

FigSIM: A Dataset for Fine-grained Suicide Severity and Figurative Language in Suicide Memes

Liulu Chen¹, Elise R. Carrotte^{2,3}, Brian E. Chapman⁴, Jo Robinson^{2,3}, Mike Conway¹,

¹School of Computing and Information Systems, University of Melbourne, Australia,

²Orygen, The National Centre of Excellence in Youth Mental Health, Australia,

³Centre for Youth Mental Health, University of Melbourne, Australia,

⁴O'Donnell School of Public Health, UT Southwestern Medical Center, United States

Correspondence: liuliuc@student.unimelb.edu.au

Abstract

Suicide memes are memes used to express suicide-related thoughts or comment on suicide-related issues. Suicide memes are increasingly common on social media, yet remain poorly understood and potentially harmful. There is an urgent need to better understand their characteristics and to develop appropriate content moderation strategies that limits users' exposure to potentially harmful content. Currently, the absence of annotated datasets of suicide memes remains a key barrier to developing and evaluating automated moderation approaches. In this paper, we introduce FigSIM, the first dataset designed for fine-grained analysis of suicide memes. The dataset consists of 1049 memes, each annotated for (1) fine-grained suicide severity levels, (2) figurative phenomena (e.g. metaphors), and (3) suicide-related content (e.g. suicide method depiction). We benchmark 16 unimodal and multimodal models across three tasks: figurative language, suicide severity, and suicide-related content detection. Overall, FigSIM demonstrates that suicide memes pose unique challenges for both modeling and content moderation. Analysis revealed biases, such as underprediction of higher suicide severity levels, especially for figurative memes. The dataset (including splits used for analyses) is available at: <https://github.com/LiuluChen/FigSIM>.

Content Warning: This paper contains suicide-related content that may be triggering.

1 Introduction

Memes, as a multimodal form of online expression, are increasingly attracting attention from social media researchers (Nguyen and Ng, 2024). By combining visual templates with user-generated text, memes are widely used to express ideas, emotions, and humor across diverse online communities and topics (Literat and Kligler-Vilenchik, 2019). However, their rapid dissemination also raises risks,

especially when they are used to spread harmful content, whether deliberately (e.g., toxic, racist) or unintentionally through engagement with sensitive topics. In particular, mental health memes, especially those involving suicide-related content, present complex challenges for both research and social media moderation.

Suicide is a serious public health problem with cascading impacts on families, communities, and society (Pirkis et al., 2024), and has become the third leading cause of death among 15–29-year-olds worldwide (WHO, 2024). Social media plays an increasingly important role in suicide prevention (Rice et al., 2016). Young people use social media to seek help for mental health concerns, disclose suicidal thoughts, and connect with peers (La Sala et al., 2025). These behaviors offers opportunities to better understand suicidal behavior and identify risk signals for early intervention (Rice et al., 2016). However, research consistently shows that exposure to suicide-related content online could be distressing or harmful for some audiences, especially when posts include explicit, graphic, or method-related details (Robinson et al., 2018). Such content can be emotionally challenging and triggering (Susi et al., 2023), which increase the likelihood of imitative behavior (e.g., self-harm) (Robinson et al., 2025; Susi et al., 2023). Importantly, the impact of suicide-related content may differ between individuals, and content shared without harmful intent may still inadvertently cause distress to those who are exposed to it (Thorn et al., 2023).

The use of suicide memes further blurs the boundary between harmful and supportive communication. The inherent humor characteristic of memes makes it challenging to interpret the images' underlying intent (Vásquez and Aslan, 2021). Suicide jokes are often perceived as dark humor or hyperbole that may not reflect actual suicidality (Smith and Linker, 2021). Further, meme templates are typically visually stylized (Nguyen and

Ng, 2024), and suicide-related content (e.g., methods) may be depicted in stylized forms that affect interpretation. Prior studies report mixed impact of suicide memes: while they may serve as a coping mechanism, they can also obscure personal distress, decrease sensitivity to suicide, or cause negative emotions among viewers (Perez, 2019; Nicomedes et al., 2024). This ambiguity introduces additional challenges for accurate moderation, for both human reviewers and automated systems. Yet, there is no moderation strategy or algorithm that is specifically targeted at suicide memes.

Over the last decade, studies on suicide risk on social media have focused primarily on text (Ji et al., 2021; Matero et al., 2019; Sawhney et al., 2020), and more recently on multimodal data such as naturalistic images and audio (Chatterjee et al., 2022; Badian et al., 2023), typically as binary classification tasks (suicidal vs. non-suicidal). However, assessing different levels of suicide severity risk is crucial for designing more precise and effective interventions (Gaur et al., 2019). The meme format further increases the difficulty of accurately interpreting such content, as traditional machine learning models often struggle to capture implicit cues such as metaphor and irony (Mazhar et al., 2025). Within the mental health domain, prior research has explored the detection of depressive and anxiety symptoms in memes (Huang et al., 2024; Mazhar et al., 2025). Recent quantitative studies have revealed the distinguishable characteristics of suicide memes (Chen et al., 2025). More broadly, numerous studies have focused on the detection of hateful (Huang et al., 2024), toxic memes (Kiela et al., 2020), and figurative detection of political memes (Liu et al., 2022).

However, to the best of our knowledge, no existing studies have yet focused on (1) identifying figurative language in suicide memes, (2) identifying fine-grained suicide risks within suicide memes, (3) assessing the transferability of current memes classification models to suicide meme.

To address the above gaps, we present the first novel and challenging multi-label suicide memes dataset called **FigSIM (Fine-grained Figurative Suicide Meme)**. We propose three tasks: figurative language, suicide severity risk, and suicide-related content detection in suicide memes. We benchmark **FigSIM** against 16 competitive baseline models (20 configurations). The best-performing multimodal model achieves macro-F1 scores of 70.21%, 71.60%, and 58.51% on the three tasks, respec-

tively. Further analysis highlights the difficulty and challenges posed by suicide memes. In particular, we make the following **key contributions**:

- (1) A novel dataset consists of 1,049 suicide memes with fine-grained manual annotations related to figurative phenomena, suicide severity grounded in clinical scales, and suicide-related content.
- (2) A comprehensive benchmark of unimodal and multimodal models on proposed tasks.
- (3) Thorough analyses of the dataset and benchmark results, highlighting the challenges and biases of current models on suicide memes.

2 Related Work

2.1 Meme Understanding

Memes have become widely used for communication on social media, especially among young people (YPulse, 2019). Existing studies on memes can be broadly divided into: classification, interpretation, and explanation (Nguyen and Ng, 2024). In this work, we focus specifically on meme classification. Classification studies have mainly focused on detecting harmful memes, such as those that are abusive (Das and Mukherjee, 2023; Jha et al., 2024), hateful (Kiela et al., 2020), or offensive (Suryawanshi et al., 2020), in the context of binary classification. Other works have also explored additional aspects of memes through multi-class classification, such as figurative language types (e.g., irony, metaphor) (Liu et al., 2022), mental health symptoms (Yadav et al., 2023; Mazhar et al., 2025), humor (Tanaka et al., 2022), and targets (e.g., race) (Mathias et al., 2021).

2.2 Mental health and Social Media

There is extensive research on mental health in social media contexts. Many studies have analyzed language patterns and online engagement to identify symptoms of depression (De Choudhury et al., 2021), anxiety (Shen and Rudzicz, 2017), and suicide ideation (Matero et al., 2019; Ji et al., 2021). A few efforts have been made to classify fine-grained depressive symptoms using post text (Yadav et al., 2020) and suicide risk severity (Gaur et al., 2019). Beyond text, researchers have also explored using visual information in shared images to capture psychological cues (Reece and Danforth, 2017; Manikonda and De Choudhury, 2017). Recent work has increasingly adopted multimodal approaches combining language, vision, and speech to improve detection performance (Chatterjee et al.,

2022; Badian et al., 2023).

However, memes as an emerging multimodal social media format have received limited attention in mental health research. Only a few recent studies have developed models that use mental health memes to identify fine-grained depressive or anxious symptoms (Yadav et al., 2023; Mazhar et al., 2025). Chen et al. (2025) examined the unique characteristics of suicide-related memes and developed a binary classification model. Yet, the literature still lacks a deeper understanding of suicide meme and robust fine-grained classification models designed specifically for them.

2.3 Suicide Memes Moderation

Beyond detection, designing effective moderation strategies for suicide memes remains an open problem. Existing guidelines for online suicide communication, such as #chatsafe (Robinson et al., 2023), rarely address meme-based content, largely due to the lack of empirical evidence regarding which types of suicide memes should be considered harmful. Coarse binary classification provides insufficient granularity to understand the diverse intent and impacts of suicide-related memes (Chen et al., 2025). In addition, the mixed impacts of suicide memes further complicate moderation decisions (Mueller and Abrutyn, 2015).

Therefore, there is a clear gap in existing moderation models and frameworks for addressing suicide memes. An important goal for researchers, social media platforms, and society is to design and develop a context-aware and domain-specific moderation framework that can shape safer online environments. As a first step, we introduce the **FigSIM** dataset, a fine-grained multimodal dataset that captures figurative phenomena and suicide risks in memes. We hope this dataset will provide a foundation for future research on multimodal understanding and fine-grained suicide risk detection, and provide insights for designing moderation frameworks specifically targeting suicide memes.

3 Dataset Construction

In this section, we present a new benchmark dataset: **FigSIM**, for identifying **Fine-grained Figurative Suicide Memes**. The dataset was developed based on clinical assessment scales, online suicide communication guidelines, and figurative language theories. Psychologists and researchers specializing in youth suicide prevention contributed

to each phase of the annotation development and process.

3.1 Dataset Collection

We collected images from r/SuicideMeme, a subreddit dedicated to sharing suicide-related memes that ranks among the top 5% of Reddit communities by size (Reddit, 2024b). Using Pushshift Reddit API (Baumgartner et al., 2020), we retrieved all submissions, extracted image URLs, and downloaded images using the Requests library. We filtered memes by requiring embedded text (if any) to be in English and readable by OCR, and by removing near-duplicate memes (same image with fewer than three-word differences). Memes without overlay text were kept, as purely visual memes are also important and often more difficult to interpret. After filtering, we obtained 1967 suicide memes posted between April 2018 and December 2022. Due to limited annotation resources, we randomly sampled 1050 memes for annotation.

3.2 Annotation Scheme and Categories

We developed a multidimensional annotation scheme consisting of the following **five categories**: (1) **figurative phenomenon**, (2) **suicide severity scale**, (3) **suicide-related content**, (4) **modality**, and (5) **context**. Figure 1 shows the examples for each label in each category. The detailed label definitions for each category see Appendix B.

Category: Figurative Phenomenon. This category captures whether a meme contains figurative expressions. The labels include: (1) *Metaphor*; (2) *Pun/Double meaning*; (3) *Irony/Sarcasm*. Memes are labeled as *None* if no figurative expressions apply. As a meme may contain multiple figurative types, this category is treated as multi-label.

Category: Suicide Severity Scale. The suicide severity scale category was designed using the Columbia-Suicide Severity Rating Scale (C-SSRS) (Posner et al., 2011), a clinical scale widely used to assess the presence and intensity of suicidal ideation and behavior. An additional label *suicide death* is introduced to better capture content referencing fatal outcomes. The final labels include: (1) *None*; (2) *Wish to be dead*; (3) *Suicide ideation*; (4) *Suicide planning*; (5) *Suicide attempts*; (6) *Suicide death*. This category is treated as ordinal.

Category: Suicide-related Content. To enhance the scheme’s applicability and capture the nature of memes and social media platforms, we designed another category: **suicide-related con-**

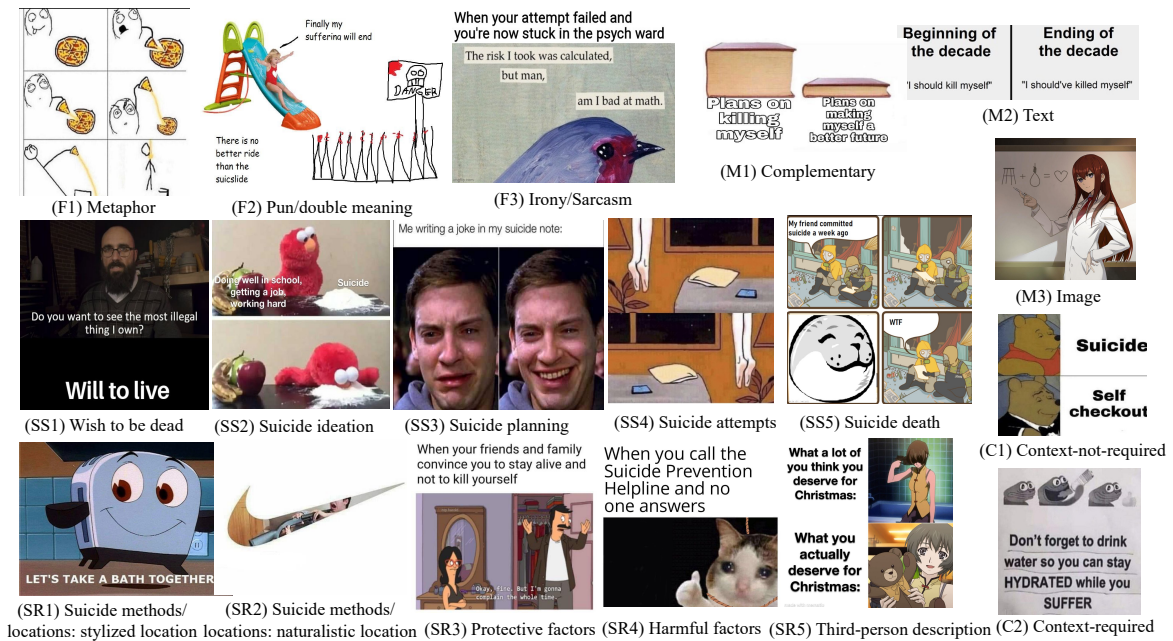


Figure 1: Examples of memes from each annotation category: Figurative Phenomenon (F1–F3), Suicide Severity (SS1–SS5), Suicide-related Content (SR1–SR5), Modality (M1–M3), and Context (C1–C2).

tent, which is treated as multi-label. We used two media guidelines: #chatsafe (Robinson et al., 2023) and the Mindframe (McTernan et al., 2018), which were developed to support safe online communication about suicide. Additional labels were introduced to capture potential cues that are harmful or protective in the context of suicide. The final labels include: (1) *Suicide methods/locations: stylized depiction*; (2) *Suicide methods/locations: naturalistic depiction*; (3) *Protective factors*; (4) *Harmful factors*; (5) *Third-person description*. Memes are labeled as *None* if none apply.

Category: Modality and Context. Two categories, Modality and Context, were introduced for computational analysis purposes. For **Modality**, memes were labeled as *text*, *image*, or *complementary*, depending on whether the meme can be interpreted using a single modality or requires both image and text jointly. The **Context** category identifies whether interpreting suicide-related information is straightforward or requires external or background knowledge (e.g., understanding a suicide-related metaphor). Each meme is labeled as either *context-required* or *no-context-required*.

3.3 Annotation Process

We used Argilla¹ as our annotation platform, hosted in the secure cloud environment at the University of Melbourne.

¹<https://argilla.io/>

Pilot Annotation and Scheme Refinement. We conducted two pilot rounds (50 memes each) to refine the annotation scheme. Four annotators, a registered psychologist and researcher specializing in youth suicide prevention, and three computing and health informatics researchers, independently annotated the pilot batches. We held regular adjudication meetings to resolve disagreements, ensure conceptual clarity, and align interpretations, which informed iterative scheme refinement. During the pilot rounds, we documented brief image descriptions and annotation explanations to capture annotators’ reasoning (Appendix A). Prior to large-scale annotation, a calibration batch (N=50) was annotated to assess inter-annotator agreement (IAA) between annotators.

Large-Scale Annotation and Adjudication. For large-scale annotation, two annotators from the pilot rounds (one from psychology, one from computing) independently labeled each meme. Disagreements were adjudicated by a third annotator from the pilot study, who determined the final gold-standard labels. IAA was monitored for each batch to track annotation consistency.

IAA. We computed IAA separately for each annotation category. For single-label categories, we used Cohen’s κ (Cohen, 1960) to measure chance-corrected agreement. For multi-label categories, we report the average Cohen’s κ across individual labels (excluding *None*). Final agreement was cal-

Category	Type	κ	Agreement
Suicide Severity Scale	Single	0.65	Substantial
Modality	Single	0.88	Almost perfect
Context	Single	0.58	Moderate
Figurative Phenomenon	Multi	0.56	Moderate
Suicide-Related Content	Multi	0.63	Substantial

Table 1: IAA across annotation dimensions. Interpretation following Landis and Koch (1977).

culated on the independently annotated large-scale dataset (N=900), excluding pilot and calibration batches, as shown in Table 1.

4 Dataset Analysis

The final dataset contains 1049 suicide memes annotated across five categories: figurative phenomenon, suicide severity, suicide-related content, modality, and context. One meme was discarded due to quality control. In this section, we analyze the dataset to highlight its complexity and associated modeling challenges.

4.1 Data Distribution and Statistics

Figure 2 shows the label distributions across annotation categories. For Suicide Severity Scale, the dataset is dominated by *suicide ideation*, while very few memes indicate *suicide death*. Most memes are labeled as *irony/sarcasm*, while 146 out of 1049 contain no figurative expression. For Suicide-related Content, *stylized method depictions* and *harmful factors* are most common, whereas *protective factors* and *third-person descriptions* occur less often. On average, each meme contains 1.06 figurative labels and 0.69 suicide-related labels (excluding *None*). Most memes are *complementary* and *no-context-required*. Cross-category relationships are analyzed in Appendix C.

4.2 Inter-annotator Agreement

Table 1 reports IAA across the annotation categories. Overall, Suicide Severity Scale, Suicide-related Content, and Modality show substantial to almost perfect agreement, suggesting consistent agreement in identifying suicide-related risks and the visual-textual structure of memes. In contrast, Figurative Phenomenon and Context show lower agreement (moderate), potentially due to the inherent subjectivity of figurative interpretation and the nuanced judgments required to determine whether external context is needed, which may also vary across annotator demographics. Supplementary metrics and label-wise agreement see Appendix D.

4.3 Visual and Language Analysis

We conducted a visual and language analysis focusing on the category Suicide Severity Scale.

Visual Analysis: We extracted image features, including brightness, colors (HSV), contrast (standard deviation of grayscale values), and face presence (binary Haar-cascade detection). We compared these features across suicide severity scale labels using Kruskal-Wallis tests with Dunn’s post-hoc comparisons (Bonferroni adjusted, $\alpha = .05$). Two features showed significant differences: *Suicide death* memes show lower saturation than all other labels, while contrast is significantly higher in *suicide planning* and *suicide death* memes compared to *wish to be dead* memes. Other features also showed non-significant but interesting trends. For example, face presence was lower in *suicide attempts* than in *suicide ideation* ($p = .063$).

Language Analysis: Using LIWC (Pennebaker et al., 2015) and EmoLex (Mohammad and Turney, 2013) lexicons, we quantified psycholinguistic and emotional features on meme text. The same statistical tests used in the visual analysis were applied. Overall, 35 features differ significantly across the Suicide Severity Scale. Linguistic complexity varies across severity. For instance, *suicide planning* memes contain longer sentences than *wish to be dead* and *suicide attempts*, while *suicide ideation* shows higher word count than *suicide attempts*. Affective features, such as sadness and fear, are more prevalent in *suicide ideation* memes than in other levels. Finally, death-related terms show the strongest distinctions across severity levels, with higher usage in *suicide ideation* memes.

4.4 Moderation Analysis

For moderation analysis, we evaluate three multimodal moderation models: OpenAI omni-moderation-latest, Azure AI Content Safety, Llama-Guard-3-8B. Although our dataset lacks harmfulness labels, these models’ predictions provides insights to how current moderation models behave on suicide memes. As shown in Table 2, the proportion of flagged memes increases with suicide severity across all models. Azure AI Content Safety is the most sensitive, with the highest flag rates for both suicide/self-harm and overall harm categories. Notably, the overall harm flag rate is consistently higher than the suicide/self-harm flag rate, indicating that some memes are flagged due to other potential risks (e.g.,

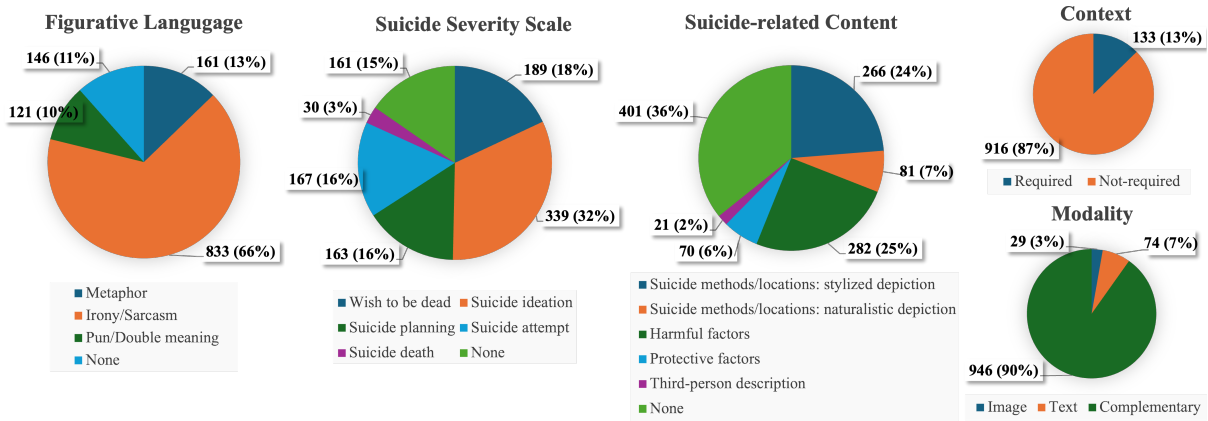


Figure 2: Distribution of annotations across all categories in the FigSIM dataset. Pie charts show the label frequencies and proportions for each annotation category.

	Suicide/Self-harm			Overall Harm		
	Azure	Llama	OpenAI	Azure	Llama	OpenAI
None	0.304	0.304	0.143	0.437	0.304	0.180
Wish to be dead	0.271	0.312	0.243	0.372	0.312	0.339
Suicide ideation	0.619	0.386	0.475	0.678	0.386	0.501
Suicide planning	0.560	0.509	0.497	0.660	0.509	0.521
Suicide attempts	0.661	0.428	0.575	0.745	0.428	0.653
Suicide death	0.633	0.333	0.500	0.767	0.333	0.667

Table 2: Moderation flag rates across suicide severity. “Suicide/Self-harm” refers to flags triggered specifically by self-harm, whereas “Overall Harm” includes any flag raised by other safety categories (e.g., violence).

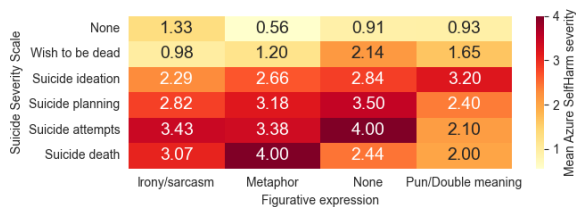


Figure 3: Azure moderation outcomes by suicide severity and figurative expression. Values represent mean self-harm severity score (scale: 0-7).

violence) rather than suicide-related concerns.

Figure 3 shows Azure’s suicide/self-harm severity score across suicide severity levels and figurative phenomenon. At the same severity level, figurative memes tend to receive lower scores than non-figurative memes. The results suggest that figurative language may impact how moderation models interpret suicide-related content.

5 Experiments, Results, and Analysis

We proposed three tasks for the FigSIM dataset: (1) figurative language detection, (2) suicide severity detection, and (3) suicide-related content detection.

Below, we present baseline results for a set of state-of-the-art unimodal and multimodal models.

5.1 Baselines

We evaluate baseline models across text-only, image-only, and multimodal configurations. Our selection aims to provide a representative benchmark including widely used pretrained text and vision, multimodal models, recent multimodal large language models (MLLMs), and task-adjacent meme models from mental health research.

Text only: We extract OCR text from each meme using Google Vision API and fine-tune three pretrained language models: BERT-base (Devlin et al., 2019), MentalBERT (Ji et al., 2022), and RoBERTa (Liu et al., 2019).

Image only: We fine-tune ResNet-50 (He et al., 2016), ViT-Base (Dosovitskiy et al., 2021), and DINOv2 (Oquab et al., 2024).

Multimodal: We finetune CLIP (Radford et al., 2021) and BLIP-2 (Li et al., 2023). We additionally evaluate MLLMs using zero- and few-shot prompting: gpt-5-mini (OpenAI, 2025), claude-sonnet-4-5 (Anthropic, 2025), gemini-3-pro (Google DeepMind, 2025), Llama-3.2-11B-Vision-Instruct (Dubey et al., 2024), Qwen3-VL-8B-Instruct (Bai et al., 2025), and InternVL3_5-8B (Wang et al., 2025).

Specialized meme models: We further evaluate mental-health meme models from adjacent domains to examine their cross-domain generalization to suicide memes. Specifically, we include the depression symptom detection model proposed by Yadav et al. (2023) and the M3H framework (Mazhar et al., 2025), a recent model for mental-health meme classification that incorporates figura-

Input	Model	Figurative detection		Suicide severity detection		Suicide-related detection	
		F1 _{macro}	F1 _{weighted}	F1 _{macro}	F1 _{weighted}	F1 _{macro}	F1 _{weighted}
Text-only	BERT	51.25±4.34	75.98±1.44	48.45±2.78	49.73±2.46	35.22±3.01	44.31±2.66
	MENTALBERT	54.09±4.89	77.06±1.75	48.07±2.46	50.03±2.49	40.69±1.95	49.33±0.96
	ROBERTA	53.05±1.99	76.41±0.91	49.97±3.39	51.81±3.20	38.37±1.12	48.58±1.01
Image-only	ResNet-50	43.18±4.74	72.23±1.77	21.70±3.67	22.55±3.52	25.58±1.32	34.88±0.58
	ViT	55.53±1.78	77.20±0.78	25.88±1.11	27.99±0.69	32.36±3.39	36.82±3.99
	DINOv2	53.02±2.63	76.07±1.50	26.75±4.75	28.13±5.73	31.86±1.90	37.09±1.62
Multimodal	CLIP	62.06±1.27	80.21±0.36	49.97±1.08	52.10±1.32	33.70±1.67	43.75±1.86
	BLIP-2	55.36±4.08	77.46±1.80	39.47±0.34	41.95±0.41	38.81±3.37	44.25±2.65
	GPT-5-mini zero-shot	65.46±2.87	80.40±2.03	65.78±2.28	66.22±2.29	52.52±3.99	53.36±3.03
	GPT-5-mini few-shot	63.58±1.46	76.04±0.91	67.41±1.46	67.95±1.31	53.87±0.46	54.54±0.47
	Claude-sonnet-4-5 zero-shot	70.21±0.82	80.69±0.42	62.51±2.44	63.27±2.31	55.72±1.85	59.96±1.56
	Claude-sonnet-4-5 few-shot	66.87±1.81	70.81±1.20	62.02±0.38	62.80±0.44	55.15±1.50	61.21±1.36
	Gemini-3-pro zero-shot	60.24±0.24	73.15±0.92	66.09±2.21	65.67±2.24	58.51±4.06	58.16±2.53
	Gemini-3-pro few-shot	63.79±1.78	74.38±1.39	71.60±2.43	71.73±2.52	54.57±2.06	57.03±1.26
	Llama-3.2 Vision zero-shot	50.59±0.32	71.73±0.13	34.62±6.79	33.51±0.00	23.98±3.40	37.60±0.00
	Llama-3.2 Vision few-shot	34.42±0.00	66.10±0.00	26.03±0.00	30.83±0.00	22.86±0.00	32.70±0.00
	QWen3-VL zero-shot	59.13±0.11	76.48±0.14	42.03±0.41	44.62±0.52	44.45±6.79	49.00±0.01
	QWen3-VL few-shot	58.85±4.33	75.85±2.95	52.55±0.21	54.45±0.11	42.80±0.00	48.14±0.00
	InternVL3.5 zero-shot	42.81±4.18	74.71±5.74	31.03±0.00	33.87±0.00	45.13±1.38	45.91±1.78
InternVL3.5 few-shot	49.09±0.00	72.07±0.00	31.88±0.00	34.93±0.68	45.18±1.79	48.42±0.60	
Specialized	Yadav et al. (2023)	–	–	49.01±2.98	50.89±2.99	–	–
	M3H	–	–	61.47±1.74	62.58±1.95	–	–

Table 3: Comparison of model performance on the test set ($N = 208$).

tive and commonsense signals.

Implementation details We split the dataset into training, validation, and test sets using a stratified strategy to preserve label distributions across all annotation categories. The data was divided using a 60:20:20 ratio, resulting in 633 training memes, 208 validation memes, and 208 test memes (for detailed label-wise statistics see Appendix E). We trained all supervised models using CrossEntropyLoss for single-label categories and BCEWithLogitsLoss for multi-label categories. To mitigate class imbalance, we applied class-reweighted losses and tuned decision thresholds for multi-label tasks. For MLLMs, prompts were constructed using the annotation definitions and guidelines. We report macro-F1 and weighted-F1 scores on test set. For detailed training configurations see Appendix E.

Due to extreme class imbalance, we merged *Suicide Attempt* and *Suicide Death* into a single class, ***Suicide Attempt/Death***, for suicide severity classification, and **exclude** the *Third-person description* label from suicide-related content detection.

5.2 Performance Comparison

Table 3 shows baseline model performance across tasks. Language models consistently outperform vision models on suicide severity and suicide-related content detection. This suggests that linguistic cues play an important role in inferring suicide intent, while visual information alone is often insufficient.

MLLMs achieve the strongest performance across all three tasks, with commercial MLLMs consistently outperforming open-source MLLMs. Notably, LLaMA refuses to generate predictions for several images; therefore, these instances are excluded from evaluation for LLaMA. This behavior reflects the safety filtering mechanisms of LLMs, which may limit their applicability in this context. Interestingly, few-shot prompting often reduces performance for figurative detection, likely due to prompt sensitivity and subjective interpretation. In contrast, suicide severity detection benefits more consistently from few-shot examples, which indicates that explicit guidance helps models distinguish between fine-grained risk levels. Across models, suicide-related content detection is the most challenging task, with the lowest macro-F1 scores.

Existing models for mental-health memes detection demonstrate certain transferability for suicide severity detection. The M3H framework outperforms many baseline models, but there is still a substantial gap (10%) between M3H and the best-performing model. This gap suggests the need for suicide-specific domain knowledge beyond what is captured by current mental-health meme models.

Table 4 shows per-label F1 scores for the best-performing model selected based on macro-F1 for each task. For figurative detection, performance is highest for *irony/sarcasm*, potentially due to its dominance in the dataset. Suicide severity detec-

Label	F1	Avg. F1	Model
<i>Figurative Phenomenon</i>			
Irony/sarcasm	86.70±0.60	70.21	Claude sonnet zero-shot
Metaphor	65.93±1.28		
Pun/double meaning	58.00±1.54		
<i>Suicide Severity Scale</i>			
Wish to be dead	75.68±4.80	71.60	Gemini 3 pro few-shot
Suicide ideation	71.57±3.13		
Suicide planning	68.14±3.23		
Suicide attempt/death	71.62±1.92		
None	70.98±1.71		
<i>Suicide-related Content</i>			
Protective factors	61.76±8.02	58.51	Gemini 3 pro zero-shot
Harmful factors	47.19±1.05		
Stylized depiction	69.92±2.45		
Naturalistic depiction	55.18±5.97		

Table 4: Per-label F1 scores and Average (per-label F1) for the best-performing model in each task.

tion shows relatively balanced performance across severity levels, with the lowest F1 in *suicide planning*. Suicide-related content detection exhibits lower F1 scores for *harmful factors* and *naturalistic depictions*, indicating greater modeling difficulty.

5.3 Error Analysis

To better understand model performance, we conduct an error analysis on the best-performing models for each task.

Severity calibration bias. Figure 4 shows over- and underprediction of suicide severity conditioned on the true label. For suicide severity detection, models tend to over-predict lower-severity categories (*None* and *Wish to be dead*) but underpredict higher suicide severity (*Suicide planning* and *Suicide attempts/death*). Figurative language is associated with this underprediction. When suicidal risk is expressed figuratively, severity is substantially underestimated more often than for non-figurative memes. This pattern is most evident for metaphorical expressions, though the number of such cases is limited. Importantly, this effect persists even when figurative language is correctly identified, suggesting that the primary challenge lies not only in detecting figurative cues but also in mapping indirect expressions to appropriate severity levels. Detailed analyses are provided in Appendix F.

Context-required cases are challenging. Although most misclassified memes are *no-context-required* due to dataset imbalance, *context-required* memes exhibit relatively higher error rates across tasks. Similarly, while complementary memes dominate the dataset, unimodal cases (e.g., text-

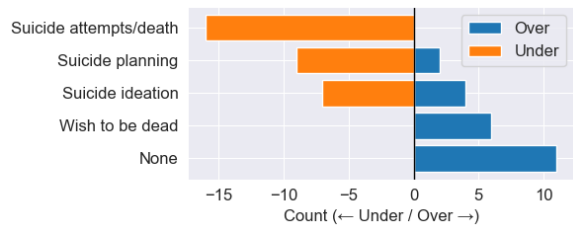


Figure 4: Counts of over- and underprediction for suicide severity, conditioned on true labels.

only) show higher misclassification rates. These observations suggest that models may struggle without context or multimodal cues, though conclusions are constrained by the small number of such cases. Detailed analyses are provided in Appendix F.

6 Discussion and Implications

Our dataset analysis highlights the unique challenges of suicide meme understanding. Lower IAA on figurative phenomenon and context reflects the inherent subjectivity and ambiguity of humor and implicit meaning in memes, which complicates suicide severity interpretation. Existing moderation models exhibit varying sensitivity across suicide severity levels. Memes with higher suicide severity are more likely to be flagged, while figurative expressions are often assessed as less severe. However, future qualitative validation is necessary to determine if flagged memes are truly harmful. Across baseline models and tasks, MLLMs achieve the strongest results, though task difficulty varies substantially within the dataset. Error analysis shows that MLLMs tend to underpredict higher severity categories, and performances may be affected when the context is required. Overall, these findings highlight the complexity of suicide memes and the need for models that better integrate figurative language, multimodal signals, and fine-grained suicide risk.

Our dataset highlights the importance of explicitly considering meme formats in online safety and suicide content moderation. We hope this dataset can support hypothesis generation and evidence-building regarding the characteristics of suicide memes, inform the development of more context-aware moderation strategies, and contribute to guidance on safer online communication about suicide beyond text-based communication.

7 Conclusion

We introduce **FigSIM**, the first dataset designed for fine-grained analysis of suicide memes. By inte-

grating annotations of figurative language, suicide scale severity, and suicide-related content, FigSIM captures the complexity and ambiguity of suicide communication in meme format. Benchmark results across unimodal and multimodal models show that suicide memes pose unique challenges to existing models. Overall, FigSIM provides a foundation for studying suicide memes and for developing more context-aware detection and moderation frameworks beyond text-based communication.

Limitations

There are several limitations in this study. First, the dataset is derived from a single data source, r/SuicideMeme on Reddit. Suicide-related communication, humor norms, and commonsense knowledge may vary across communities, cultures, and platforms. Future work aims to extend this dataset to additional sources to improve coverage and generalizability.

Second, while our annotation scheme was developed in consultation with youth suicide prevention researchers including a registered psychologist, the labels for figurative phenomenon and suicide severity are inherently subjective. The interpretation and annotation rely solely on the content within each meme, under the assumption that the meme reflects indirect self-expression unless explicitly indicated otherwise. Moreover, while our scheme introduces more granularity than prior work, the categories are not exhaustive and do not capture all nuances of suicide-related expression in memes. “Fine-grained” should therefore be understood as relative rather than comprehensive.

Accordingly, the annotations should not be interpreted as clinical judgments, nor do they provide any diagnostic insights about the meme creators. Our main goal is to understand how suicidality is communicated through humor and figurative expression in online memes, to characterize their varying levels of harmfulness, and support the development of more context-aware moderation strategies.

Ethical Considerations

This study involves the collection and analysis of suicide memes, which contain highly sensitive and potentially distressing material. Our primary motivation in constructing the dataset is to support research on the detection of suicide-related online content and to inform responsible moderation

strategies. However, we acknowledge that there is potential for misuse by bad actors in ways we cannot fully anticipate.

All data used in this study were collected exclusively from the publicly accessible subreddit r/SuicideMemes. One ethical challenge is that we can not seek individual consent from Reddit users. To better protect user privacy and confidentiality, we followed the relevant standard outlined by Benton et al. (2017). We removed all personally identifiable information, including usernames, metadata. Only meme images and their corresponding annotations were used for analysis.

This study received ethical approval from the University of Melbourne Human Research Ethics Committee (Project ID: 29894).

To minimize risks related to data distribution, we will release only the annotations and a retrieval script that enables approved researchers to re-download images directly from Reddit, in compliance with Reddit’s data policies (Reddit, 2024a). Access will be restricted to academic, non-commercial use. The dataset (including splits used for analyses) is available at: <https://github.com/LiuliuChen/FigSIM>.

References

- Anthropic. 2025. *Claude sonnet 4.5*. Accessed: 2025-12-07.
- Yael Badian, Yaakov Ophir, Refael Tikochinski, Nitay Calderon, Anat Brunstein Klomek, Eyal Fruchter, and Roi Reichart. 2023. *Social media images can predict suicide risk using interpretable large language-vision models*. *The Journal of Clinical Psychiatry*, 85(1):50516.
- Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, and 45 others. 2025. *Qwen3-VL technical report*. *arXiv preprint arXiv:2511.21631*.
- Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. 2020. *The Pushshift Reddit dataset*. In *Proc. Int. AAI Conf. Web Soc. Media (ICWSM '20)*, volume 14, pages 830–839.
- Adrian Benton, Glen Coppersmith, and Mark Dredze. 2017. *Ethical research protocols for social media health research*. In *Proc. 1st ACL Workshop on Ethics in Nat. Lang. Process. (EthicsNLP '17)*, pages 94–102, Valencia, Spain. ACL.

- Moumita Chatterjee, Piyush Kumar, Poulomi Samanta, and Dhruvasish Sarkar. 2022. [Suicide ideation detection from online social media: A multi-modal feature-based technique](#). *International Journal of Information Management Data Insights*, 2(2):100103.
- Liuliu Chen, Jo Robinson, and Mike Conway. 2025. [What do you meme? – identifying characteristics and user perceptions of suicide memes in social media](#). *Proceedings of the International AAAI Conference on Web and Social Media*, 19(1):385–402.
- Jacob Cohen. 1960. [A coefficient of agreement for nominal scales](#). *Educational and Psychological Measurement*, 20:37–46.
- Mithun Das and Animesh Mukherjee. 2023. [BanglaAbuseMeme: A dataset for Bengali abusive meme classification](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15498–15512, Singapore. Association for Computational Linguistics.
- Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. 2021. [Predicting depression via social media](#). *Proceedings of the International AAAI Conference on Web and Social Media*, 7(1):128–137.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: Pre-training of deep bidirectional transformers for language understanding](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. [An image is worth 16x16 words: Transformers for image recognition at scale](#). In *International Conference on Learning Representations*.
- Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Amy Yang, Angela Fan, and 1 others. 2024. [The Llama 3 herd of models](#). *arXiv preprint arXiv:2407.21783*.
- Manas Gaur, Amanuel Alambo, Joy Prakash Sain, Ugur Kursuncu, Krishnaprasad Thirunarayan, Ramakanth Kavuluru, Amit Sheth, Randy Welton, and Jyotishman Pathak. 2019. [Knowledge-aware assessment of severity of suicide risk for early intervention](#). In *The World Wide Web Conference, WWW '19*, page 514–525, New York, NY, USA. Association for Computing Machinery.
- Google DeepMind. 2025. Gemini: A family of highly capable multimodal models. <https://deepmind.google/technologies/gemini/>. Accessed: 2025-12-02.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. [Deep residual learning for image recognition](#). In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.
- Jianzhao Huang, Hongzhan Lin, Liu Ziyang, Ziyang Luo, Guang Chen, and Jing Ma. 2024. [Towards low-resource harmful meme detection with LMM agents](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 2269–2293, Miami, Florida, USA. Association for Computational Linguistics.
- Prince Jha, Krishanu Maity, Raghav Jain, Apoorv Verma, Sriparna Saha, and Pushpak Bhattacharyya. 2024. [Meme-ingful analysis: Enhanced understanding of cyberbullying in memes through multimodal explanations](#). In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 930–943, St. Julian's, Malta. Association for Computational Linguistics.
- Shaoxiong Ji, Shirui Pan, Xue Li, Erik Cambria, Guodong Long, and Zi Huang. 2021. [Suicidal ideation detection: A review of machine learning methods and applications](#). *IEEE Transactions on Computational Social Systems*, 8(1):214–226.
- Shaoxiong Ji, Tianlin Zhang, Luna Ansari, Jie Fu, Prayag Tiwari, and Erik Cambria. 2022. [MentalBERT: Publicly available pretrained language models for mental healthcare](#). In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7184–7190, Marseille, France. European Language Resources Association.
- Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2020. [The hateful memes challenge: Detecting hate speech in multimodal memes](#). In *Advances in Neural Information Processing Systems*, volume 33, pages 2611–2624. Curran Associates, Inc.
- Louise La Sala, Amanda Sabo, Maria Michail, Pinar Thorn, Michelle Lamblin, Vivienne Browne, and Jo Robinson. 2025. [Online safety when considering self-harm and suicide-related content: Qualitative focus group study with young people, policy makers, and social media industry professionals](#). *J Med Internet Res*, 27:e66321.
- J. Richard Landis and Gary G. Koch. 1977. [The measurement of observer agreement for categorical data](#). *Biometrics*, 33 1:159–74.
- Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. [BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models](#). In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 19730–19742. PMLR.

- Ioana Literat and Neta Kligler-Vilenchik. 2019. Youth collective political expression on Social Media: The role of affordances and memetic dimensions for voicing political views. *New Media & Society*, 21(9):1988–2009.
- Chen Liu, Gregor Geigle, Robin Krebs, and Iryna Gurevych. 2022. FigMemes: A dataset for figurative language identification in politically-opinionated memes. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7069–7086, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.
- Lydia Manikonda and Munmun De Choudhury. 2017. Modeling and understanding visual attributes of mental health disclosures in social media. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, CHI '17, page 170–181, New York, NY, USA. Association for Computing Machinery.
- Matthew Matero, Akash Idnani, Youngseo Son, Salvatore Giorgi, Huy Vu, Mohammad Zamani, Parth Limbachiya, Sharath Chandra Guntuku, and H. Andrew Schwartz. 2019. Suicide risk assessment with multi-level dual-context language and BERT. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, pages 39–44, Minneapolis, Minnesota. Association for Computational Linguistics.
- Lambert Mathias, Shaoliang Nie, Aida Mostafazadeh Davani, Douwe Kiela, Vinodkumar Prabhakaran, Bertie Vidgen, and Zeerak Waseem. 2021. Findings of the WOAHA 5 shared task on fine grained hateful memes detection. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 201–206, Online. Association for Computational Linguistics.
- Abdullah Mazhar, Zuhair Hasan Shaik, Aseem Srivastava, Polly Ruhnke, Lavanya Vaddavalli, Sri Keshav Katragadda, Shweta Yadav, and Md Shad Akhtar. 2025. Figurative-cum-commonsense knowledge infusion for multimodal mental health meme classification. In *Proceedings of the ACM on Web Conference 2025*, WWW '25, page 637–648, New York, NY, USA. Association for Computing Machinery.
- Niall McTernan, Ailbhe Spillane, Grace Cully, Eimear Cusack, Theresa O'Reilly, and Ella Arensman. 2018. Media reporting of suicide and adherence to media guidelines. *International Journal of Social Psychiatry*, 64(6):536–544. PMID: 29972096.
- Saif M. Mohammad and Peter D. Turney. 2013. Crowdsourcing a word-emotion association lexicon. *Computational Intelligence*, 29(3):436–465.
- Anna S. Mueller and Seth Abrutyn. 2015. Suicidal disclosures among friends: Using social network data to understand suicide contagion. *Journal of Health and Social Behavior*, 56(1):131–148. Publisher: SAGE Publications Inc.
- Khoi P. N. Nguyen and Vincent Ng. 2024. Computational meme understanding: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21251–21267, Miami, Florida, USA. Association for Computational Linguistics.
- Christian Jasper C. Nicomedes, Christopher F. Sasot, Geraldine F. Santos, John Mark S. Distor, Priscila B. Marzan, and Aimee Rose Manda. 2024. A convergent-mixed method study on the attitudes and perception towards suicide memes and suicidality. *The Open Psychology Journal*, 17.
- OpenAI. 2025. Gpt-5 model documentation. Accessed: 2025-12-07.
- Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, and 7 others. 2024. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*. Featured Certification.
- James W Pennebaker, Ryan L Boyd, Kayla Jordan, and Kate Blackburn. 2015. The development and psychometric properties of LIWC2015.
- Jessamine Perez. 2019. *Suicide memes: Internet users' anti-future expressions*. Ph.D. thesis, Georgetown University in Qatar.
- Jane Pirkis, Rakhi Dandona, Morton Silverman, Murad Khan, and Keith Hawton. 2024. Preventing suicide: a public health approach to a global problem. *The Lancet Public Health*, 9(10):e787–e795.
- Kelly Posner, Gregory K. Brown, Barbara Stanley, David A. Brent, Kseniya V. Yershova, Maria A. Oquendo, Glenn W. Currier, Glenn A. Melvin, Laurence Greenhill, Sa Shen, and J. John Mann. 2011. The Columbia–Suicide Severity Rating Scale: Initial validity and internal consistency findings from three multisite studies with adolescents and adults. *American Journal of Psychiatry*, 168(12):1266–1277.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763.
- Reddit. 2024a. Reddit privacy policy. <https://www.reddit.com/policies/privacy-policy>. [Accessed 07-09-2024].

- Reddit. 2024b. [Suicidememe. best meme](https://www.reddit.com/r/SuicideMeme/). <https://www.reddit.com/r/SuicideMeme/>. [Accessed 07-09-2024].
- Andrew G Reece and Christopher M Danforth. 2017. [Instagram photos reveal predictive markers of depression](#). *EPJ Data Science*, 6(1):15.
- Simon Rice, Jo Robinson, Sarah Bendall, Sarah Hetrick, Georgina Cox, Eleanor Bailey, John Gleeson, and Mario Alvarez-Jimenez. 2016. [Online and social media suicide prevention interventions for young people: A focus on implementation and moderation](#). *Journal of the Canadian Academy of Child and Adolescent Psychiatry*, 25(2):80–86.
- Jo Robinson, Nicole T. M. Hill, Pinar Thorn, Rikki Battersby, Zoe Teh, Nicola J. Reavley, Jane Pirkis, Michelle Lamblin, Simon Rice, and Jaelea Skehan. 2018. [The #chatsafe project. Developing guidelines to help young people communicate safely about suicide on social media: A Delphi study](#). *PLOS ONE*, 13(11):e0206584.
- Jo Robinson, Louise La Sala, Bridget Kenny, Charlie Cooper, Michelle Lamblin, Matthew Spittal, Caroline Gao, Marina Kunin, Angela Nicholas, Atria Rezwani, Maddox Gifford, Jane Pirkis, and Ann John. 2025. [How do Australian social media users experience self-harm and suicide-related content? A National cross-sectional survey comparing young people and adults](#). *BMC Public Health*.
- Jo Robinson, Pinar Thorn, Samuel McKay, Laura Hemming, Rikki Battersby-Coulter, Charlie Cooper, Maria Veresova, Angela Li, Nicola Reavley, Simon Rice, and 1 others. 2023. [#chatsafe 2.0. updated guidelines to support young people to communicate safely online about self-harm and suicide: A delphi expert consensus study](#). *PLoS one*, 18(8):e0289494.
- Ramit Sawhney, Harshit Joshi, Saumya Gandhi, and Rajiv Ratn Shah. 2020. [A time-aware transformer based model for suicide ideation detection on social media](#). In *Proc. Conf. Empir. Methods Nat. Lang. Process. (EMNLP '20)*, pages 7685–7697, Online. Association for Computational Linguistics.
- Judy Hanwen Shen and Frank Rudzicz. 2017. [Detecting anxiety through Reddit](#). In *Proceedings of the Fourth Workshop on Computational Linguistics and Clinical Psychology — From Linguistic Signal to Clinical Reality*, pages 58–65, Vancouver, BC. Association for Computational Linguistics.
- Nicholas Smith and Shannah Linker. 2021. [Suicide-memes as exemplars of the everyday inauthentic relationship with death](#). *Mortality*, 26:1–16.
- Shardul Suryawanshi, Bharathi Raja Chakravarthi, Michael Arcan, and Paul Buitelaar. 2020. [Multimodal meme dataset \(MultiOFF\) for identifying offensive content in image and text](#). In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pages 32–41, Marseille, France. European Language Resources Association (ELRA).
- Karima Susi, Francesca Glover-Ford, Anne Stewart, Rebecca Knowles Bevis, and Keith Hawton. 2023. [Research Review: Viewing self-harm images on the internet and social media platforms: systematic review of the impact and associated psychological mechanisms](#). *Journal of Child Psychology and Psychiatry*, 64(8):1115–1139.
- Kohtaro Tanaka, Hiroaki Yamane, Yusuke Mori, Yusuke Mukuta, and Tatsuya Harada. 2022. [Learning to evaluate humor in memes based on the incongruity theory](#). In *Proceedings of the Second Workshop on When Creative AI Meets Conversational AI*, pages 81–93, Gyeongju, Republic of Korea. Association for Computational Linguistics.
- Pinar Thorn, Louise La Sala, Sarah Hetrick, Simon Rice, Michelle Lamblin, and Jo Robinson. 2023. [Motivations and perceived harms and benefits of online communication about self-harm: An interview study with young people](#). *DIGITAL HEALTH*, 9:20552076231176689.
- Camilla Vásquez and Erhan Aslan. 2021. [“Cats be outside, how about meow”: Multimodal humor and creativity in an internet meme](#). *Journal of Pragmatics*, 171:101–117.
- Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, and 1 others. 2025. [InternVL3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency](#). *arXiv preprint arXiv:2508.18265*.
- WHO. 2024. [Suicide — who.int](https://www.who.int/news-room/fact-sheets/detail/suicide). <https://www.who.int/news-room/fact-sheets/detail/suicide>. [Accessed 13-01-2025].
- Shweta Yadav, Cornelia Caragea, Chenye Zhao, Naincy Kumari, Marvin Solberg, and Tanmay Sharma. 2023. [Towards identifying fine-grained depression symptoms from memes](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8890–8905, Toronto, Canada. Association for Computational Linguistics.
- Shweta Yadav, Jainish Chauhan, Joy Prakash Sain, Krishnaprasad Thirunarayan, Amit Sheth, and Jeremiah Schumm. 2020. [Identifying depressive symptoms from tweets: Figurative language enabled multitask learning framework](#). In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 696–709, Barcelona, Spain (Online). International Committee on Computational Linguistics.
- YPulse. 2019. [Topline: Social media behavior](https://www.ypulse.com/report/2019/02/20/topline-social-media-behavior2/). <https://www.ypulse.com/report/2019/02/20/topline-social-media-behavior2/>. [Accessed 07-09-2025].

A Pilot Rounds: Image Description and Annotation Note

Two pilot rounds (50 memes each) were conducted to refine the annotation scheme. Four annotators (described in Section 3.3) independently annotated the pilot batches. Disagreements were resolved through regular adjudication meetings to ensure conceptual clarity and align interpretations. During these sessions, one annotator documented the adjudication discussions by writing short image descriptions (i.e., literal descriptions of the image and the meme template sources when identifiable) and summarizing the annotation notes for each meme. These notes were then reviewed and approved by a second annotator.

We emphasize that these explanations were created to clarify annotators' interpretations during guideline development and calibration. They are not intended to be used as ground-truth reasoning, as more rigorous assessment would be required for that purpose. Figure 5 shows an example of the image descriptions and annotation notes.

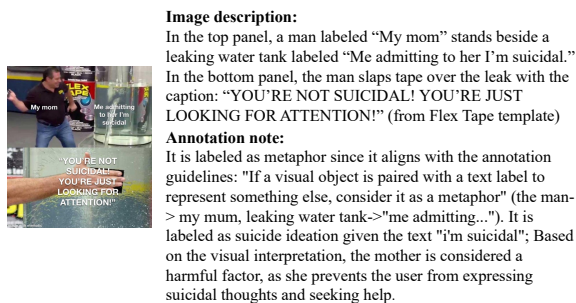


Figure 5: Examples of memes with image description and annotation note.

B Annotation Categories and Label Definitions

B.1 Figurative Phenomenon

The following are the definitions of labels in the Figurative Phenomenon category. Memes are labeled as *None* (F4) if no figurative expressions apply.

- **Metaphor (F1):** A cross-domain mapping across conceptual domains (source to target). One domain of our experience (thoughts) can be understood in terms of another domain.
- **Pun/Double meaning (F2):** A form of expression that suggests multiple meanings through a term, similar-sounding words, or, in some

cases, visual elements that create layered interpretations.

- **Irony/Sarcasm (F3):** Profess to hold beliefs or attitudes contrary to one's underlying belief or attitude for humorous or emphatic effect.

B.2 Suicide Severity Scale

The following are the definitions of labels in the Suicide Severity Scale category. Memes are labeled as *None* (SS6) if no severity applies.

- **Wish to be dead (SS1):** Thoughts about a wish to be dead or not alive anymore.
- **Suicide ideation (SS2):** Thoughts of wanting to end their life/die by suicide/kill oneself.
- **Suicide planning (SS3):** planning one's suicide such as thinking about or selecting methods, location or timing.
- **Suicide death (SS5):** Content that suggests a fatal outcome resulting from a suicide attempt.

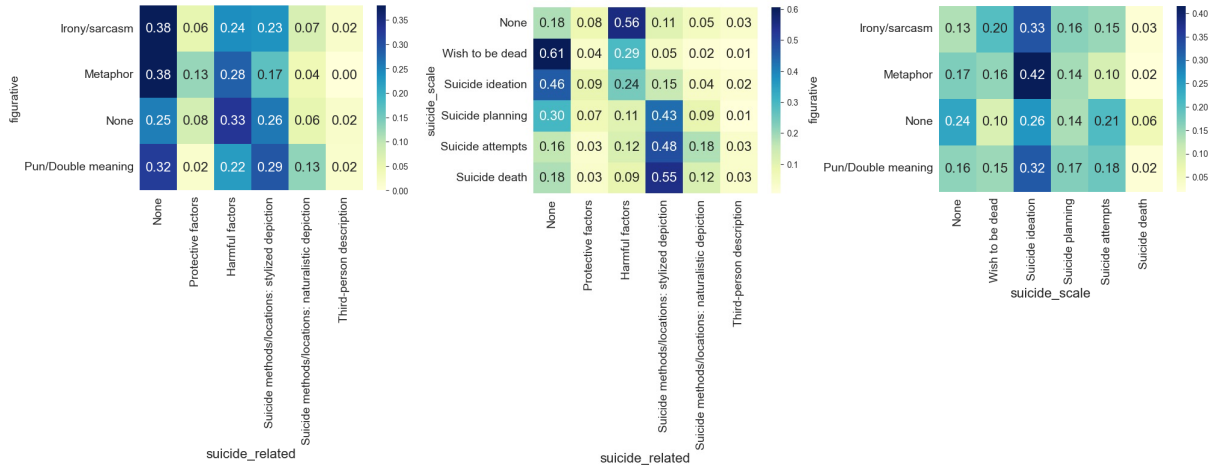
B.3 Suicide-Related Content

The following are the definitions of labels in the Suicide-Related Content category. Memes are labeled as *None* (SR6) if no suicide-related content is present.

- **Suicide methods/locations: stylized depiction (SR1):** Designed elements, symbolic patterns, or text description that could indicate a method/location of suicide.
- **Suicide methods/locations: naturalistic depiction (SR2):** Contain methods or locations where actual suicide could occur.
- **Protective factors (SR3):** Contain content that protects or prevents the user from suicidal thinking or acting on suicidal thoughts.
- **Harmful factors (SR4):** Contain content that reflects potential factors contributing to the emergence or worsening of suicidal thoughts.
- **Third-person description (SR5):** Mentions of suicide-related content that involve someone other than the author, which do not necessarily indicate the author's suicidality.

C Additional Dataset Statistics

Figure 6 shows heatmaps capturing pairwise relationships between suicide severity, figurative phenomenon, and suicide-related content. Figure 6 (a) shows that figurative phenomenon co-occur more frequently with the labels *None*, *harmful factors*, and *stylized method depictions* than with other suicide-related content labels. Figure 6 (b) indi-



(a) Figurative x Suicide-related

(b) Suicide scale x Suicide-related

(c) Figurative x Suicide scale

Figure 6: Cross-category heatmaps (normalized by row; darker colors indicate higher within-row proportions).

	Observed Agreement	Jaccard Similarity
Context	0.92	\
Modality	0.98	\
Suicide Severity Scale	0.72	\
Suicide-related Content	\	0.75
Figurative Phenomenon	\	0.78

Table 5: Observed agreement and Jaccard similarity across annotation categories.

cates that across all severity levels, a large proportion of memes are annotated with no suicide-related content, while stylized method depictions account for a higher proportion in higher-severity categories (e.g., *suicide planning* and *suicide attempts*) relative to other suicide-related labels. In Figure 6 (c), all figurative categories show their highest proportions at the *suicide ideation* level, while labels corresponding to *suicide death* remain sparse. Overall, these heatmaps reveal asymmetric relationships across annotation dimensions, reflecting overlap patterns within the dataset.

D Additional Inter-Annotators Agreement Analysis

In addition to the main IAA results reported in the dataset analysis section (Section 4.2), we provide supplementary agreement metrics for all annotation categories. Table 5 presents the observed agreement for the three single-label categories and the Jaccard similarity for the two multi-label categories. Observed agreement reflects the raw proportion of matching annotations, while Jaccard similarity captures set-level overlap for multi-label decisions.

	Cohen κ	Agreement
Metaphor	0.61	Substantial
Pun / Double meaning	0.57	Moderate
Irony/sarcasm	0.52	Moderate

Table 6: Label-wise IAA: Figurative Phenomenon. Interpretation following (Landis and Koch, 1977).

	Cohen κ	Agreement
Suicide methods/locations: stylized depiction	0.71	Substantial
Suicide methods/locations: naturalistic depiction	0.63	Substantial
Protective factors	0.62	Substantial
Harmful factors	0.62	Substantial
Third-person description	0.59	Moderate

Table 7: Label-wise IAA: Suicide-related Content. Interpretation following (Landis and Koch, 1977).

To further examine consistency within multi-label categories, we additionally report label-wise Cohen’s κ scores in Tables 6 and 7.

E Implementation Details

E.1 Data Split

We split the dataset into training, validation, and test sets using a stratified strategy to preserve label distributions across all annotation categories. The dataset is divided using a 60:20:20 ratio, resulting in 633 training memes, 208 validation memes, and 208 test memes. Table 8 reports detailed label-wise statistics for each split.

E.2 Hyperparameters

All supervised models are implemented in PyTorch. The hyperparameters used in our benchmark ex-

Split	Figurative				Suicide Severity Scale						Suicide Related Content						Modality			Context		Total
	F1	F2	F3	F4	SS1	SS2	SS3	SS4	SS5	SS6	SR1	SR2	SR3	SR4	SR5	SR6	M1	M2	M3	C1	C2	
Train	97	73	499	88	113	207	98	100	18	97	159	49	42	169	13	240	572	44	17	553	80	633
Val	32	24	165	29	37	66	33	34	6	32	54	16	14	56	4	79	187	15	6	181	27	208
Test	32	24	169	29	39	66	32	33	6	32	53	16	14	57	4	82	187	15	6	182	26	208

Table 8: Dataset statistics across Figurative Phenomenon (F1–F4), Suicide Severity (SS1–SS6), Suicide-related Categories (SR1–SR6), Modality (M1–M3), and Context (C1–C2). Label codes (e.g., F1: Metaphor) and full definitions are provided in Appendix B.

Name	Value
Optimizer	AdamW
Learning rate	2e-5
Learning rate schedule	linear
Warmup ratio	0.1
Weight decay	0.01
Gradient norm clipping	1.0
Batch size	16
Epochs	[10, 20]

Table 9: Training hyperparameter setup.

Parameter	Value
Decoding strategy	Deterministic (temperature=0)
Temperature	0.0
Top-p	1.0
Max output tokens	128
Stop sequences	} (custom eos token)

Table 10: Decoding parameters for MLLM inference.

periments are reported in Table 9. For CLIP-based models, we use concatenation-based fusion of image and text representations. All models are trained using the same data splits to ensure fair comparison across configurations. Models are evaluated after each training epoch, and the checkpoint achieving the best macro F1 score on the validation set is saved. The checkpoint with the best validation macro F1 is then used for evaluation on the test set.

MLLMs are evaluated using zero-shot and few-shot prompting without task-specific fine-tuning. The decoding parameters are shown in Table 10. Prompts are constructed using the category definitions and practical guidelines specified in the annotation scheme. Table 11 shows an example for figurative detection. For few-shot prompting, we randomly sample six instances for each task, ensuring that the examples cover all defined labels, and provide both images and their corresponding text labels as input to the models.

Experiments involving open-source and local models were run on NVIDIA A100 GPUs, while API-based LLMs were accessed via commercial endpoints.

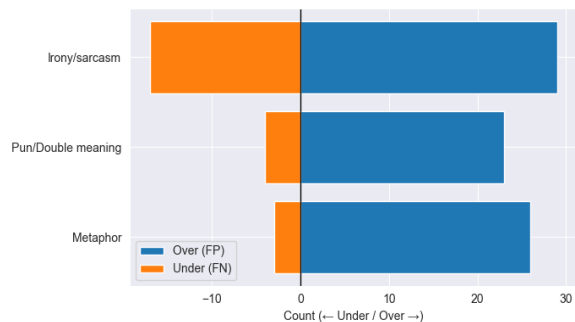


Figure 7: Over- and underprediction by figurative phenomenon.

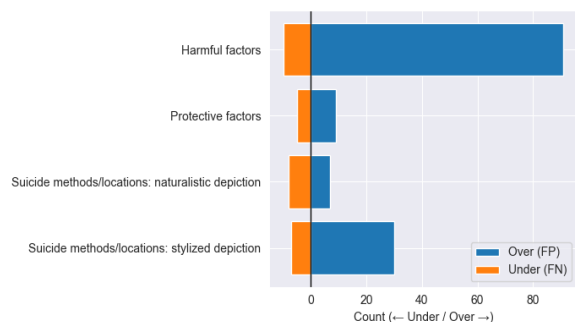


Figure 8: Over- and underprediction by suicide-related content.

F Error Analysis

To better understand model performance, we conduct an error analysis on the best-performing models for each task: Claude-sonnet-4-5 zero-shot for figurative phenomenon detection, Gemini 3 Pro few-shot for suicide severity classification, and Gemini 3 Pro zero-shot for suicide-related content detection.

Label-wise Error Patterns Figures 7 and 8 present label-wise over- and underprediction patterns for figurative phenomenon detection and suicide-related content detection, respectively. Here, over-prediction refers to false positive errors, where a label is predicted but not annotated, while underprediction refers to false negative errors, where an annotated label is not predicted by

Figurative zero-shot prompt

You are annotating a SUICIDE-RELATED meme.

TASK: Identify which figurative language the meme uses.

Allowed labels (you may choose MULTIPLE):

- Metaphor
- Pun/Double meaning
- Irony/sarcasm

DEFINITIONS:

- Metaphor: A cross-domain mapping across conceptual domains (source to target). One domain of our experience (thoughts) can be understood in terms of another domain.
- Pun/Double meaning: A form of expression that suggests multiple meanings through a term, similar-sounding words, or, in some cases, visual elements that create layered interpretations.
- Irony/sarcasm: Professing to hold beliefs or attitudes contrary to one’s underlying belief or attitude for humorous or emphatic effect.

PRACTICAL GUIDELINES:

- Contrast effect in memes: Many memes use a contrast effect such as a happy image paired with a negative or dark caption, expectation vs. reality. This type of memes should be considered as Irony/sarcasm.
- Metaphor: If a visual object is paired with a text label (e.g., “me”, “life”) to represent something else, consider it as a metaphor. This style commonly appears in memes where meaning is constructed by assigning roles or abstract concepts to image elements. Exception: If the text label is simply adding context or clarification to the visual object, rather than representing something else conceptually, it should not be annotated as Metaphor.

RESPONSE FORMAT:

Return ONLY a JSON object of the form:

```
{"figurative": [<zero or more labels from the allowed list>]}
```

Rules:

- Use EXACT strings from the allowed labels (case-sensitive).
 - If no figurative device is present, return an empty list: {"figurative": []}
 - Do NOT include any extra keys or text.
-

Table 11: Zero-shot prompt: figurative phenomenon detection

the model. Across both tasks, models tend to over-predict more salient or frequently occurring categories, such as *irony/sarcasm* in figurative detection and *harmful factors, stylized depiction of suicide methods* in suicide-related content detection. Suicide severity prediction is analyzed separately in the main paper (Section 5.3), where over- and underprediction are defined with respect to the ordinal severity scale.

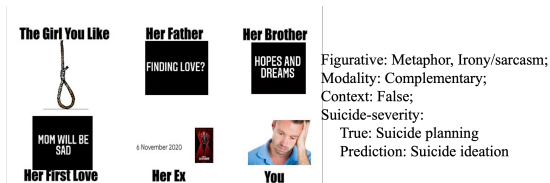
Severity Underprediction by Figurative Subtype. Table 12 reports suicide severity underprediction rates conditioned on the presence of different figurative subtypes. Overall, severity is underpredicted more frequently when figurative language is present than when it is absent. Among figurative subtypes, metaphorical expressions exhibit the highest underprediction rate, followed by irony/sarcasm and puns/double meanings. We note that the number of metaphor and pun/double meaning samples in the test set is limited, and therefore subtype-level patterns should be interpreted descriptively rather than as definitive comparisons.

Context-required cases are challenging. Due to dataset imbalance, most misclassified memes are *no-context-required* and *complementary* in ab-

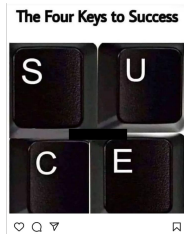
Figurative Property	Underprediction Rate	N
Irony/Sarcasm		
Absent	0.429	7
Present	0.604	48
Metaphor		
Absent	0.551	49
Present	0.833	6
Pun/Double meaning		
Absent	0.571	49
Present	0.667	6

Table 12: Severity underprediction rates conditioned on figurative language properties. Underprediction is computed as the proportion of samples in which the predicted suicide severity is strictly lower than the annotated ground truth, treating severity labels as an ordered scale.

solute terms. However, within the misclassified test set, a substantial proportion of *context-required* memes are misclassified across tasks (5/26 for suicide severity, 16/26 for figurative phenomenon, and 12/26 for suicide-related content detection). Similarly, while *complementary* memes dominate the dataset, unimodal cases exhibit relatively higher error rates: 4/15 text-only memes are misclassified



(a)



Figurative: Pun/double meaning, Irony/sarcasm;
 Modality: Complementary;
 Context: True;
 Suicide-severity:
 True: Suicide ideation
 Prediction: None

(b)

Figure 9: Examples of misclassified cases

for suicide severity detection, 7/15 for figurative phenomenon detection, and 8/15 for suicide-related content detection. Additionally, 3/6 image-only memes are misclassified for suicide severity detection. These results are reported descriptively and suggest increased difficulty when contextual information or multimodal cues are limited, however, conclusions are constrained by the small number of such cases.

Qualitative Examples. Figure 9 shows examples of misclassified memes. In example (a), a suicide method is depicted metaphorically and ironically in a stylized format. Human annotators assigned the label *Suicide planning*, as a specific method is mentioned, indicating the planning phase. Although the figurative labels are correctly identified, the model underpredicts severity, assigning *Suicide ideation* instead of the annotated *Suicide planning*. In example (b), suicidal ideation is expressed through a visual pun that relies on contextual interpretation (success vs. suicide), leading the model to predict *None*. These examples illustrate how indirect, figurative, or context-dependent expressions can obscure severity cues.