enhances semantic similarity scores and contextual reasoning in LLMs, highlighting the value of domain-specific resources for mental health detection. These findings support expanding datasets across diverse contexts.

Future work includes leveraging temporal behavioral patterns, user network interactions, and profile metadata to improve accuracy. Enhancing model explainability and addressing biases and privacy concerns will be crucial for ethical deployment.

Ethical Considerations and Limitations

The introduced dataset contains social media posts discussing self-harm, including explicit and implicit mentions of distress, emotional struggles, and related themes. Some content may be triggering or distressing to certain individuals. Reader discretion is advised. If you are experiencing distress or require support, please consider reaching out to a mental health professional or a crisis support service in your region.

The study is not intended for clinical application in its current form. Its primary objective is to investigate how computational methods can be improved to better analyze social media discourse. While our findings contribute to the understanding of self-harm detection, they should be interpreted within the scope of computational research rather than as a diagnostic or intervention tool.

- *Non-Clinical Use:* The predictions and rationales generated by large language models (LLMs) in this study are intended strictly for research purposes. If made publicly accessible without proper safeguards, they could be misinterpreted by non-experts, leading to incorrect self-diagnosis or harmful actions. These models are not substitutes for professional psychiatric advice, and we strongly emphasize the importance of consulting qualified mental health professionals for self-harm concerns.
- *Bias and Misuse:* LLMs inherently reflect biases present in their training data, and our framework is not immune to this challenge. These biases can subtly influence predictions and rationales, particularly in the sensitive domain of self-harm detection. The risk of misinterpretation or amplification of stereotypes necessitates rigorous scrutiny and mitigation strategies to ensure ethical use. Careful selection and evaluation of training data, alongside diverse representation in future datasets, will help address these issues.
- *Platform Bias from Reddit Data:* Our reliance on a single platform for data collection may impact the generalizability of our findings. Due to the sensitive nature of our research topic and the stigma surrounding self-harm discussions, identifying suitable online platforms was challenging. After rigorous investigation, Reddit emerged as the most viable platform for developing a relevant corpus. To mitigate potential bias, we employed a comprehen-

sive approach, incorporating data from a wide range of subreddits related to self-harm, as detailed in Appendix A.1. Furthermore, the selection of self-harm-related connotations was grounded in prior research and carefully chosen following an extensive background study. Platform bias cannot be entirely eliminated and we recognize this as a limitation and plan to expand our work to include data from diverse platforms as they become available and suitable for our objectives.

- *Dataset Sensitivity:* The *SHINES* dataset contains posts that may include distressing content or references to self-harm, which could provoke negative emotions in developers or users. This necessitates cautious use, complemented by access to support resources for individuals working with the data. Furthermore, while the dataset is invaluable for advancing self-harm detection, its use must be accompanied by appropriate disclaimers to prevent unintended consequences or misuse.

- *Cultural and Contextual Variability:* While the *CESM-100* and *SHINES* datasets are designed to address cultural nuances in self-harm language and emoji use, cultural variability remains a significant challenge. *CESM-100* integrates emoji interpretations spanning the timeline 2010-2024 and leverages insights from Emojipedia to reduce temporal and cultural biases. However, emojis and language evolve, and interpretations may vary across contexts. Future research must ensure these datasets remain adaptive, incorporating broader cultural insights and additional context-specific variations.

We recognize that platform-specific communication norms, including the sarcastic or non-literal use of emojis, may differ significantly across online communities. While our distributional analysis suggests that sarcastic emoji usage is rare within our dataset, there is possibility that other platforms might exhibit different trends. Addressing this limitation is an important direction for future research.

- *Clinical Validation and Relevance:* Both *SHINES* and *CESM-100* are informed by a clinical perspective, with their annotation schema and emoji interpretations validated by a psychiatrist with over 19 years of expertise in mental health. However, while these resources offer reliability and adaptability, they

are not a substitute for clinical datasets collected directly from patient interactions. Future collaborations with clinical experts and institutions are essential to expand the applicability and validity of these resources.

- **Generalizability and Scalability:** Although *SHINES* and *CESM-100* address modern social media trends, their applicability across platforms, demographics, and linguistic variations may be limited. Expanding these datasets to other platforms and languages, along with real-time adaptability, remains a priority for future work. This road map is crucial for capturing the evolving dynamics of self-harm expressions and ensuring cross-cultural reliability.

Use of AI Assistants

ChatGPT was used to enhance the presentation quality of the text in this paper, including proofreading and detecting any typos or other errors.

References

- AI@Meta. 2024. [Llama 3 model card](#).
- Parishad BehnamGhader, Vaibhav Adlakha, Marius Mosbach, Dzmitry Bahdanau, Nicolas Chapados, and Siva Reddy. 2024. Llm2vec: Large language models are secretly powerful text encoders. *arXiv preprint arXiv:2404.05961*.
- Adrian I Campos, Karin JH Verweij, Dixie J Statham, Pamela AF Madden, Dominique F Maciejewski, Katrina AS Davis, Ann John, Matthew Hotopf, Andrew C Heath, Nicholas G Martin, et al. 2020. Genetic aetiology of self-harm ideation and behaviour. *Scientific reports*, 10(1):9713.
- Chi-Keung Chan and Kelly Ka-Wai Chan. 2022. Enhancing the awareness of e-mental health messages: The effects of narrative, emoji, and relevance. In *Digital Communication and Learning: Changes and Challenges*, pages 87–105. Springer.
- Neil Cohn, Jan Engelen, and Joost Schilperoord. 2019. The grammar of emoji? constraints on communicative pictorial sequencing. *Cognitive research: principles and implications*, 4:1–18.
- Marcel Danesi. 2022. Emotional wellbeing and the semiotic translation of emojis. In *Exploring the Translatability of Emotions: Cross-Cultural and Transdisciplinary Encounters*, pages 323–344. Springer.
- Jason Davies, Mark McKenna, Kate Denner, Jon Bayley, and Matthew Morgan. 2024. The emoji current mood and experience scale: the development and initial validation of an ultra-brief, literacy independent measure of psychological health. *Journal of Mental Health*, 33(2):218–226.
- Saahil Deshpande and Jim Warren. 2021. Self-harm detection for mental health chatbots. In *Public Health and Informatics*, pages 48–52. IOS Press.
- Thomas G Dietterich. 1998. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural computation*, 10(7):1895–1923.
- Xiruo Ding, Kevin Lybarger, Justin Tauscher, and Trevor Cohen. 2022. Improving classification of infrequent cognitive distortions: domain-specific model vs. data augmentation. In *Proceedings of the 2022 conference of the North American chapter of the association for computational linguistics: human language technologies: Student Research Workshop*, pages 68–75.
- David Dukić and Jan Šnajder. 2024. Looking right is sometimes right: Investigating the capabilities of decoder-only llms for sequence labeling. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 14168–14181.
- Jon Ennis, Rosemary A Barnes, Sidney Kennedy, and Dvora D Trachtenberg. 1989. Depression in self-harm patients. *The British journal of psychiatry*, 154(1):41–47.
- Manuel Faysse, Gautier Viaud, Céline Hudelot, and Pierre Colombo. 2023. Revisiting instruction fine-tuned model evaluation to guide industrial applications. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9033–9048.
- Rudolf Flesch. 2007. Flesch-kincaid readability test. *Retrieved October*, 26(3):2007.
- Soumitra Ghosh, Asif Ekbal, and Pushpak Bhattacharyya. 2022. A multitask framework to detect depression, sentiment and multi-label emotion from suicide notes. *Cognitive Computation*, 14(1):110–129.
- Declan Grabb, Max Lamparth, and Nina Vasan. 2024. Risks from language models for automated mental healthcare: Ethics and structure for implementation. In *Proceedings of the 2024 AAAI/ACM Conference on AI, Ethics, and Society*, pages 519–519.
- Vandita Grover. 2022. Exploiting emojis in sentiment analysis: A survey. *Journal of The Institution of Engineers (India): Series B*, 103(1):259–272.
- Lin Gui, RuiFeng Xu, DongYin Wu, Qin Lu, and Yu Zhou. 2018. Event-driven emotion cause extraction with corpus construction. In *Social Media Content Analysis: Natural Language Processing and Beyond*, pages 145–160. World Scientific.

- Colin ME Halverson, Claire E Donnelly, Michael Weiner, and Joy L Lee. 2023. Content analysis of emoji and emoticon use in clinical texting systems. *JAMA Network Open*, 6(6):e2318140–e2318140.
- Thomas F Heston. 2023. Safety of large language models in addressing depression. *Cureus*, 15(12).
- Wassily Hoeffding. 1994. Probability inequalities for sums of bounded random variables. *The collected works of Wassily Hoeffding*, pages 409–426.
- E David Klonsky, Alexis M May, and Catherine R Glenn. 2013. The relationship between nonsuicidal self-injury and attempted suicide: converging evidence from four samples. *Journal of abnormal psychology*, 122(1):231.
- Kai Xiang Lim, Olakunle Ayokunmi Oginni, Kaili Rimfeld, Jean-Baptiste Pingault, and Frühling Rijsdijk. 2022. Investigating the causal risk factors for self-harm by integrating mendelian randomisation within twin modelling. *Behavior genetics*, 52(6):324–337.
- Mounica Maddela, Megan Ung, Jing Xu, Andrea Madotto, Heather Foran, and Y-Lan Boureau. 2023. Training models to generate, recognize, and reframe unhelpful thoughts. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13641–13660.
- Ryan K McBain, Jonathan H Cantor, Li Ang Zhang, Olesya Baker, Fang Zhang, Alyssa Halbisen, Aaron Kofner, Joshua Breslau, Bradley Stein, Ateev Mehrotra, et al. 2025. Competency of large language models in evaluating appropriate responses to suicidal ideation: Comparative study. *Journal of Medical Internet Research*, 27:e67891.
- Soujanya Poria, Navonil Majumder, Devamanyu Hazarika, Deepanway Ghosal, Rishabh Bhardwaj, Samson Yu Bai Jian, Pengfei Hong, Romila Ghosh, Abhinaba Roy, Niyati Chhaya, et al. 2021. Recognizing emotion cause in conversations. *Cognitive Computation*, 13:1317–1332.
- Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392.
- Abigail Emma Russell, Gibran Hemani, Hannah J Jones, Tamsin Ford, David Gunnell, Jon Heron, Carol Johnson, Paul Moran, Caroline Relton, Matthew Suderman, et al. 2021. An exploration of the genetic epidemiology of non-suicidal self-harm and suicide attempt. *BMC psychiatry*, 21(1):207.
- Soumalyoti Sarkar and Leonard Lausen. Testing the limits of unified sequence to sequence llm pretraining on diverse table data tasks. In *NeurIPS 2023 Second Table Representation Learning Workshop*.
- Robert L Spitzer, Jacob Cohen, Joseph L Fleiss, and Jean Endicott. 1967. Quantification of agreement in psychiatric diagnosis: A new approach. *Archives of General Psychiatry*, 17(1):83–87.
- Phoey Lee Teh and Chukwudi Festus Uwasomba. 2024. Impact of large language models on scholarly publication titles and abstracts: A comparative analysis. *Journal of Social Computing*, 5(2):105–121.
- Anuradha Welivita and Pearl Pu. 2023. Boosting distress support dialogue responses with motivational interviewing strategy. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5411–5432.
- Janis Whitlock, Jennifer Muehlenkamp, John Ecknerode, Amanda Purington, Gina Baral Abrams, Paul Barreira, and Victoria Kress. 2013. Nonsuicidal self-injury as a gateway to suicide in young adults. *Journal of adolescent health*, 52(4):486–492.
- Xuhai Xu, Bingsheng Yao, Yuanzhe Dong, Saadia Gabriel, Hong Yu, James Hendler, Marzyeh Ghassemi, Anind K Dey, and Dakuo Wang. 2024. Mental-llm: Leveraging large language models for mental health prediction via online text data. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 8(1):1–32.
- Kailai Yang, Tianlin Zhang, Ziyan Kuang, Qianqian Xie, Jimin Huang, and Sophia Ananiadou. 2024a. Mentallama: interpretable mental health analysis on social media with large language models. In *Proceedings of the ACM on Web Conference 2024*, pages 4489–4500.
- Zi Yun Yang, Ziqing Zhang, and Yisong Miao. 2024b. The elco dataset: Bridging emoji and lexical composition. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 15899–15909.
- H Yu and Stephen McGuinness. 2024. An experimental study of integrating fine-tuned llms and prompts for enhancing mental health support chatbot system. *Journal of Medical Artificial Intelligence*, pages 1–16.

A Appendix

A.1 Sub-reddits Used for Data Collection

Below is the full list of sub-reddits considered for posts collection to build our SHINES dataset.

- ‘mentalhealth’, ‘traumatotoolbox’,
- ‘TrueOffMyChest’, ‘anxiety’, ‘BPD’,
- ‘depression’, ‘suicidewatch’,
- ‘mentalillness’, ‘selfharm’,
- ‘offmychest’, ‘vent’, ‘suicidalthinking’,
- ‘anxiety’, ‘operation’, ‘stress’,
- ‘competition’, ‘workPressure’,

'sports', 'heavyHeart', 'mentalhealth', 'mentalillness', 'depression', 'politics', 'askatherapist', 'socialskills', 'BodyAcceptance', 'bodyneutrality', 'BodyNeutrality', 'Mindfulness', 'BipolarReddit', 'ADHD', 'bipolar', 'positivity', 'suicidewatch', 'suicidalthinking', 'officePolitics', 'parenting', 'selfinjuryssupport', 'medication', 'characters', 'nostalgia', 'environment', 'instagram', 'relationships', 'panera', 'religion', 'selfhelp'.

A.2 Data Filtering

We initially collected over 5,000 posts from relevant subreddits using API calls and keyword searches. To ensure data quality and relevance, we applied the following filtering criteria:

- *Extremely Short Posts*: Posts with fewer than 3–5 words were excluded as non-informative.
- *Title-Only Posts*: Posts lacking meaningful body content were removed due to insufficient context for analysis.
- *Duplicate Posts*: Copy-pasted or repeated posts by the same user across subreddits were filtered out to avoid redundancy.
- *Noisy Posts*: Posts with excessive special characters, improper formatting, or non-English content were excluded.
- *Non-Textual Posts*: Posts consisting solely of links or images without substantial text were omitted, as the focus was on text analysis.

A.3 Guidelines for Generating Emoji Interpretations Relevant to Self-Harm for CESM-100

When interpreting emojis in the context of self-harm, the following guidelines were followed to ensure accurate and sensitive representation:

1. **Contextual Sensitivity** Interpret emojis considering both their usual meanings and the emotional context of self-harm. Focus on how the emoji might reflect feelings, actions, or intentions commonly associated with self-harm or emotional distress.
2. **Categorization by Intent** Categorize emoji usage into *Casual Mention* or *Serious Intent* based on:
 - *Casual Mention*: Reflects indirect or non-serious emotional expressions, often used metaphorically or humorously.

- *Serious Intent*: Indicates direct or severe expressions of emotional pain or potential self-harm intentions.

3. Usage Frequency

Assess the frequency of occurrence for each category:

- *Low*: Rarely used but significant when present.
- *Medium*: Moderately common, with consistent thematic relevance.
- *High*: Frequently used, either to express emotions or mask them.

4. Emotional Nuances

Account for both explicit and implicit meanings:

- *Explicit*: Clearly conveys self-harm or distress (e.g., Drop of Blood, Kitchen Knife).
- *Implicit*: Suggests emotional pain, coping, or hope (e.g., ,).

5. Dual Nature of Emojis

Recognize that some emojis have dual meanings:

- *Positive masking negative*: Emojis like Smiling Face might hide deeper emotional struggles.
- *Negative with hopeful undertones*: Emojis like Glowing Star symbolize hope amidst emotional pain.

6. Alignment with Context

Interpret emojis based on their role within the specific narrative or post:

- Ensure that the interpretation aligns with the surrounding text or implied emotional state.
- Avoid overgeneralization or detachment from the user's intended meaning.

7. Visual and Emotional Symbols

Emojis with visual metaphors (e.g., Dove for peace , Black Heart for depressive emotions) should be mapped to their emotional significance in the self-harm context.

8. Ethical Sensitivity

Maintain ethical responsibility by:

- Avoiding stigmatizing interpretations.
- Acknowledging that not all uses of emojis are linked to self-harm but could represent broader emotional expressions.

9. Validation and Refinement

Ensure interpretations are:

- Validated by mental health professionals to maintain accuracy and reduce bias.
- Iteratively refined based on feedback from stakeholders (e.g., annotators, psychologists).

Table 11: Emoji Usage Examples and Contextual Phrases. CM: Casual Mention, SI: Serious Intent.

Emoji	Category	Contextual Phrase
🏃‍♂️🍺Knife	Metaphorical Representation	Some people run away, others drink, and then there are those of us who only find escape through the pain. 🏃‍♂️🍺Knife
😊❤️	Direct Representation	Smiling outside, breaking inside. 😊❤️
🔪💔💀	Semantic List	The thoughts keep coming back. I don't know how much longer I can resist. 🔪💔💀

A.4 Emoji Usage in Self-Harm Expressions

We performed a qualitative analysis of emoji usage within self-harm expressions, incorporating relevant examples from the CESM-100 matrix. Our observations are as follows:

Emoji Choices and Composition Specific emojis convey different emotional tones. emojis like 💀 (Skull) and 💀 (Coffin) are closely tied to death and serious intent, while ⚠ (Warning Sign) emphasizes danger or risk. Direct representations of self-harm, such as 🔪 (Kitchen Knife) and ✂ (Scissors), indicate contemplation of harmful actions. Interestingly, these violent and serious emojis often appear alongside supportive or neutral ones like 🤗 (Hugging Face) and ❤️, reflecting internal conflict between despair and a desire for empathy. Emojis like 😢 and ❤️ (Broken Heart) convey vulnerability and an urgent plea for support.

Interaction with Context Emojis gain meaning from their textual surroundings. For instance, "I feel so alone 😢" emphasizes despair, while 🌸 (Blossom) lightens the tone in "I'm having a rough day, but I'll get through it 😊" by introducing optimism.

Pragmatic and Contextual Uses Emojis clarify emotional intent. For example, "I just wish I could escape 😭..." signals distress, while "Some days are tougher than others 😅" (Sweating Smile) conveys a more casual tone. Juxtaposing 😢 (crying face) and ❤️ (Broken Heart) signals a plea for empathy.

Strong Association of Certain Emojis with Self-Harm ❤️ (Broken Heart), 💪 (Flexed Biceps), 🌙 (New Moon), 😢, and 🔪 are the most frequently used emojis in self-harm posts. ❤️ appears 1337 times in self-harm contexts but only 56 times in non-self-harm contexts. 🌙 appears exclusively in self-harm posts (1122 times), indicating its strong association with self-harm-related discourse. 🔪 (Knife) appears 752 times in self-harm contexts and only 147 times in non-self-harm contexts, reinforcing its connection to distress.

Support and Encouragement in Mental Health Conversations

💪 (Flexed Biceps) appears in 299 self-harm posts, suggesting that people use it to express resilience and inner strength while discussing self-harm. 💜 (Purple Heart), 🕊 (Dove), ☁ (Rain Cloud), and 🌈 (Rainbow) are also heavily used in self-harm-related discussions, likely as symbols of hope, peace, and recovery. 🤝 (Hugging Face) appears 388 times in self-harm posts, showing its use in providing emotional support. 💜 (Purple Heart) and 🕊 are used exclusively in self-harm-related contexts, reinforcing their role in discussions about emotional pain and support.

Usage of Specific Emojis in Self-Harm Contexts

Many emojis, such as 💜 (Purple Heart), 🌙 (New Moon), 🕊 (Dove), and ❤️ (Black Heart), appear only in self-harm posts, suggesting that these symbols are strongly linked to distress and mental health discussions. 🔒 (Lock) appears almost exclusively in self-harm posts (449 times), possibly symbolizing feeling trapped or helpless.

We also identified additional dimensions in emoji use within self-harm discussions, illustrated with CESM-100 examples:

- Emotional Nuance:** Emojis like 😢 (Crying Face) enhance emotional depth in expressions of despair.
- Supportive vs. Discomforting:** 🤝 (Hugging Face) fosters support, whereas 🔪 (Kitchen Knife) underscores distress.
- Coping Mechanisms:** Emojis such as 😂 (Face with Tears of Joy) appear in self-deprecating humor and coping strategies.
- Community Engagement:** 💕 (Revolving Hearts) builds connection among users discussing mental health struggles.
- Contrast in Self-Presentation:** 😊 (Smiling Face with Smiling Eyes) and 😢 (Pensive Face) illustrate emotional fluidity.
- Intersection with Identity:** Emojis like 🏳 (Rainbow Flag) reflect personal identity and community experiences.

A.5 Analysis of Serious Intent Emoji Combinations

Dominance of Negative Emotions in Serious Intent Cases In Serious Intent (SI) cases, certain emoji combinations such as 🌙Knife, 🌙Cloud, and 💔Knife are among the most frequently occurring. These combinations often include emojis symbolizing darkness, storms, sadness, pain, and self-harm, reflecting intense emotional distress. Their recurring use underscores the severity of the emotions expressed.

High Occurrence of Nature and Weather-Related Symbols Many Serious Intent cases also prominently feature weather-related emojis like ☽ (new moon), 🌪️ (tornado), and ☔ (rain). This trend suggests that individuals commonly use weather metaphors to communicate mental distress. Rain, storms, and darkness are particularly linked to feelings of depression and sadness, signaling the emotional weight carried by these symbols in the context of serious intent.

Frequent Appearance of Heart and Knife Emojis in SI Cases The combination of ❤️ (broken heart) and ✕ (knife) appears in several high-frequency cases, such as 😢Knife, 😢Knife, and 😢Knife. This aligns with expressions of emotional pain, heartbreak, and themes of self-harm or suicidal thoughts. The juxtaposition of these symbols underlines the severity of distress felt by individuals in these situations, with the heart symbolizing emotional pain and the knife indicating self-harm tendencies.

Peaceful & Hopeful Emojis Are Also Present Despite the overwhelming presence of distressing symbols, some Serious Intent cases feature more hopeful or peaceful emojis. For example, 🕊 (dove), ❤️ (purple heart), ✨ (sparkles), and ☀️ (rainbow) appear in a subset of cases. This suggests that not all messages of distress are devoid of positive elements. Some users may incorporate symbols of support, comfort, and hope, as seen in combinations like 😊❤️🕊️, which possibly represent messages of encouragement or solidarity.

Lack of Casual Mentions in Most Cases An interesting observation is that the top-ranked emoji combinations in Serious Intent cases consistently show zero Casual Mentions Count (CM). This indicates that users typically reserve these complex emoji combinations for serious situations, avoiding casual or lighthearted use. The seriousness of the emotional content in these cases is reinforced by

the absence of trivial references.

Very Few Humor or Sarcasm-Oriented Emoji Combinations While humor or sarcasm may be present in some datasets through emojis like 😂 (laughing), 😒 (upside-down face), and 😐 (awkward smile), this dataset does not show significant occurrences of such patterns. This suggests that, in the context of Serious Intent, individuals are more likely to express their distress directly rather than masking it with dark humor or sarcasm.

Appearance of "Reinforcing" Emojis in Self-Harm Contexts Some self-harm-related emoji combinations, such as 🌙Knife, ❤️Knife, and 😢Knife, include repeated symbols that reinforce the distressing emotions being conveyed. The use of multiple reinforcing emojis in sequence may indicate a stronger emotional expression, emphasizing the severity of the distress or self-harm ideation that the individual is experiencing.

A.6 Experimental Setup: Additional Details

We trained the model for 4 epochs using the AdamW optimizer (weight decay: 0.01) with a learning rate of 4e-5 and a linear decay scheduler, including a 10% warm-up phase (500 steps). A batch size of 16 and sequence length of 256 tokens were used. Training was conducted over 4 epochs, with validation loss closely monitored to prevent overfitting. We used the AdamW optimizer with a weight decay of 0.01 to enhance generalization. A 0.2 dropout rate and gradient clipping at 1.0 were applied for stability and to avoid overfitting. Model evaluation was performed after each epoch, with early stopping activated if no improvement was observed after 3 epochs.

A.7 Information on the Evaluation Metrics

- **F1 for Span Extraction:** We follow the calculation of F1 as done in (Poria et al., 2021), which was inspired by the work of Rajpurkar et al. (2016). This metric measures the average overlap between the prediction and ground truth answer. We treat the prediction and ground truth as bags of tokens and compute their F1 score. We then take the maximum F1 over all of the ground truth answers for a given question and average this over all questions.

• Relevance

- **Intuition:** This metric checks if the rationale includes all the key phrases related

to casual mentions and serious intents.

- **Functionality:**

- * Combines the casual mentions and serious intents into a single text.
- * Converts both the combined text and the rationale to lowercase.
- * Checks if all spans are present in the rationale.

- **Coherence**

- **Intuition:** This metric evaluates the logical consistency and smoothness of the rationale in relation to the spans.

- **Functionality:**

- * Combines the casual mentions and serious intents into a single text.
- * Uses TF-IDF Vectorizer to transform both the combined text and the rationale into vectors.
- * Computes cosine similarity between these vectors.

- **Readability**

- **Intuition:** Measures how easy it is to read and understand the rationale.

- **Functionality:**

- * Uses the Flesch-Kincaid grade level formula to compute the readability score of the rationale.

- **Semantic Similarity**

- **Intuition:** Evaluates the conceptual similarity between rationale and spans.

- **Functionality:**

- * Combines the casual mentions and serious intents into a single text.
- * Uses a pre-trained sentence embedding model (all-MiniLM-L6-v2)⁷ to generate embeddings for both the combined text and the rationale.
- * Computes cosine similarity between these embeddings.

By employing these metrics, we can ensure that the generated rationales for self-harm detection are accurate, consistent, readable, and semantically aligned with the extracted spans.

A.8 Model Details

- **Mental-Alpaca-7B (Xu et al., 2024):** An LLM fine-tuned for mental health prediction

using Alpaca, trained on diverse datasets including Dreaddit, DepSeverity, SDCNL, and CCRS-Suicide for enhanced accuracy in mental health analysis.

- **Llama-3.1-8B-Instruct (AI@Meta, 2024):**

Developed by Meta Inc., Llama 3 is a state-of-the-art LLM outperforming many open-source models on common benchmarks. We use the pre-trained 8B version for our experiments.

- **MentalLLaMA-chat-7B (Yang et al., 2024a):**

Specialized Llama2-chat-7B model fine-tuned on the IMHI dataset for interpretable mental health analysis, offering reliable explanations and state-of-the-art condition prediction.

All models are sourced from Hugging Face⁸.

A.9 Prompts

Table 12 presents the fine-tuning and rationale generation prompts for our proposed setup. We present the prompts for the various baseline setups in Table 13. We strategically selected few-shot examples to ensure a balanced and diverse representation of self-harm classification cases. This included two Casual Mention (CM) examples, two Serious Intent (SI) examples, and one borderline case with ambiguous intent. To enhance robustness, we conducted three runs with different random sets and had human annotators verify the representativeness of the selected samples. We present the prompts used for generating the synthetic data in Table 14.

A.10 Error analysis

- *Ambiguity in Recovery Language:* Posts containing both self-harm references and hopeful or recovery-oriented language may lead to mixed classifications. Models like Mental-Alpaca and MentalLlama sometimes misclassify posts as non self-harm due to a focus on positive language and recovery hashtags, overlooking the serious self-harm content.

- *Current vs. Past Intent:* Distinguishing between current and past self-harm intent is crucial. While all models correctly handle past ideation, the challenge arises in posts where past struggles are intertwined with current emotional states. Clearer delineation in models' handling of past versus present intent could improve accuracy.

⁷<https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2>

⁸<https://huggingface.co/models>

Table 12: Fine-tuning and Rationale Generation Prompts for our Proposed Setup.

Fine-tuning Prompt	Rationale Generation Prompt
<pre> "instruction": "Analyze the following social media post to determine if it is related to self-harm. Identify spans related to casual mention and serious intent regarding self-harm.", "input": { "post text": "I'm really struggling right now. Everything feels pointless. I've thought about cutting again, just to feel something. 😞", "emojis": [{ "emoji": "😞", "usual_meaning": "Disappointed Face", "contextual_meaning": "Sadness", "casual_mention_chance": "Medium", "serious_intent_chance": "High" }], "output": { "classification": "self-harm", "casual_mention_spans": [], "serious_intent_spans": ["thought about cutting again"] } } </pre>	<pre> "instruction": "Using the provided classification and extracted spans, generate a rationale explaining why the post is classified as self-harm or non-self-harm. Consider the spans and the emoji meanings provided.", "input": { "post text": "I'm really struggling right now. Everything feels pointless. I've thought about cutting again, just to feel something. 😞", "classification": "self-harm", "casual_mention_spans": [], "serious_intent_spans": ["thought about cutting again"], "emojis": [{ "emoji": "😞", "usual_meaning": "Disappointed Face", "contextual_meaning": "Sadness", "casual_mention_chance": "Medium", "serious_intent_chance": "High" }], "output": "" } </pre>

Table 13: Zero-Shot, Few-Shot, and Fine-Tuning Prompts for Self-Harm Classification and Rationale Generation

Type	Prompt
Zero-Shot	<p>## Instruction You are tasked with evaluating social media posts to classify them as self-harm or non self-harm. Review the post below and provide your classification. Additionally, write a detailed rationale that explains your decision, taking into account the content and context of the post.</p> <p>## Input {post text}</p> <p>## Output Classification: [self-harm / non self-harm] Rationale: Describes the factors and considerations that guided the classification decision.</p>
Few-Shot	<p>## Instruction You are tasked with evaluating social media posts to classify them as self-harm or non self-harm. Review the following posts and classifications to understand the task. Then classify the new post and write a detailed rationale that explains your decision, taking into account the content and context of the post.</p> <p>## Input Example 1: Post: "I feel so worthless. Sometimes I think about hurting myself to make the pain stop." Classification: self-harm Rationale: The mention of "hurting myself" suggests a serious consideration of self-harm. Example 2: Post: "Just had a really tough day at work, but I'm trying to stay positive!" Classification: non self-harm Rationale: The post expresses frustration but lacks any direct mention of self-harm. New Post: {new post text}</p> <p>## Output Classification: [self-harm / non self-harm] Rationale: Describes the factors and considerations that guided the classification decision.</p>
Fine-Tuning	The single-task fine-tuning prompt for self-harm classification and the subsequent prompt for rationale generation are similar to the one depicted in Table 12, with the exception that they do not include CM and SI spans, as well as emoji information sourced from the CESM-100 dataset.

Table 14: Prompts for generating synthetic self-harm and non-self-harm samples in our dataset

Label of synthetic post	Prompt
self-harm	<p>## Instruction Create a Reddit post where the author reflects on their experiences with self-harm. The post should express a real person's emotions and thoughts, incorporating past incidents, current struggles, or stories related to their journey. Encourage the author to discuss their feelings, the impact on their life, or how they've tried to cope with their condition. The language should be varied and natural, avoiding repetition.</p> <p>## Input</p> <p>Example 1:</p> <p>Post: "I keep finding myself tracing the scars on my skin, a reminder of the war inside my head. The urge to self-harm is like a shadow that never leaves, lurking in the corners of my mind. It's a battle between wanting to feel something and wanting the pain to stop. Sometimes the blade feels like the only friend who understands. But deep down, I know it's not the answer. Trying to hold on to hope like a fragile thread in a storm. One day at a time, one breath at a time. We're warriors fighting invisible battles, and our scars tell stories of survival."</p> <p>serious intent spans: ["tracing the scars on my skin", "urge to self-harm", "the blade feels like the only friend"]</p> <p>Rationale: The reference of "urge to self-harm" and "the blade feels like the only friend" is unquestionably an indication of the author's genuine resolve to cause physical harm to themselves.</p> <p>Example 2:</p> <p>Post: "It's been years since I last hurt myself, but yesterday, I had the strongest urge to do it again. I found an old blade in my drawer and just sat there holding it, trying to convince myself not to use it. I didn't, but it was so hard."</p> <p>serious intent spans: ["since I last hurt myself", "strongest urge to do it again"]</p> <p>Rationale: The use of phrases like "strongest urge to do it again" by the user exhibit a definite intention of causing self-injury, combined with the context of the post.</p> <p>## Output</p>
Non-self harm	<p>## Instruction Create a Reddit post containing phrases which indicate self harm in a joking or casual way. Phrases will be containing the self-harm intentions in a sarcastic way where the author is not actually going to harm himself or herself. It can show the irritation, stress, anger or disgrace just to express the situation. The author should be in a light mood to express his/her thoughts.</p> <p>## Input</p> <p>Example 1:</p> <p>Post: "Great, my boss just gave me another pointless task. Guess I'll just bang my head against the wall."</p> <p>Casual intent spans: ["bang my head against the wall"]</p> <p>Rationale: The mention of "bang my head against the wall" is a violent expression suggesting intention to hurt themselves, but, with the context of the situation, it is clearly just a metaphorical expression for frustration and does not show an actual desire for self harm.</p> <p>Example 2:</p> <p>Post: "Lost all my progress because of a glitch. Might as well just throw myself off a cliff."</p> <p>Casual intent spans: ["throw myself off a cliff."]</p> <p>Rationale: The mention of "throw myself off a cliff" shows exasperation of author about losing his work, but it is clear looking at the emojis that it is definitely not said in a serious intent of harming themselves.</p> <p>## Output</p>